SessionIntentBench: A Multi-task Inter-Session Intention-Shift Modeling Benchmark for E-commerce Customer Behavior Understanding

Anonymous ACL submission

Abstract

Session history is a common way of recording user interacting behaviors throughout a browsing activity with multiple products. For example, if a user clicks a product webpage and then leaves, it might be because there are certain features that don't satisfy the user, which serve as an important indicator of on-the-spot user preferences. However, all prior works fail to capture and model customer intention effectively because insufficient information exploitation and only apparent information like descriptions and titles are used. There is also a lack of data and corresponding benchmark for explicitly modeling intention in E-commerce product purchase sessions. To address these issues, we introduce the concept of an *intention tree* and propose a dataset curation pipeline. Together, we construct a sibling multimodal benchmark, SessionIntentBench, that evaluates L(V)LMs' capability on understanding inter-session intention shift with four subtasks. With 1,952,177 intention entries, 1,132,145 session intention trajectories, and 13,003,664 available tasks mined using 10,905 sessions, we provide a scalable way to exploit the existing session data for customer intention understanding. We conduct human annotations to collect ground-truth label for a subset of collected data to form an evaluation gold set. Extensive experiments on the annotated data further confirm that current L(V)LMs fail to capture and utilize the intention across the complex session setting. Further analysis show injecting intention enhances LLMs' performances.

1 Introduction

Modeling and analyzing customer intention is of great importance in the E-commerce domain (Dai et al., 2006; Jammalamadaka et al., 2009; Li et al., 2020). This enables us to give better product recommendations and provide more personalized services (Hu et al., 2008; Zhao et al., 2015; Zhu et al., 2024). Conventional ways of understanding user intention always rely on analyzing user profiles or purchasing records, but such information is not easily retrievable or even missing in real world applications. Therefore, we need a data source with better accessibility and applicability, such as the product purchase sessions, which concludes the user behavior throughout a series of sequential browsing activities. By analyzing the interaction history in this short period of time, we are able to infer the user intention and how it changes over time. The shifting intent behind product searches and inspections can further affect future user interactions. For example, in Figure 1, the customer exposes his intention when he switches from flashy red shoes to plain white ones. After that, browsing for shoes at a much lower price shows customers' need for cheap and cheerful products. By modeling customer session intention and adjusting inferred results when needed, we can provide more customized services in an accurate and timely manner.

Figure 1: An example of customer intention-shift in the session. At each session step, the customer interacts with a new product, may change his purchase intent, and then looks for items with desired features.

Existing work either covers session or intention, but not collectively. There has been an experiment focusing on exploiting the product information within one session and using it to make direct predictions (Jin et al., 2023b), which assembles useful information based on specific product attributes like titles and prices. While some other works explicitly model the user intention behind the single purchase or co-buy behaviors (Xu et al., 2024; Ding et al., 2024). They leverage the most recent user actions for intention understanding and inference, covering only one or two products, but fall short of exploring user preference shifts over a longer horizon, such as sessions. However, Jin et al. (2023b) have shown that session information and fine-grained attribute analysis would help LLMs to give better next-product recommendations. Considering these aspects, it is essential to formulate a method to explicitly model intention over a session period.

But when modeling intention dynamically in more complex purchase contexts, such as sessions, several gaps remain. Firstly, current works only use short-term information and focus on single or co-buy purchases. This approach overlooks the potential motivational intention embedded in earlier user interactions, therefore hindering the models' capability of making reasonable inferences. Furthermore, among various attributes, only product titles and images are used as product inference hints, which omits important dimensions of product information and results in a waste of information from the collected knowledge base. Last but not least, we lack an automated pipeline to streamline the construction of such intention data, there hasn't been any formulation of such tasks or benchmark data to evaluate L(V)LM systems.

To combat this, we first propose SessionIntentBench tasks, consisting of four sequential subtasks tailored to systematically evaluate L(V)LMs' capability in understanding customer intention within session browsing records. Then, we design an automated framework to streamline the collection of detailed product metadata, customer intention, and intention shift within the session by prompting L(V)LM in a multi-step manner.

By applying our method to Amazon-M2 (Jin et al., 2023b), we first filter and collect 10,905 sessions with complete textual and visual data. We enrich the original session with intention entries and obtain 1,132,145 possible intention pathways. After that, we further conduct human annotations to 8,980 sampled intention trajectories to form an evaluation benchmark. Then, we carry out extensive experiments over more than 20 L(V)LMs by applying different evaluation settings and prompting techniques, along with extra fine-tunings. Our findings indicate that current L(V)LMs struggle with the proposed tasks. Further analyses reveal potential underlying causes behind the observed low model accuracy and introduce intention injection as a possible way of assisting models' understanding of session intent and improving performances. We will make our code, data, and models publicly available after acceptance.

2 Related Works

2.1 Intention Understanding

Intention is the internal mental state that affects people's decision-making (Alford and Biswas, 2002). By analyzing the inner intention states of the users, service providers are able to present more personalized products (Dai et al., 2006) and give back more accurate responses (Zhang et al., 2016). In E-commerce, customer intention is crucial in understanding their purchase behaviors and preferences (Shim et al., 2001). There has been ongoing research trying to decode how to model shopping intention. For example, using history information like tags (Wang et al., 2025) and co-buy behaviors (Yu et al., 2023; Xu et al., 2024). Recently, studies show that LLMs are struggling to connect the dots between intended products and user intention (Ding et al., 2024). However, figuring out the items the user wanted is even more difficult when it comes to more complex settings like session histories (Jin et al., 2023b). To bridge the gap between understanding intention and providing more precise shopping aids, we formulate SessionIntentBench tasking L(V)LMs to infer intent by leveraging session metadata from multiple angles.

2.2 Purchase Session in E-commerce

Purchase session is a record of customer interaction history, which has been becoming an increasingly hot area of research (Alves Gomes et al., 2022; Jia et al., 2023; Wang et al., 2024b). Various methods are proposed trying to exploit the abundant information contained here, such as using deep reinforcement learning models (Bharadwaj et al., 2022), leveraging graph neural networks (Jin et al., 2023a), and carrying out complex logical reasoning techniques (Liu et al., 2023b). While Jin et al. (2023b) systematically introduces session information as an important factor for understanding sequential interacting behavior, Liu et al. (2023b) points out that product attributes play a pivotal role in enhancing user intent capture. This shows that a more fine-grained framework of session intention evaluation is needed. Furthermore, recognizing that multiple intentions can coexist within a session, researchers have explored various approaches to enhance product recommendations. Sun et al. (2024) iteratively updates an intention ranking prompt to optimize recommendations, while Choi et al. (2024) train a neural network to learn intention embedding representations and refine selections accordingly. While these works aim to provide more precise product recommendation, our research focuses on improving language models' intention understanding and reasoning ability using semantic intention representation. Using the summarization and generation ability of L(V)LMs, in SESSIONINTENTBENCH, we extract and incorporate session intent metadata from multiple aspects for more comprehensive intention capturing.

3 Problem Definition

3.1 SESSIONINTENTBENCH Task Definitions

First of all, we give definitions to the tasks in SESSIONINTENTBENCH. We propose to model the intention shift from four aspects as a comprehensive formulation, as outlined in Figure 2, to facilitate the creation of a L(V)LM shopping agent that is able to: (i) Detect the attribute that is decisive in the intention shift. (ii) Model intention trajectories with mined attributes and leverage them to give better predictions on future interactions. (iii) Compare between the most recently viewed product with previously interacted ones and use this comparison to validate the plausibility of the inferred intent. (iv) Leverage modeled intention trajectories to predict future product interaction preferences.

To this end, we propose tasks that each emphasize a different angle of analysis. Assume we have collected the customer interaction history over time steps $t = 1, 2, \ldots, T$, i.e., the interacted products $P_1, P_2, \ldots, P_T$ and attributes that affect customer decision-making at each step $A_1, A_2, \ldots, A_T$. Then the history information up till time step t can be summarized as $\mathcal{H}_t = \{(P_j, A_j)\}_{j=1}^{t}$. We denote inferred customer intention as $I_1, I_2, \ldots, I_T$, and comparisons between interacted items and internal intent of the current step and previous step $C_1, C_2, \ldots, C_T$. Further discussions on the theory, intuition, task clarifications, and additional details can be found in the Appendix B, C.

TASK 1: Intent-Based Purchasing Likelihood Estimation: The first task asks the model to verify whether the last proposed intention is a good alignment with the new product we are going to interact with. The model will be given history information $\mathcal{H}_{t-1}$, the proposed intention I_{t-1}, and new product P_t. It is asked to output a likelihood estimation score $\mathcal{S}_1(P_t, I_{t-1}) \mid_{\mathcal{H}_{t-1}} \in \{0, 1, 2, 3\}$ for the customer to interact with P_t, where 3 means the most likely and 0 means the least probable.

TASK 2: Purchasing Likelihood Inference via Valued Attributes Regularization: The second task requires the model to verify whether the proposed valued attributes of the user are an essential element of the actual unseen product. The model is provided with the history information $\mathcal{H}_{t-1}$, the proposed valued attribute A_{t-1}, and the new unseen product P_t. The model is required to output an estimated interaction likelihood score $\mathcal{S}_2(P_t, A_{t-1}) \mid_{\mathcal{H}_{t-1}} \in \{0, 1, 2, 3\}$ for the user to interact with P_t under the assumption that the user values the product feature A_{t-1}, where 3 means the most likely and 0 means the least probable.

TASK 3: Intention Justification via Comparison: To ensure that the proposed intent is reasonable and to verify against potential hallucinations, the third task asks the model to justify whether the proposed C_t provides a reasonable justification for the user to interact with P_t after seeing P_{t-1}. Formally, the model is tasked to output a score $\mathcal{S}_3(C_t, P_{t-1}, I_{t-1}, P_t, I_t) \mid_{\mathcal{H}_{t-1}} \in \{0, 1, 2, 3\}$ indicating the plausibility of the generated comparison.

TASK 4: Intention Evolution Modeling: The final task we proposed aims to test the model's ability to help the recommendation systems decide whether to further recommend similar products or not. Providing the model with all the historical information and inferred purchasing intent, we ask it to choose from exposing the user to (a) Similar products under the same category, (b) Products with different features but still under the same category, (c) Products under different category (exploring more to figure out user preferences). If we map the choices to numerical score $\{1, 2, 3\}$, then we formalize the task as questioning for $\mathcal{S}_4(exploration, I_t) \mid_{\mathcal{H}_t} \in \{1, 2, 3\}$. Note that the degree of exploitation decreases and exploration increases as the score increases.

3.2 Dataset

We obtain products in series of sequential interactions from Amazon-M2 dataset (Jin et al., 2023b) and product image information from Amazon Review Dataset (Hou et al., 2024). We leverage the abundant textual information (such as titles, price, color, material, etc.) mentioned in Amazon-M2 and retrieve corresponding product images from Amazon Review Dataset to curate the dataset. After filtering out products whose links are missing or not accessible, we obtained 10,905 sessions with complete textual and visual components.

4 SessionIntentBench Construction

In this section, we present our methodology of constructing the intention tree from the source data we collected and how we curated the SessionIntentBench. An overview is presented in Figure 2. Our framework consists of four steps: (i) Extract attributes for session products to provide aids for model inference in later steps. (ii) For each time step, prompt the models to mimic customer behavior and infer multiple intentions from previous interactions. (iii) Enrich intention tree structure with more nuanced inter-session intention metadata analyses, which is taken from multiple perspectives on how and why intention shifts over time. (iv) Conduct human annotations for the constructed tree

4.1 Multi-modal Attribute Extraction

The first step aims to extract product attributes that can better assist LVLMs in analyzing user intention shift in later stages. To achieve this, we use GPT-4o-mini (OpenAI, 2024a) as the extraction tool and provide it with ensembled textual and visual information of session products. The LVLM is then asked to output a general classification of the product itself, and to categorize then instantiate the extractable features of the product, for example (e.g., *color: white, size: 7.5 inches*).

4.2 Customer Intention Generation

To build up the intention tree based on the product purchase session, we first fill up the tree bones with predicted user intention using L(V)LMs. The intentions are inferred at each time step following the session time frame. Starting with the first item in the session, we ask the model to infer a list of possible intention $< I_{t1}, I_{t2}, I_{t3}, \dots > \big|_{t=1}$ based on textual and visual information of the product user interacted, where the prompt is demonstrated below. Then, repeat the inference every step as we add the next new session product into the visible list of items to the model.

Genre	Property	Train	Test
Basic Info	# Sessions (uni.)	8963	5306
	# Sampled Tasks	28736	7184
	Avg. # Products	3.4163	3.4123
	Avg. # Intention	3.4163	3.4123
Session Len	# $Len = 3$	18956	4752
	# $Len = 4$	7598	1902
	# $Len = 5$	2182	530
Task Num	# Task 1	7153	1827
	# Task 2	7171	1809
	# Task 3	7154	1826
	# Task 4	7258	1722

Table 1: Statistics of the sampled and annotated data for the SessionIntentBench benchmark. *uni.* means unique sessions are included. Note that the dataset is randomly shuffled and then sampled by 13,003,664 tasks or 1,132,145 intention trajectories, not by sessions.

To make the intention instantiation successional, we add the intention information of the previous time step $\{I_i\}_{i=1}^{t-1}$ (<Prev Intent>) to facilitate the model to do the reasoning. And at each time we do the inference, we will only use one intention chosen from the previous step intention, to ensure the coherent intention trajectory sampling. More specifically, the model is constraint to output the five most possible user intentions, denoted as $\{\text{<New Intent i>}\}_{i=1}^{5}$, prior to the fifth product at each iteration. This process is referred to as *branching*, as it resembles the growth of a tree, wherein each new intention branches out from the initial concept, akin to twigs dividing into finer branches. And starting from the fifth product, we only infer one possible intention at a time to control the exponential growth of the tree size (by setting $|\text{<New Intent>}|=1$).

```
<TASK-PROMPT>
<INPUT:>
<Prev Intent><Prev Products><New Product>
<OUTPUT:>
<New Intent 1><Attr 1><Rationale 1><Comp 1>
<New Intent 2><Attr 2><Rationale 2><Comp 2>
...
<New Intent 5><Attr 5><Rationale 5><Comp 5>
<INPUT:>
<Prev Intent><Prev Products><New Product>
<OUTPUT:>
```

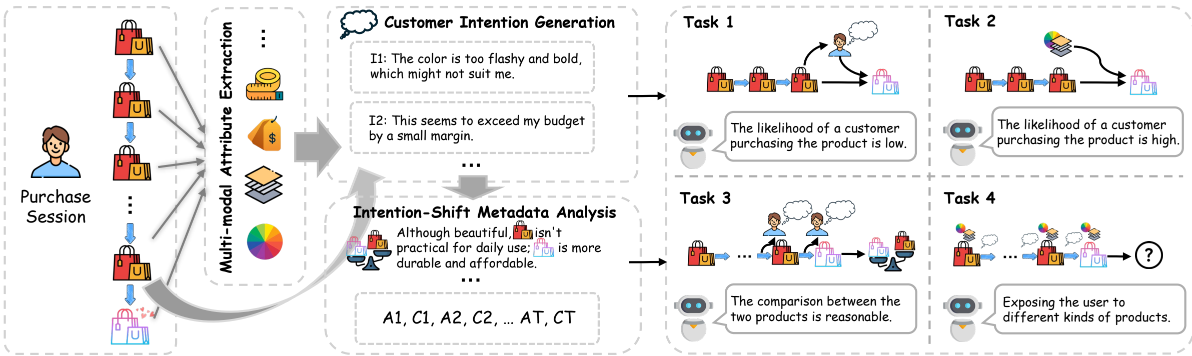

Figure 2: Overview of SESSIONINTENTBENCH and the construction pipeline. Multi-modal attribute extraction is conducted first as an aid for further step intention generation. Metadata analyses are conducted afterward to provide a more fine-grained and detailed inspection of intention shifts in the session interaction. Here, A_i, C_i stands for attributes and comparisons at i-th step. Different task is associated with different collections of metadata.

4.3 Inter-Sessions Intention Shift Analysis

Following this, we want to investigate the specific reasons behind each intention shift before and after the customer sees each product and how that might influence the further decision-making of the customer. The prompt we used for generation is given above. To ground the reasoning on the actual product metadata, we require the model to point out the most likely feature <Attr> A_t that affects the user choices. Furthermore, we ask for a more comprehensive comparison <Comp> C_t between the last product P_t the previous one P_{t-1}, so that it provides logical support for the modeled intention pathways. In order to help models reason better, we require the model to provide rationales (<Rationale>) behind the generations as part of the output. We collect this analysis metadata in the format of one general categorization plus one detailed instantiation, e.g., *book type: fiction, price: $20*.

4.4 Human Annotation

We hire Amazon Mechanical Turk annotators to label a randomly sampled subset of our data to balance cost and quality. We ask the workers to annotate emphasizing the following perspectives: (1) The alignment of proposed intention I_t and session products P_{t+1}. (2) The consistency between the inferred valued attribute A_t and actual interacted products P_{t+1}. (3) The plausibility of the generated intention comparison C_t. (4) Predictions on further intention pathways based on historical information. In this way, the session intention could not only provide insights into the thinking process of customers but also meaningful references for when to explore and when to exploit product recommendation systems. To simplify the annotation process, the annotators are only asked to assign a likelihood score or plausibility score for each task in a format roughly similar to *yes, maybe yes, maybe no, no* (corresponds to $\mathcal{S} = 3, 2, 1, 0$). We carried out multiple rounds of annotation worker selections with different criteria to ensure high annotation quality. More details are in Appendix D.

5 Evaluations and Analyses

5.1 Intrinsic Evaluations

We present our detailed statistics in Table 1. By filling up the tree with intention on 10,905 sessions, we obtain more than 1,950,000 intention entries and 1,100,000 intention trajectories. The majority of these sessions contains less than four products, though long sessions also exist with up to 18 products. To sample a subset of sessions to form the SESSIONINTENTBENCH, we first retrieve candidate sessions with lengths three to five. We then sample 2,000 sessions with 2 trajectories per session and later add another disjoint 1,445 sessions with 4 trajectories per session. That gives 9,780 trajectories in total. To grant the model with full information available, we only query the tasks at the end of each session time step, that is, using all the products available and masking the last product when querying the TASK 1, 2.

5.2 Baselines and Model Selections

Evaluation Metric We use accuracy and Macro-F1 score as evaluation metrics. Accuracy is defined as the percentage of questions that are correctly answered. We regard scoring 0,1 in TASK 1-3 as the true positive label and scoring 0 as the one for TASK 4. To start with, we include the Random Selection and Majority Vote score of each task.

Models	Intent-Based Inference		Valued Attributes Reg.		Comparison Just.		Evolution Modeling	
	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
Random	50.00	50.00	50.00	50.00	50.00	50.00	54.38	35.00
Majority	62.30	76.77	54.35	NaN	71.80	83.58	63.15	NaN
LLM *(Zero-Shot)*								
Meta-Llama-3.1-8B	56.87	70.98	49.36	55.10	71.30	83.24	39.26	53.01
Gemma-2-9B	57.03	69.37	52.18	49.44	41.68	44.19	53.77	34.54
Mistral-7B-v0.3	**62.17**	76.52	47.65	**64.08**	71.30	83.24	39.61	53.53
Ministral-8B	56.98	69.33	51.58	50.48	68.02	80.48	38.27	54.08
Mistral-Nemo-12B	53.09	63.82	51.63	35.04	56.79	69.71	47.15	45.11
Falcon-3-7B	57.31	71.74	52.24	49.17	67.36	79.41	44.36	49.68
Falcon-3-10B	54.95	66.93	51.35	48.59	65.49	78.24	43.84	45.89
Qwen-2.5-3B	54.19	64.42	51.96	41.87	68.62	81.01	37.63	53.98
Qwen-2.5-7B	58.62	71.92	51.02	56.18	70.59	82.61	40.07	51.86
LVLM *(Zero-Shot)*								
LLaVA-v1.6-mistral-7b	58.29	71.90	47.48	62.27	62.94	75.11	37.62	54.20
LLaVA-v1.6-vicuna-7b	62.01	76.55	46.93	63.88	71.27	83.22	37.21	**54.24**
Qwen-2-VL-7B	58.73	71.48	50.63	56.37	70.61	82.73	37.67	53.95
Meta-Llama-3.2-11B-V	45.10	61.38	38.41	52.35	42.11	59.20	36.33	53.23
L(V)LM *(Few-shots)*								
Mistral-7B-v0.3	60.43	74.60	50.64	61.39	67.09	79.08	43.44	49.85
Qwen-2-VL-2B	58.02	73.46	40.63	58.40	66.70	79.92	36.99	53.45
LLaVA-v1.6-vicuna-7b	51.06	**77.26**	22.61	62.92	66.81	82.99	27.99	54.07
L(V)LM *(Fine-tuned)*								
Meta-Llama-3.1-8B	52.82	63.84	51.46	46.27	70.76	82.82	51.92	33.01
Meta-Llama-3.2-3B	55.67	66.80	51.80	46.70	69.61	81.93	51.63	32.66
Mistral-7B-v0.3	57.47	68.56	50.64	44.64	67.69	79.88	55.69	31.69
Ministral-8B	58.35	69.55	51.24	45.01	66.54	79.10	55.57	35.11
Mistral-Nemo-12B	56.10	66.80	52.02	46.68	67.74	79.81	55.81	32.95
Qwen-2.5-7B	54.02	65.63	52.02	46.75	69.50	81.66	54.47	31.59
Falcon-3-7B	55.77	65.02	**52.85**	48.46	**71.41**	**83.30**	54.65	36.86
L(V)LM *(Proprietary API)*								
GPT4o-mini	57.44	69.34	51.95	43.81	71.19	83.13	38.39	53.90
GPT4o-mini (5-shots)	58.83	71.86	49.32	53.01	65.25	78.11	46.51	46.96
GPT4o-mini (COT)	57.26	69.02	51.87	43.33	68.86	81.22	42.81	49.42
GPT4o	55.05	65.33	49.75	36.27	56.30	67.51	41.64	52.39
GPT4o (5-shots)	53.10	63.58	44.20	38.61	54.94	65.01	43.44	48.41
GPT4o (COT)	53.30	61.91	52.00	36.08	49.50	50.87	**58.42**	13.73

Table 2: Evaluation results (%) of various (L)LMs on the annotated testing sets of SessionIntentBench. The best performances within each method are underlined, and the best among all methods are **bold-faced**.

Model Selections Then, we test out a diverse set of L(V)LMs on SessionIntentBench. Since all the tasks we proposed belong to classification setups, we choose accuracy and Macro-F1 score as evaluation metrics. The models we selected, as given in Table 2, can be classified into three genres: **(I) Open L(V)LMs with Zero-shot**: Firstly, we select a vast collection of models from different companies or organizations. Text-to-text models includes Llama3.1, Llama3.2 (Grattafiori et al., 2024), Gemma2 (Team et al., 2024), Mistral (Jiang et al., 2023), Falcon (Almazrouei et al., 2023), and Qwen2.5 (Qwen et al., 2025). Image-text-to-text models includes LLaVA (Liu et al., 2023a), Qwen2-VL (Wang et al., 2024a), and Llama with Vision (Grattafiori et al., 2024). Models under this category are prompted using zero-shot. **(II) fine-tuned L(V)LMs with Zero-shot**: Following that, we fine-tuned Llama3.1, Llama3.2, Mistral, Falcon3, Qwen2.5 on partitioned training set and evaluate them on the testing set. **(III) Proprietary L(V)LM API with Several Different Prompting Techniques** : Lastly, we test out GPT-4o and GPT-4o-mini (OpenAI et al., 2024; OpenAI, 2024a) using zero-shot prompting, 5-shots prompting and Chain-of-Thought prompting (Wei et al., 2023).

5.3 Main Evaluation Results

Intention Evolution Modeling (task 4) is the most challenging task. Our experiments show that the average accuracy of the zero-shot models on Task 4 is 42.34%. Compared to the second hardest task (*Purchasing Likelihood Inference via Valued Attributes Regularization*), which models scored 49.63%, there is a great gap of 7.29% on Task 4. After being fine-tuned, all open models are able to achieve a minimum accuracy of 51.92%, while the top performing one (Mistral-Nemo-12B) scores 55.81%, just above the Random vote accuracy. It is worth noticing that GPT-4o with Chain-of-Thought prompting is able to achieve the high-

Training Data	Backbone	Intent-Based Inference		Valued Attributes Reg.		Comparison Just.		Evolution Modeling	
		Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
Zero-shot	Llama-3.1-8B	56.87	70.98	49.36	55.10	71.30	83.24	39.26	53.01
	Llama-3.2-3B	54.68	63.97	52.02	43.48	33.13	49.48	51.34	36.61
	Mistral-7B-v0.3	**62.17**	**76.52**	47.65	64.08	71.30	83.24	39.61	53.53
	Ministral-8B	56.98	69.33	51.58	50.48	68.02	80.48	38.27	**54.08**
	Falcon-3-7B	57.31	71.74	52.24	49.17	67.36	79.41	44.36	49.68
	Qwen-2.5-7B	58.62	71.92	51.02	56.18	70.59	82.61	40.07	51.86
SIB	Llama-3.1-8B	52.82	63.84	51.46	46.27	70.76	82.82	51.92	33.01
	Llama-3.2-3B	55.67	66.80	51.80	46.70	69.61	81.93	51.63	32.66
	Mistral-7B-v0.3	57.47	68.56	50.64	44.64	67.69	79.88	55.69	31.69
	Ministral-8B	58.35	69.55	51.24	45.01	66.54	79.10	55.57	35.11
	Qwen-2.5-7B	54.02	65.63	52.02	46.75	69.50	81.66	54.47	31.59
	Falcon-3-7B	55.77	65.02	52.85	48.46	**71.41**	**83.30**	54.65	36.86
MIND + SIB	Llama-3.1-8B	60.10	68.81	55.33	48.67	70.54	82.54	57.72	39.74
	Llama-3.2-3B	59.88	67.92	55.28	50.15	64.02	75.48	58.54	40.50
	Mistral-7B-v0.3	60.04	69.96	52.90	45.87	67.69	79.56	**59.93**	37.16
	Ministral-8B	58.24	67.33	53.95	47.44	65.44	77.01	58.77	40.93
	Qwen-2.5-7B	59.00	67.65	53.95	48.62	63.09	74.98	57.84	39.30
	Falcon-3-7B	58.57	68.42	**55.94**	**50.22**	71.30	83.25	58.36	40.00

Table 3: Evaluation results (%) of transferring knowledge from MIND to aid SESSIONINTENTBENCH. The best performances among each method are underlined, and the best ones among all methods are **bold-faced**. We abbreviate SESSIONINTENTBENCH as SIB.

est rate of 58.42% among all models and methods. This might be because the larger model size and the trick of enabling reasoning at run time could help the model to better mimic the thinking process of a real-life customer. This result shows that more works need to be done to level up the model's capability of capturing long-term user intention trends.

Fine-tuning can greatly improve the poor performing models, but struggle to help the mediocre ones. Poor performing models, which we referred to as the ones that receive a low score compared to models under the same category in some evaluation tasks, can quickly pick up relevant capabilities by being fine-tuned on the training set before testing. For example, LLAMA-3.2-3B shows poor performance on TASK 3 (*Intention Justification via Comparison*), but after being fine-tuned on SESSIONINTENTBENCH, it shows a leap of performance by 36.5% and demonstrates comparable outcome with other larger 7B or 8B models. The mediocre performing models, which we referred to as the ones that score near the highest among the models but still struggle to surpass the top accuracy records. Among the proposed tasks, the largest maximum accuracy raise from zero-shot to fine-tuned happens at TASK 4, with a lift of 2.04% in the highest score. As a result of these two factors, the variance between different models shrinks after fine-tuning. See the Appendix F.2 for additional discussions on fine-tuning.

LVLMs struggle to make good usage of visual signals. In comparison to LLMs, which only use textual signals as the input, LVLMs can refer

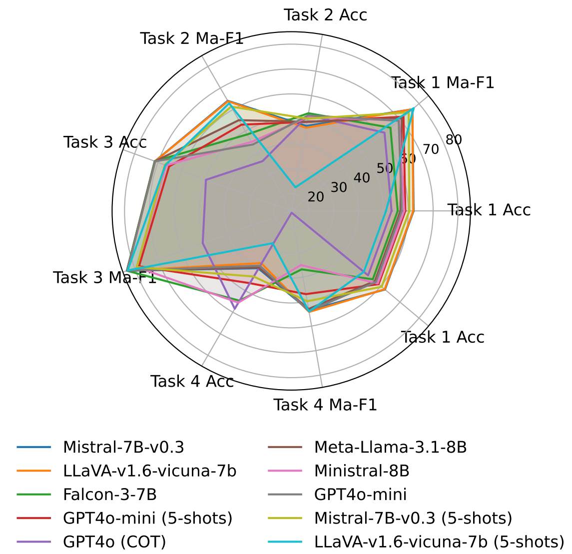

Figure 3: Radar chart of models that have the best performances in multiple tasks within each method. No single model can produce a boundary that encompasses all the data points from other models.

to image information to facilitate their question-answering and inference reasoning. However, as shown in Table 2, the highest accuracy scores of LVLMs all lag behind compared to the highest ones of LLMs. When evaluated on TASK 4 using direct zero-shot, the best LVLM outcome is even behind the best LLM by a huge gap of 11.27%. Possible issues could be the low signal-noise ratio of the images collected, and sellers usually include more comprehensive and concise features of products in text format.

No model dominates. The overall scoring result is quite close, especially after fine-tuning, where the variance between models shrinks. The open mod-

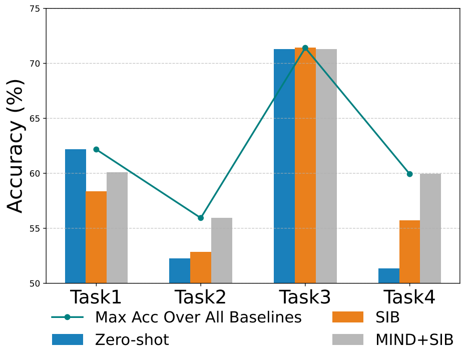

Figure 4: Comparison between best performances across different methods on different tasks. All baseline's max accuracy line is consistent with the max accuracy line over open models of all methods (i.e., Zero-shot, fine-tuning with SESSIONINTENTBENCH (SIB), and sequential fine-tuning with MIND then SIB). Paying for proprietary API does not add extra value in this case.

els that achieve the best accuracy within their category are Mistral-7B-v0.3 (zero-shot LLM), LLaVA-v1.6-vicuna-7b (zero-shot LVLM), and Falcon-3-7B (fine-tuned LLM). They belong to different companies or organizations, and none of them achieves the best accuracy in more than two tasks, as shown in Figure 3. It is an indicator that no specific model can dominate the session intention modeling game. Although the GPT-4o with Chain-of-Thought techniques gives the best Intention Evolution Modeling results, surpassing the best fine-tuned one (Mistral-Nemo-12B) with 2.59%, it might not be cost-effective due to expensive pricing and large token consumption required.

5.4 The Impact of Intention Injection

Observing Table 2, we find L(V)LMs struggling with directly leveraging intention for next product inference (*Intent-Based Inference*) and mastering long term trend of shifting intention from session history (*Intention Evolution Modeling*). However, this can be caused by many factors, such as failing to capture diverse user characteristics and preferences. Given this, we hypothesize that fine-tuning models on intention knowledge bases beforehand might enhance their ability to adapt according to different background setups and help them generalize better. Therefore, we tried to first fine-tune models on MIND (Xu et al., 2024), a large-scale intention knowledge base grounded on co-buy behaviors.

After sequentially fine-tuning first on MIND and then on SESSIONINTENTBENCH, we find a leap in performance in the proposed tasks. By injecting intention from MIND to aid SESSIONINTENTBENCH, we improved TASK 1 by 1.75%, TASK 2 by 3.09%, TASK 4 by 4.24% (another great improvement compared to zero-shot baseline), as demonstrated in Figure 4. This demonstrates that intention injection can be an effective technique to improve the model's ability to identify user intention from a short yet complex series of interactions.

5.5 Error Analyses

We randomly sample 200 tasks where GPT-4o with Chain-of-Thought commits an error. And we recruit experts to analyze the causes behind them manually. Our results show that:

- 47.5% errors are caused by incorrect understanding of the provided metadata. This may be because the model fails to incorporate past product information for deeper comprehension.
- 24% errors are caused by incorrect ground-truth labels. For objective factors, this might be due to internal conflict of session products and metadata from the intention tree or incorporating complicated metadata in the label instruction.
- 7% errors are due to models' failure to capture important product features contained in the session products, which might be aligned with or different from the metadata described in the problem assumption.
- 6.5% errors are due to irrelevant reasoning or model hallucinations, where the model is often heading towards a different reasoning direction due to some misleading, unimportant features.
- 15% the errors are due to models' inability to capturethe overall intention of the customer when the provided metadata is vague or not decisive when estimating the likelihood.

6 Conclusions

In conclusion, we propose an automated pipeline to construct a large-scale knowledge base and further construct a sample dataset SESSIONINTENTBENCH for L(V)LMs evaluations. Extensive experiments show that current models struggle to understand and infer customers' intentions while injecting intention from other knowledge bases can level up the performance. We hope our work can bridge the gap between intention understanding in simplified research cases like co-buy intention and more complex yet practical scenarios like session history. We hope this framework can benefit the community by providing better services with future models.

Limitations

We implemented our intention tree construction pipeline using GPT-4o-mini as the metadata generator. As LLM space advances, more advanced models like GPT-o1 (OpenAI, 2024b), GPT-o3-mini (OpenAI, 2025) will become more accessible to researches, which would potentially better mimic customer thinking process and behavior and generate intention and metadata in higher standard. This would enable our knowledge base and dataset generation with even higher quality.

Our current intention modeling process does not incorporate additional personalized factors such as past purchases, user characteristics, and social relationships with other customers. Incorporating these variables, which can be precomputed, could enhance model reasoning during inference, thereby providing more accurate modeling of session intent for specific customers.

The modeling setting we proposed contains multiple perspectives of session intent metadata, including attributes, intention, and comparisons. However, more metadata mined from the session can possibly be added for further knowledge integrations and better utilization of available information. More work can be done to explore what other internal factors can be incorporated within the session itself.

Ethics Statement

Offensive Content Inspection We leverage the generation capability of L(V)LMs to construct a knowledge base and carry out experiments. The generated intention at the dataset construction step is closely related to the session product information itself. The remaining metadata is based on the reasoning and comparison within products and related intentions. As the experiment setting, we only ask models to give out specific scores of likelihood or generate content with constraint reasoning, which is also closely related to sessions and products.

Annotation Wage The annotators are paid a wage in compliance with the local law, on an average of 15 USD per hour. They have all agreed to participate in annotation voluntarily.

Licenses Amazon-M2 dataset are released under the license of Apache 2.0. This grants our access to the dataset for free. Our code and data will be shared under the MIT license. It will allow the free distribution of assets we proposed and curated. All associated licenses permit user access for research purposes, and we have agreed to follow all terms of use.

References

Bruce L Alford and Abhijit Biswas. 2002. The effects of discount level, price consciousness and sale proneness on consumers' price perception and behavioral intention. *Journal of Business Research*, 55(9):775–783.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Miguel Alves Gomes, Richard Meyes, Philipp Meisen, and Tobias Meisen. 2022. Will this online shopping session succeed? predicting customer's purchase intention using embeddings. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2873–2882, New York, NY, USA. Association for Computing Machinery.

Diddigi Raghu Ram Bharadwaj, Lakshya Kumar, Saif Jawaid, and Sreekanth Vempati. 2022. Fine-grained session recommendations in e-commerce using deep reinforcement learning. *Preprint*, arXiv:2210.15451.

Minjin Choi, Hye-young Kim, Hyunsouk Cho, and Jongwuk Lee. 2024. Multi-intent-aware session-based recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2532–2536, New York, NY, USA. Association for Computing Machinery.

Honghua (Kathy) Dai, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, and Ying Li. 2006. Detecting online commercial intention (oci). In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, page 829–837, New York, NY, USA. Association for Computing Machinery.

Wenxuan Ding, Weiqi Wang, Sze Heng Douglas Kwok, Minghao Liu, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Junxian He, and Yangqiu Song. 2024. IntentionQA: A benchmark for evaluating purchase intention comprehension abilities of language models in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2247–2266, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso,

Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Derek Hao Hu, Qiang Yang, and Ying Li. 2008. An algorithm for analyzing personalized online commercial intention. In *Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '08, page 27–36, New York, NY, USA. Association for Computing Machinery.

Ravi Chandra Jammalamadaka, Naren Chittar, and Sanjay Ghatare. 2009. Mining product intention rules from transaction logs of an ecommerce portal. In *Proceedings of the 2009 International Database Engineering & Applications Symposium*, IDEAS '09, page 311–314, New York, NY, USA. Association for Computing Machinery.

Bohan Jia, Jian Cao, Shiyou Qian, Nengjun Zhu, Xin Dong, Liang Zhang, Lei Cheng, and Linjian Mo. 2023. Smone: A session-based recommendation model based on neighbor sessions with similar probabilistic intentions. *ACM Trans. Knowl. Discov. Data*, 17(8).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Luzhi Wang, Yizhen Zheng, Guojie Song, Fei Jiang, Xiang Li, Wei Lin, and Shirui Pan. 2023a. Dual intent enhanced graph neural network for session-based new item recommendation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 684–693, New York, NY, USA. Association for Computing Machinery.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, and Xianfeng Tang. 2023b. Amazon-m2: A multilingual multi-locale shopping session dataset for recommendation and text generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 755–764, New York, NY, USA. Association for Computing Machinery.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Xin Liu, Zheng Li, Yifan Gao, Jingfeng Yang, Tianyu Cao, Zhengyang Wang, Bing Yin, and Yangqiu Song. 2023b. Enhancing user intent capture in session-based recommendation with attribute patterns. *Preprint*, arXiv:2312.16199.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. *OpenAI*.

OpenAI. 2024b. Introducing openai o1. *OpenAI*.

OpenAI. 2025. Openai o3-mini. *OpenAI*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Soyeon Shim, Mary Ann Eastlick, Sherry L Lotz, and Patricia Warrington. 2001. An online prepurchase intentions model: The role of intention to search. *Journal of Retailing*, 77(3):397–416.

Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew Soon Ong. 2024. Large language models for intent-driven session recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 324–334, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A.

Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Dongjing Wang, Haojiang Yao, Dongjin Yu, Shiyu Song, He Weng, Guandong Xu, and Shuiguang Deng. 2025. Graph intention embedding neural network for tag-aware recommendation. *Neural Networks*, 184:107062.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Yu Wang, Amin Javari, Janani Balaji, Walid Shalaby, Tyler Derr, and Xiquan Cui. 2024b. Knowledge graph-based session recommendation with session-adaptive propagation. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 264–273, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Long Chen, and Yangqiu Song. 2024. MIND: Multimodal shopping intention distillation from large vision-language models for E-commerce purchase understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7800–7815, Miami, Florida, USA. Association for Computational Linguistics.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. *Preprint*, arXiv:2211.08316.

Chenwei Zhang, Wei Fan, Nan Du, and Philip S. Yu. 2016. Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 1373–1384, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Qi Zhao, Yi Zhang, Daniel Friedman, and Fangfang Tan. 2015. E-commerce recommendation with personalized promotion. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, page 219–226, New York, NY, USA. Association for Computing Machinery.

Xi Zhu, Fake Lin, Ziwei Zhao, Tong Xu, Xiangyu Zhao, Zikai Yin, Xueying Li, and Enhong Chen. 2024. Multi-behavior recommendation with personalized directed acyclic behavior graphs. *ACM Trans. Inf. Syst.*, 43(1).

Appendices

A Implementation Details

A.1 Attribute Extraction

To extract attributes from the given session products using GPT-4o-mini, we use the following 3-shot prompt as a general template:

```
Your goal is to extract the attribute type and attribute values of the product.
You will be provided with the product names and their corresponding product images, and you will output for the product:
Category: general category name of the product. Keep the category name simple and within 3 words.
Attributes: attribute(s) of the product. You can infer new ones from the image. Keep the attribute simple and within 3 words each. Separate different attributes by |. Generate in the format of attribute: value

Below are three examples:
...
Input:
Product Name: Adidas Ultraboost 21 Women's Running Shoes on sale, White/Pink special, Size 8 only, best for daily runs!
Output:
Category: Clothing
Attributes: brand: Adidas | model: Ultraboost 21 | gender: Women's | type: Running Shoes | color: White/Pink | size: 8
Input:
Product Name: Lightweight and powerful Dell XPS 13 Laptop, with newly released Intel i7, 16GB RAM, enhanced 512GB SSD, Silver version
Output:
Category: Electronics
Attributes: brand: Dell | model: XPS 13 | processor: Intel i7 | RAM: 16GB | storage: 512GB | color: Silver
Input:
Product Name: baking enthusiasts' good friend - KitchenAid Artisan Series 5-Quart Stand Mixer, Empire Red
Output:
Category: Kitchen Applianc
Attributes: brand: KitchenAid | model: Artisan Series | capacity: 5-Quart | type: Stand Mixer | color: Empire Red
...
Input:
<INPUT MESSAGE>
Output:
```

A.2 Intention Tree Construction

To construct the intention tree by filling in the necessary intention entries, pivotal attributes and underlying comparison, we use the following 5-shots template:

```
Act as a customer who is browsing a series of products.
For each input, you are required to generate several intentions as output, and each intention should only contain the following lines of information:
New Intention: new intention you may have after interacting with the new product
Attribute: attribute(s) of the new product that caused the change in intention. You can infer new ones from the image. Generate in the format of attribute: value
Rationale: a short rationale explaining why the attribute of the new product reflects the new intention. Generate in the format of facets: reasoning
Comparison: a comparison between the new product and the previous product to justify why the new product caused the change in intention. Generate in the format of aspects: comparison

Here is one example with five intentions:
...
Input:
Previous Intention: Looking for stylish and modern footwear that complements their athletic look.
Previous Product: Nike Free Metcon 5 Women's Workout Shoes (varieties: runner, target consumers: women, size: 3.5, price: $100).
New Product: LV Glove Loafer (varieties:
```

```
loafer, target consumers: men, size: 3.5, price: $200, structure: cushioned insole).
Output:
New Intention: Invest in premium quality footwear for long-lasting style and comfort.
Attribute: design: luxury material and craftsmanship.
Rationale: durability: The LV Glove Loafer is crafted from high-quality materials, offering durability and style that ensures it will last longer than ordinary shoes.
Comparison: collectability: compared to the Nike Free Metcon 5, which focuses on performance, the LV offers a blend of luxury and longevity, making it a worthy investment.

New Intention: Own a versatile pair of shoes suitable for both casual and formal settings.
Attribute: varieties: loafer.
Rationale: usages: The loafer style of the LV Glove Loafer makes it versatile enough to be worn in both casual and formal settings, unlike the more specialized athletic design of the Nike Free Metcon 5.
Comparison: versatility: While the Nike Free Metcon 5 is primarily designed for workouts, the LV Glove Loafer's loafer style offers versatility for various occasions.

New Intention: Enhance your wardrobe with a statement piece that reflects personal style.
Attribute: design: unique and luxurious.
Rationale: aesthetics: The unique and luxurious design of the LV Glove Loafer makes it a statement piece that can elevate any outfit, reflecting personal style.
Comparison: uniqueness: Unlike the more common athletic design of the Nike Free Metcon 5, the LV Glove Loafer stands out as a unique and stylish addition to the wardrobe.

New Intention: Prioritize comfort without compromising on style.
Attribute: comfort: cushioned insole.
Rationale: comfort: The cushioned insole of the LV Glove Loafer ensures comfort for long periods, making it a practical choice without compromising on style.
Comparison: comfort: While the Nike Free Metcon 5 is designed for athletic performance, the LV Glove Loafer offers a balance of comfort and style for everyday wear.

New Intention: Choose a high-end brand to reflect social status.
Attribute: brand: Louis Vuitton.
Rationale: status: Owning a product from a high-end brand like Louis Vuitton reflects social status and prestige.
Comparison: brand prestige: Compared to Nike, which is known for athletic wear, Louis Vuitton is a luxury brand that signifies higher social status.
...
Input:
Previous Intention: <Previous Intention>
Previous Product: <PREVIOUS PRODUCTS>
New Product: <THE LAST PRODUCT>
Output:
```

For smaller branching sizes, we just need to delete some of the examples provided. And for extending to more branches every step, we add additional examples as needed.

A.3 Intention Generator Model Selection

We first tried out free open LVLM models like Mantis and LLaVA families. However, the models fail to achieve the desired outcome since most of them cannot output in pre-assigned formatting. This is possible because models are not able to handle the potentially long, complex product textual descriptions and attributes provided. For example, repeatedly generating a single word or outputting a large number of special symbols like "###." After careful examination, it is not caused by prompts and product information included. We switched to GPT-4o-mini later and found the generated intention and metadata result is in the desired format, and it demonstrates comparable results with GPT-4o. Therefore, we opt for GPT-4o-mini as the major generating force for the intention tree construction part.

A.4 Few-shot Example Curation

When selecting examples for prompt templates, our primary criterion is clarity, ensuring the examples are easily understandable by large language models (LLMs). Additionally, examples must be concise, free of conflicts or ambiguities, and possess a relatively clear answer from a human perspective.

Given that small, finite discrete point masses cannot fully approximate the continuous product and intention distribution space, the chosen examples are not intended to be representative of the *entire* product distribution.

We initially generated a large set of examples randomly using GPT-4o (8 examples across 5 attempts of different categories and features, totaling 40 examples). Our researchers then manually selected, refined, and annotated these examples, followed by a thorough validation process.

Impact of Prompt Variations: To identify the optimal prompt, we experimented with multiple templates by (1) varying example categories (e.g., switching from clothing to electronic devices) and (2) increasing the number of examples. Testing on the GPT-4o API showed that these modifications resulted in minimal accuracy fluctuations (within 1%).

A.5 Model Evaluation

We evaluate the model on the tasks using different prompt techniques including zero-shot prompts (see Table 8), 5-shots prompts (see Table 9, 10, 11, 12) and Chain-of-Thought prompts (see Table 13). The detailed prompts we used can be found in the corresponding tables.

B Theory Formalization

B.1 Intention Tree

The Intention Tree $\mathbf{T}$ is defined inductively. At each discrete time step $t \in \mathbb{N}^+$, we model the branching process of the tree by extending the existing $\mathbf{T}_{1,2,3,\ldots,t-1}$ to $\mathbf{T}_{1,2,3,\ldots,t}$. We denote the information stored in the intention nodes up to time step t as $\mathcal{H}_t$, which represents the session interaction history observed up to time t.

Note that traditional models typically infer intentions based on direct transitions of products or intentions, such as $\mathbb{P}(P_{t+1}|P_t)$ or $\mathbb{P}(I_{t+1}|I_t)$.

Our model instead adopts a two-step prediction process:

$$\begin{aligned}\mathbb{P}(P_{t+1}|\mathcal{H}_t) =& \mathbb{P}(\mathcal{M}_t|\mathcal{H}_t) \cdot \mathbb{P}(P_{t+1}|\mathcal{H}_t, \mathcal{M}_t) \\ \sim& \mathbb{P}(\mathcal{M}_t^\phi|\mathcal{H}_t) \cdot \mathbb{P}(P_{t+1}|\mathcal{H}_t, \mathcal{M}_t^\phi)\end{aligned} \tag{1}$$

Here, $\mathcal{M}_t^\phi$ is the model's approximation of the session-level information $\mathcal{M}_t$ at time t.

Rather than modeling the rich information in I_t as a monolithic whole, we explicitly separate out the key components. This gives rise to the inferred elements:

- I_t^ζ: the extracted intention,
- C_t^ζ: relevant comparisons, and
- A_t^ζ: associated attributes,

which together serve as an approximation of the full information space at time step t:

$$\mathcal{M}_t^\phi = (A_t^\zeta, I_t^\zeta, C_t^\zeta)$$

From the model's perspective, the inference process is thus simplified to:

$$\mathbb{P}(P_{t+1}|\mathcal{H}_t, I_t^\zeta, C_t^\zeta, A_t^\zeta)$$

Here, the components $(A_t^\zeta, I_t^\zeta, C_t^\zeta)$ are inferred from $(P_{t+1}, \mathcal{H}_t)$, with the superscript ζ indicating a branched approximation of consumer intentions, generated by GPT-4o-mini.

As an example of how this formulation enables targeted modeling, consider the task of Valued Attribute Regularization. We first sample a subset of attribute profiles $\mathcal{A}^\pi$ from the full attribute space $\mathcal{A}$. The model is tasked with distinguishing which attributes are of crucial influence based on the provided profiles. This task is represented as a cascade of inference steps:

$$P_{t+1}, \mathcal{H}_t \mid \mathcal{A}^\pi \Rightarrow I_t^\zeta \mid P_{t+1}, \mathcal{H}_t, \mathcal{A}^\pi \Rightarrow A_t^\zeta \mid \mathcal{A}^\pi, I_t^\zeta$$

constraint to $A_t^\zeta \in \mathcal{A}^\pi$. with the constraint that $A_t^\zeta \in \mathcal{A}^\pi$.

B.2 Intuition

Consider a session with two products and a 5-branching scheme. One intention is generated for the first product, while five intentions are generated for the second product. These five intentions for the second product serve as branches stemming from the intention of the first product.

When a third product is added to the session, each intention associated with the second product becomes a new initial intention, branching out into five additional intentions linked to the third product. This process continues iteratively for subsequent products.

Each intention is uniquely associated with a set of metadata (e.g., comparison, justification, attributes) in a one-to-one correspondence.

C Task Design: Additional Clarifications

C.1 Design Criteria for Choice Options

The scores ranging from 0 to 3 in the task definition serve as symbolic representations of specific answer formulations. Concrete examples illustrating the implementation of these choices can be found in the tables on the following pages.

For example, for the first task, the score represents the likelihood estimation $\mathcal{S}_1(P_t, I_{t-1}) \mid_{\mathcal{H}_{t-1}} \in \{0, 1, 2, 3\}$, which indicates the probability of a customer interacting with P_t. A score of 3 signifies the highest likelihood, while a score of 0 represents the lowest. Specifically, as shown in Table 8 of our paper:

- Score = 3 corresponds to *A. Yes: The product is a logical and reasonable outcome of the purchasing intention.*
- Score = 2 corresponds to *B. Maybe Yes: I may consider this, but it's not a strong impulse.*
- Score = 1 corresponds to *C. Maybe No: The product is not directly related to my intention.*
- Score = 0 corresponds to *D. No: I would never purchase it if I were the customer with the given intention.*

We hope this explanation resolves any doubts regarding the formulation of the scores. For the ground truth label, we combine "Maybe Yes" (B) with "Yes" (A) and "Maybe No" (C) with "No" (D) to create a clearer decision boundary. This approach aligns with common practices in survey analysis and simplifies the classification problem into two categories. The design of the ground truth label *accounts for neutral response bias to better reflect the true distribution of opinions*. Otherwise, the overrepresentation of neutral responses can lead to unexpected model bias. By grouping the responses, the problem is reduced to a binary classification setting (e.g., approve/reject or positive/negative). This not only simplifies analysis but also makes the results more actionable for the model training and fine-tuning.

For Task 4, we designate a score of 0 as the true positive label, as it represents "firm preferences for products within the same category and feature set." The other two labels are grouped together, as they reflect the customer's inclination to explore products from different features or categories. This results in a binary label design for Task 4, distinct from the label structures used in Tasks 1–3.

For further details, please kindly refer to Table 8, where we outline the three-choice design for Task 4 in contrast to Tasks 1–3.

D Annotation Details

D.1 Worker Selection Protocol

We carry out strict quality control to ensure high quality human annotation result. To start with, we send qualification round invitations only to workers who satisfy the following constraints: (i) pass over 2,000 HITs, (ii) score over 90% on historical approval rate. We curate a qualification test with sampled sessions, whose gold label are provided by the authors. We first retain those with an accuracy of over 75% on the qualification test and complete over 20 questions at the same time.

Following that, we disqualify those spammers or underperforming workers. More specifically, we filter out those workers who are simply picking one side of the choices for the majority of the time. After conducting another round of testing, results show that least 7 people picking one side of the answer 80 percent of the time, so we eliminate out those people and proceed to main round annotations.

11 workers left after the selection process out of 300 initial candidates. This gives a worker selection rate of 3.67%.

D.2 Annotation Instructions

We give instruction to workers in layman's terms, with both detailed question definitions and specific explanations for important information included. We tried to include more information and less distractions. The question definitions are closely aligned to what we defined earlier in Section 3. For each of the first three questions, workers are asked to annotate using a four-point 0 to 3 likelihood scale, where 0 stands for the least probable or the least plausible, and 3 means the most likely or the most plausible. For the forth question, annotation results are constraint to a three point scale, from 1 to 3. Larger the number, larger the likelihood of exploring more diverse products.

E Annotation Result Analysis

E.1 Raw Label Result

The distribution of labels is reflected in the *Majority* score presented in the Table 2. We have also summarized the raw label distribution below. Instead of reporting individual choices (e.g., A or B), we grouped them to mitigate individual annotator biases observed during the annotation process, where some annotators consistently favored extreme answers while others preferred intermediate ones.

Task_Ind	Label	Count	Percentage
1	A_B	5844	62.30%
1	C_D	3536	37.70%
2	A_B	4282	45.65%
2	C_D	5098	54.35%
3	A_B	6735	71.80%
3	C_D	2645	28.20%
4	A	3456	36.84%
4	B_C	5924	63.16%

Table 4: Summarized counts and percentages of question labels. *Task_Ind* denotes the task index, ranging from 1 to 4. A_B indicates whether option A or B was selected, and similarly, C_D represents whether option C or D was chosen.

E.2 Consistency

We employ a majority vote rule to establish ground truth. In Table 5, the notations “2:1” and “3:0” indicate the label distribution for each question. A “3:0” label means all three annotators selected the same answer, while “2:1” indicates that one annotator’s choice differed from the other two (e.g., annotators 1 and 2 selected A or B, while annotator 3 selected C or D). This distribution shows that for over 50% of the questions, human annotators consistently agreed on the answer. Thus, these results further suggest that current models lag behind human performance in understanding these tasks.

Task_Ind	2:1	3:0
1	6041	7959
2	9170	4830
3	5390	8610
4	3934	10066

Table 5: Consistency analysis of binary answer label distribution. *Task_Ind* denotes the task index, ranging from 1 to 4.

E.3 Annotation Quality Filter

In addition to dataset-level analysis, we conducted a detailed examination of individual annotator statistics. For instance, during the annotator filtering process, we compiled Table 6 to summarize individual label distribution.

During the annotator filtering process, we identified individuals who consistently favored one set of choices (e.g., always selecting A/B or C/D). Such behavior may indicate a lack of engagement, where annotators select options indiscriminately to complete the task and receive payment. To ensure quality, we excluded these annotators in the official annotation round by administering a separate qualification test and filtering out those exhibiting this pattern.

Annotator_ID	A	B	C	D
A1***1A	3201	719	893	31
A2***EZ	3402	1208	437	1
A2***2M	106	5540	1950	48
A1***SU	633	186	135	18
A3***TX	2113	173	610	28
A2***BO	287	32	49	0
A2***YO	919	221	196	24
A2***E0	466	129	140	5
AF***9P	60	23	14	3

Table 6: Exclude annotators who consistently select the same option the majority of the time.

E.4 Benchmark and Data Quality Validation

As noted previously, incorrect labels may arise from objective factors, such as internal conflicts within session products and metadata in the intention tree or the inclusion of complex metadata in the labeling instructions. **These issues stem from customer behaviors**, which can exhibit self-conflicting patterns or random product-to-product jumps, introducing inherent randomness that is challenging to eliminate. However, one can expect that preprocessing to identify these patterns, though effort-intensive, can mitigate such issues and enhance benchmark quality.

F Further Notes on Model Performances

F.1 Evaluation Task Performances Metrics

We display our summarized task performance metrics for each of the tasks in Table 7. Statistics in both counting and percentage format are included.

	Metric	Task No.			
		Task 1	Task 2	Task 3	Task 4
Count	# TP	781	415	527	57
	# FN	354	434	775	585
	# TN	234	515	309	949
	# FP	458	445	215	131
Percentage	TP (%)	42.75%	22.94%	28.86%	3.31%
	FN (%)	19.38%	24.00%	42.44%	33.97%
	TN (%)	12.81%	28.47%	16.92%	55.11%
	FP (%)	25.07%	24.60%	11.77%	7.61%

Table 7: Task performance metrics for error analyses of GPT-4o with Chain-of-Thought answering SessionIntentBench. Where *TP*, *FN*, *TN*, *FP* stand for true positive, false negative, true negative, and false positive answers respectively.

F.2 Finetuning

Our analysis suggests that models fine-tuned solely on the Session Intention Benchmark (SIB) may underperform due to the diverse and widely dispersed distribution of SIB data. This diversity creates a significant gap between the training and test sets, impacting model generalization.

For instance, some models exhibit degraded performance when fine-tuned only on SIB for certain tasks. However, incorporating external intention injection (e.g., fine-tuning on SIB combined with MIND, an external intention knowledge base) improves outcomes. As an example, Llama-3.1-8B achieved 56.78% accuracy on Task 1 in a zero-shot setting, which dropped to 52.82% after fine-tuning on SIB alone but increased to 60.10% when fine-tuned on both SIB and MIND. This improvement likely results from the general intention knowledge provided by the external source, which mitigates distribution dispersion and enhances the model's ability to generalize across the training-test gap.

F.3 The BERT-based Models

We evaluated pretrained BERT-based models like RoBERTa-large-355M and DeBERTa-v3-large, but found them unsuitable for testing in our context.

For RoBERTa-large-355M, the raw output is exemplified as follows:


```
["task_counter": 25248,
"session_counter": 6311,
"question_idx": 3, "response":
"**A**", "task_counter": 27563,
"session_counter": 6890,
"question_idx": 2, "response":
"**Yes**", "task_counter": 2654,
"session_counter": 663, "question_idx":
1, "response": "**A**", "task_counter":
16969, "session_counter": 4242,
"question_idx": 0, "response":
"**A**", "task_counter": 33507,
"session_counter": 8376,
"question_idx": 2, "response":
"**Yes**", ...]
```


Upon analyzing the output, we observed that RoBERTa-large-355M consistently produces responses such as “A” or “Yes” regardless of the question, failing to align with the query's requirements. This behavior is not observed in larger models. We hypothesize that smaller models like RoBERTa-large-355M struggle with complex tasks, often defaulting to predicting the most frequent answer in the masked space based on the nearest question description, without effectively processing the provided session product information, let alone metadata restrictions.

For DeBERTa-v3-large, the raw output is entirely consisted of random strings such as "IBILITY" and "Measurement" instead of providing answers to the questions:


```
["task_counter": 25248,
"session_counter": 6311,
"question_idx": 3, "response":
"**IBILITY**", "task_counter":
27563, "session_counter": 6890,
"question_idx": 2, "response":
"**IBILITY**", "task_counter":
2654, "session_counter": 663,
"question_idx": 1, "response":
"**Measurement**", "task_counter":
16969, "session_counter": 4242,
"question_idx": 0, "response":
"**IBILITY**", "task_counter":
33507, "session_counter": 8376,
"question_idx": 2, "response":
"**IBILITY**", ...]
```


Despite experimenting with various prompting techniques, the model consistently failed to process the queries effectively. This suggests that DeBERTa-v3-large, due to its limited capacity, struggles with the complexity of these tasks. Similar issues were encountered when using Mantis family models as large vision-language models (LVLMs) for intention generation, where outputs often exhibited broken structures or degenerated into pure noise.

Based on the observed limitations, we believe that models such as RoBERTa-large-355M and DeBERTa-v3-large should not be included in the current evaluation phase due to their inability to effectively process the complex tasks in our framework.

Survey Instructions (Click to Collapse)

How Intentions Evolve with Changing Attributes?

Welcome to our Main Round HITs. Congratulations on passing the qualification test and thanks for participating in our HITs!

In this survey, you will be provided a session of products and asked to evaluate alterations in purchasing intentions as the product attributes changes.

Before the questions: You will be provided with a list of Session Products that will be used throughout the questions.

Answer each question: Select the option that best describes your evaluation of the model's output based on the criteria provided.

Question Formalization

Q1: Changing Intentions

After reviewing the listed products (including their titles, attributes, images, etc.), and assuming you have the provided purchasing intention, we want to understand how likely you are to purchase a specific product based on this intention. Your task is to decide whether you would consider purchasing the product given your current intentions.

You'll be provided with four rating options: Yes, Maybe yes, Maybe no and No.

Q2: Attribute that Matters

After reviewing the listed products, and assuming you highly value a specific attribute of the listed products, we want to understand how likely you are to purchase another product based on this valued attribute. Your task is to decide whether you would consider purchasing the product given your focus on the specific characteristic.

You'll be provided with four rating options: Yes, Maybe yes, Maybe no and No.

Q3: Comparisons

After reviewing the listed products, and assuming you have the provided purchasing intentions, we want to understand if the comparison between the products provides a detailed and reasonable justification for your purchasing impulse. Your task is to decide whether the comparison is thorough enough to justify your change in intention.

You'll be provided with four rating options: Yes, Maybe yes, Maybe no and No.

Q4: Changing Desire

After reviewing the listed products, and assuming you have the provided purchasing intention, we want to understand if you still wish to explore similar products. Your task is to decide whether you want to continue exploring products within the same category or look for products in different categories.

You'll be provided with three rating options: Yes, Maybe yes, and No.

Session Products List

Session Products List is a list of products that you browsed (possibly consider purchasing) in a short period of time on Amazon.

The list of products will contain the following information:

- (1) **Product title:** The name of the product you viewed.
- (2) **New intention:** You should imagine yourself as a customer who has the mentioned intention/impulse when browsing the products. The word "New" means it's the intention you hypothetically have after seeing the last product in the current list.
- (3) **Attributes:** The features, functions, or characteristics of the product that you may consider when making a purchase decision. They are complementary information for the title/image to facilitate your decision process.

Each Session Products List is in one-to-one correspondence with the question following it.

Additional Hints

- Read the Session Products List carefully: Understand the previous intention, previous product, and new product details.
- Submit your response: Once you have answered all questions, click the Submit button to complete the HIT.

Figure 5: The annotation instruction we shown to workers, with detailed question definitions in layman's terms and specific explanations for important information (e.g., a preview of information contained in *Session Product List*).

Task	Zero-shot Prompt
TASK 1	Act as a customer who is browsing a series of products given as follows. `<session product information>` After seeing `<previous products>`, and assuming you are a customer who has the intention of `<second last intention>`. How likely are you to purchase `<last product>` based on the assumed intention? A. Yes: The product is a logical and reasonable outcome of the purchasing intention. B. Maybe yes: I may consider this, but it's not a strong impulse. C. Maybe no: The product is not directly related to my intention. D. No: I would never purchase it if I were the customer with the given intention. Your Answer (Answer A or B or C or D only):
TASK 2	Act as a customer who is browsing a series of products given as follows. `<session product information>` After seeing `<previous products>`, and assuming you are a customer who highly value the feature `<second last intention attribute>` of `<second last product>`. How likely are you to purchase `<last product>`? A. Yes: The product logically and reasonably matches the characteristics I value. B. Maybe yes: I might consider this product, but it doesn't strongly appeal to me. C. Maybe no: The product does not directly relate to the characteristic I value. D. No: I would not purchase this product if I were focused on the given characteristic. Your Answer (Answer A or B or C or D only):
TASK 3	Act as a customer who is browsing a series of products given as follows. `<session product information>` Comparing between `<last two products>`, and assuming you have the intention of `<last two intention>`, Does this comparison `<last intention comparison>` provide an in-depth justification of your impulse? A. Yes: the comparison is reasonable and detailed enough to justify the change. B. Maybe yes: The comparison could be more detailed and thorough but can be ignored. C. Maybe no: The comparison is not entirely reasonable or lacks sufficient in-depth detail. D. No: The comparison does not provide any underlying reasons or insights. Your Answer (Answer A or B or C or D only):
TASK 4	Act as a customer who is browsing a series of products given as follows. `<session product information>` After seeing `<previous products>`, and assuming you have the intention of `<previous intention>`, do you still want to explore similar products? A. Yes: I want to explore products under the same category. B. Maybe yes: I want to explore products under the same category but with different features. C. No: I want to explore products under other categories. Your Answer (Answer A or B or C only):

Table 8: Zero-shot prompts for model evaluation. TASK 1 stands for ***Intent-Based Purchasing Likelihood Estimation***, TASK 2 stands for ***Purchasing Likelihood Inference via Valued Attributes Regularization***, TASK 3 stands for ***Intention Justification via Comparison***, TASK 4 stands for ***Intention Evolution Modeling***.

Task	5-shots Prompt
TASK 1	Act as a customer who is browsing a series of products given as follows. <session product information> You hold an assumed intention, which will be provided later. After seeing the products, you will be asked to determine the likelihood of purchasing the last product \ based on the assumed intention. You will be given four options to choose from: Yes, Maybe yes, Maybe no, No. Please select the most appropriate option based on the given context. A. Yes: The product is a logical and reasonable outcome of the purchasing intention. B. Maybe yes: I may consider this, but it's not a strong impulse. C. Maybe no: The product is not directly related to my intention. D. No: I would never purchase it if I were the customer with the given intention. Here are a few examples: Q: After seeing Eco-friendly laundry detergent, bamboo dish brush, reusable kitchen cloths, and assuming you are a customer who have the intention of Reducing household chemical usage. How likely are you to purchase A biodegradable dish soap based on the assumed intention? A: A. Yes Q: After seeing Instant Pot, KitchenAid Stand Mixer, Ninja Air Fryer, and assuming you are a customer who have the intention of Upgrading kitchen equipment for home cooking. How likely are you to purchase A set of gourmet spices based on the assumed intention? A: C. Maybe no Q: After seeing Columbia hiking boots, North Face backpack, Garmin GPS watch, and assuming you are a customer who have the intention of Planning for outdoor adventures. How likely are you to purchase A formal suit for weddings based on the assumed intention? A: D. No Q: After seeing "1984" by George Orwell, "To Kill a Mockingbird" by Harper Lee, \ "The Catcher in the Rye" by J.D. Salinger, and assuming you are a customer who have the intention of Finding new reading material for leisure. How likely are you to purchase "The Da Vinci Code" by Dan Brown based on the assumed intention? A: B. Maybe yes Q: After seeing Rolex Submariner, Omega Seamaster, Tag Heuer Monaco, and assuming you are a customer who have the intention of Finding a timeless gift for a special occasion. How likely are you to purchase A limited edition Patek Philippe watch based on the assumed intention? A: A. Yes Q: After seeing <previous products>, and assuming you are a customer who have the intention of <second last intention>. How likely are you to purchase <last product> based on the assumed intention? A:

Table 9: 5-shots prompts for model evaluation. TASK 1 stands for ***Intent-Based Purchasing Likelihood Estimation***

Task	5-shots Prompt
TASK 2	Act as a customer who is browsing a series of products given as follows. `<session product information>` You have a valued feature/attribute, which will be provided later. After seeing the products, you will be asked to determine the likelihood of purchasing the last product \ based on the valued attribute. You will be given four options to choose from: Yes, Maybe yes, Maybe no, No. Please select the most appropriate option based on the given context. A. Yes: The product logically and reasonably matches the characteristics I value. B. Maybe yes: I might consider this product, but it doesn't strongly appeal to me. C. Maybe no: The product does not directly relate to the characteristic I value. D. No: I would not purchase this product if I were focused on the given characteristic. Here are a few examples: Q: After seeing Noise-canceling headphones, wireless earbuds, Bluetooth speaker, and assuming you are a customer who highly value the feature High audio quality of Bluetooth speaker. How likely are you to purchase A premium soundbar? A: A. Yes Q: After seeing adjustable standing desk, monitor with blue light filter, Ergonomic office chair, and assuming you are a customer who highly value the feature Ergonomics of Ergonomic office chair. How likely are you to purchase A desk lamp with a USB port? A: C. Maybe no Q: After seeing Organic facial cleanser, natural moisturizer, chemical-free sunscreen, and assuming you are a customer who highly value the feature Natural ingredients of chemical-free sunscreen. How likely are you to purchase A synthetic fragrance? A: D. No Q: After seeing DSLR camera, camera tripod, external flash, and assuming you are a customer who highly value the feature Professional photography of external flash. How likely are you to purchase A photo editing software? A: A. Yes Q: After seeing High SPF sunscreen, UV-blocking sunglasses, wide-brimmed hat, and assuming you are a customer who highly value the feature Sun protection of wide-brimmed hat. How likely are you to purchase An aloe vera gel? A: B. Maybe yes Q: After seeing `<previous products>`, and assuming you are a customer who highly value the feature `<second last intention attribute>` \ of `<second last product>`. How likely are you to purchase `<last product>`? A:

Table 10: 5-shots prompts for model evaluation. TASK 2 stands for ***Purchasing Likelihood Inference via Valued Attributes Regularization***.

Task	5-shots Prompt
TASK 3	Act as a customer who is browsing a series of products given as follows. `<session product information>` You have an assumed intention, which will be provided later. You will be asked to evaluate the provided comparison between the last two products \\ based on the assumed intention. You will be given four options to choose from: Yes, Maybe yes, Maybe no, No. Please select the most appropriate option based on the given context. A. Yes: the comparison is reasonable and detailed enough to justify the change. B. Maybe yes: The comparison could be more detailed and thorough but can be ignored. C. Maybe no: The comparison is not entirely reasonable or lacks sufficient in-depth detail. D. No: The comparison does not provide any underlying reasons or insights. Here are a few examples: Q: Comparing between a budget smartphone with a long battery life and A high-end smartphone with \\ superior low-light performance, and assuming you have the intention of Finding a device with the best camera quality, Does this comparison The high-end smartphone boasts advanced camera technology \\ provide in-depth justification of your impulse? A: A. Yes Q: Comparing between A compact car and a mid-size SUV, and assuming you have the intention of Prioritizing fuel efficiency, Does this comparison the mid-size SUV, although spacious, consumes more fuel due to its larger engine \\ and heavier body provide in-depth justification of your impulse? A: B. Maybe yes Q: Comparing between A luxury wristwatch and a fitness tracker, and assuming you have the intention of Tracking health metrics, Does this comparison Finding a more affordable watch provide in-depth justification of your impulse? A: D. No Q: Comparing between A leather office chair with plush cushioning and \\ a mesh office chair with lumbar support and assuming you have the intention of Seeking maximum comfort during long working hours, Does this comparison The mesh office chair offers better breathability and ergonomic support \\ provide in-depth justification of your impulse? A: A. Yes Q: Comparing between A hardcover book and an e-reader, and assuming you have the intention of Enhancing the reading experience, Does this comparison The hardcover book provides a tactile, while the e-reader offers portability, \\ adjustable text size provide in-depth justification of your impulse? A: C. Maybe no Q: Comparing between `<last two products>`, and assuming you have the intention of `<last two intention>`, Does this comparison `<last intention comparison>` provide in-depth justification of your impulse? A:

Table 11: 5-shots prompts for model evaluation. TASK 3 stands for ***Intention Justification via Comparison***.

Task	5-shots Prompt
TASK 4	Act as a customer who is browsing a series of products given as follows. `<session product information>` You will be provided with a sequence of intention. You will be asked to determine whether you still want to explore similar products \\ based on the sequence of intention. You will be given three options to choose from: Yes, Maybe yes, No. Please select the most appropriate option based on the given context. A. Yes: I want to explore products under the same category. B. Maybe yes: I want to explore products under the same category but with different features. C. No: I want to explore products under other categories. Here are a few examples: Q: After seeing Stainless steel kitchen knives, non-stick frying pans, silicone spatulas, and assuming you have the intention of Upgrading kitchen tools for home cooking, do you still want to explore similar products? A: B. Maybe yes Q: After seeing Fitness tracker, yoga mat, resistance bands, and assuming you have the intention of Tracking fitness progress, do you still want to explore similar products? A: A. Yes Q: After seeing Stainless steel refrigerator, smart oven, induction cooktop, and assuming you have the intention of Making the kitchen more energy efficient, do you still want to explore similar products? A: C. No Q: After seeing Smart thermostat, LED light bulbs, energy-efficient washing machine, and assuming you have the intention of Saving on utility bills, do you still want to explore similar products? A: B. Maybe yes Q: After seeing Indoor plants, plant stands, watering can, and assuming you have the intention of Creating a greener living space, do you still want to explore similar products? A: A. Yes Q: After seeing `<previous products>`, and assuming you have the intention of `<previous new intention>`, do you still want to explore similar products? A:

Table 12: 5-shots prompts for model evaluation. TASK 4 stands for ***Intention Evolution Modeling***.

Task	Chain-of-Thought Prompt
Task 1	Act as a customer who is browsing a series of products given as follows. `<session product information>` After seeing `<previous products>`, and assuming you are a customer who have the intention of `<second last intention>`. How likely are you to purchase `<last product>` based on the assumed intention? A. Yes: The product is a logical and reasonable outcome of the purchasing intention. B. Maybe yes: I may consider this, but it's not a strong impulse. C. Maybe no: The product is not directly related to my intention. D. No: I would never purchase it if I were the customer with the given intention. Answer with a brief rationale then make your final choice \\ by answering the option alphabet A/B/C/D only in the last line of your response. Your Answer:
Task 2	Act as a customer who is browsing a series of products given as follows. `<session product information>` After seeing `<previous products>`, and assuming you are a customer who highly value the feature `<second last intention attribute>` \\ of `<second last product>`. How likely are you to purchase `<last product>`? A. Yes: The product logically and reasonably matches the characteristic I value. B. Maybe yes: I might consider this product, but it doesn't strongly appeal to me. C. Maybe no: The product does not directly relate to the characteristic I value. D. No: I would not purchase this product if I were focused on the given characteristic. Answer with a brief rationale then make your final choice \\ by answering the option alphabet A/B/C/D only in the last line of your response. Your Answer:
Task 3	Act as a customer who is browsing a series of products given as follows. `<session product information>` Comparing between `<last two products>`, and assuming you have the intention of `<last two new intention>`, Does this comparison `<last intention comparison>` provide in-depth justification of your impulse? A. Yes: the comparison is reasonable and detailed enough to justify the change. B. Maybe yes: The comparison could be more detailed and thorough but can be ignored. C. Maybe no: The comparison is not entirely reasonable or lacks sufficient in-depth detail. D. No: The comparison does not provide any underlying reasons or insights. Answer with a brief rationale then make your final choice \\ by answering the option alphabet A/B/C/D only in the last line of your response. Your Answer:
Task 4	Act as a customer who is browsing a series of products given as follows. `<session product information>` After seeing `<previous products>`, and assuming you have the intention of `<previous new intention>`, do you still want to explore similar products? A. Yes: I want to explore products under the same category. B. Maybe yes: I want to explore products under the same category but with different features. C. No: I want to explore products under other categories. Answer with a brief rationale, then make your final choice \\ by answering the option alphabet A/B/C only in the last line of your response. Your Answer:

Table 13: Chain-of-Thought prompts for model evaluation. Task 1 stands for ***Intent-Based Purchasing Likelihood Estimation***, Task 2 stands for ***Purchasing Likelihood Inference via Valued Attributes Regularization***, Task 3 stands for ***Intention Justification via Comparison***, Task 4 stands for ***Intention Evolution Modeling***.