SESSIONINTENTBENCH: A Multi-task Inter-session Intention-shift Modeling Benchmark for E-commerce Customer Behavior Understanding

Yuqi Yang[*♠], Weiqi Wang[*♠♣†], Baixuan Xu[♠], Wei Fan[♠], Qing Zong[♠], Chunkit Chan[♠], Zheye Deng[♠], Xin Liu[♣], Yifan Gao[♣], Changlong Yu[♣], Chen Luo[♣], Yang Li[♣], Zheng Li[♣], Qingyu Yin[♣], Bing Yin[♣], Yangqiu Song[♠♣‡]

[♠]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China

[♣]Amazon.com Inc, Palo Alto, CA, USA

yyangfd@connect.ust.hk; wwangbw@cse.ust.hk; yqsong@cse.ust.hk

Abstract

Session history is a common way of recording user interaction behaviors throughout a browsing activity involving multiple products. For example, if a user clicks on a product webpage and then leaves, it might be because certain features do not satisfy the user, which serves as an important indicator of on-the-spot user preferences. However, prior works fail to capture and model customer intention effectively because of insufficient information exploitation, relying only on apparent information such as descriptions and titles. There is also a lack of data and corresponding benchmarks 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 to understand inter-session intention shifts through four subtasks. With 1,952,177 intention entries, 1,132,145 session intention trajectories, and 13,003,664 available tasks mined from 10,905 sessions, we provide a scalable way to exploit existing session data for customer intention understanding. We conduct human annotations to collect ground-truth labels for a subset of the 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 intention across complex session settings. Further analysis shows that injecting intention enhances LLM performance.

Figure 1: An example of intention shift within a shopping session. As the user interacts with successive products, latent preferences become more specific and can change from one step to the next.

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

[*]Equal Contribution

[†]Work done during his internship at Amazon.com Inc.

[‡]Visiting academic scholar at Amazon.com Inc.

Findings of the Association for Computational Linguistics: ACL 2026, pages 16748–16775

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.

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; Bai et al., 2026). 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.

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., 2025a) and co-buy behaviors (Yu et al., 2023; Xu et al., 2024; Wang et al., 2025b). 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. 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., 2024c). 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

We use *intention shift* to refer to the step-to-step evolution of user preference across successive interactions in a single shopping session. In the rest of the paper, we use *intention* and *intent* interchangeably. Figure 2 summarizes our formulation.

We propose to model the intention shift from four aspects, 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.

Formally, assume a session contains products $P_1, P_2, \ldots, P_T$ observed in chronological order. Let A_t denote the valued attribute associated with the transition at step t, I_t the inferred user intention after interacting with P_t, and C_t a comparison between adjacent products that helps justify the shift from the previous state to the current one. The interaction history up to time step t is

$$\mathcal{H}_t = \{(P_j, A_j)\}_{j=1}^{t}.$$

Our goal is to evaluate whether a model can use $\mathcal{H}_t$ and the mined intention metadata to reason about future interactions and the trajectory of the session.

TASK 1: Intent-Based Purchasing Likelihood Estimation. The first task asks the model to verify whether the last proposed intention is well aligned with the new product we are going to interact with. The model will be given historical information $\mathcal{H}_{t-1}$, the proposed intention I_{t-1}, and the 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 essential elements of the actual unseen product. The model is provided with historical 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 guard against potential hallucinations, the third task asks the model to justify whether the proposed C_t provides a reasonable explanation 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 propose aims to test the model's ability to help 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 a different category (exploring further to infer user preferences). If we map the choices to the numerical scores $\{1, 2, 3\}$, 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.

Tasks 1–3 are annotated on a four-point ordered scale, while Task 4 uses a three-way exploration choice. For evaluation, these raw annotations are later mapped to binary labels; we describe this mapping in Section 5 and Appendix C.

3.2 Dataset

We construct SESSIONINTENTBENCH from Amazon-M2 (Jin et al., 2023b) and product images retrieved from the Amazon Review Dataset (Hou et al., 2024). Amazon-M2 provides session sequences and rich textual metadata such as product titles, prices, colors, and materials. We align those products with their corresponding images from the Amazon Review Dataset to obtain multimodal session records. After filtering out products with missing or inaccessible image links, we retain 10,905 sessions with complete textual and visual information.

4 SESSIONINTENTBENCH Construction

We construct SESSIONINTENTBENCH by enriching raw shopping sessions with explicit intention metadata. The pipeline in Figure 2 has four stages: (i) multimodal attribute extraction for each product, (ii) intention generation over the session timeline, (iii) metadata analysis of why intention shifts from one step to the next, and (iv) human annotation of a sampled subset for evaluation.

4.1 Multi-modal Attribute Extraction

The first stage extracts product attributes that can support later intention reasoning. We use GPT-4o-mini (OpenAI, 2024) as the extraction model and provide both textual product descriptions and product images. The model outputs a coarse product category together with normalized attribute–value pairs, for example *color: white* or *size: 7.5 inches*. This step standardizes heterogeneous product metadata before any session-level reasoning is introduced.

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 intentions 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 intentions $< I_{t1}, I_{t2}, I_{t3}, \dots > \big|_{t=1}$ based on the textual and visual information of the product the user interacted with, where the prompt is demonstrated below. Then, we repeat the inference at every step as we add the next new session product into the visible list of items for the model.

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's reasoning. At each time we perform the inference, we only use one intention chosen from the previous step's intentions to ensure coherent intention trajectory sampling. More specifically, the model is constrained to output the five most possible user intentions, denoted as $\{$<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. 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 |<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:>
```

4.3 Intention-Shift Metadata 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 customer's further decision-making. The prompt we used for generation is given above. To ground the reasoning in the actual product metadata, we require the model to point out the most likely feature <Attr> A_t that affects the user's choices. Furthermore, we ask for a more

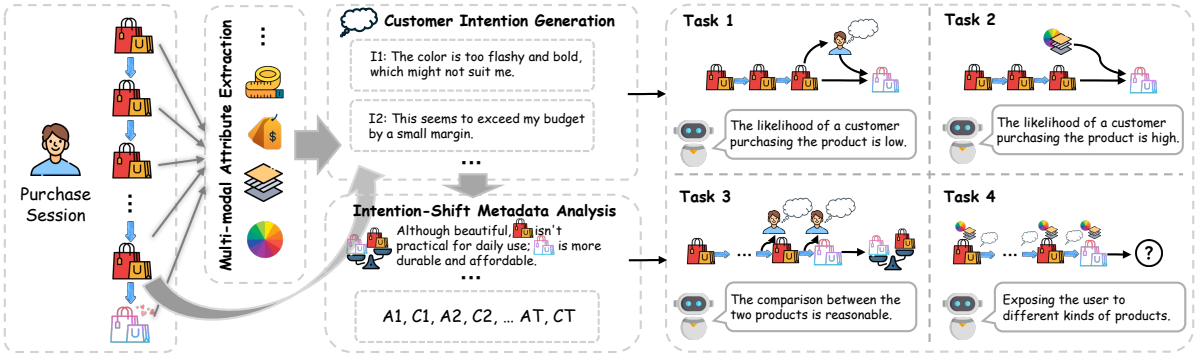

Figure 2: Overview of SESSIONINTENTBENCH and the curation pipeline. We first extract multimodal product attributes, then generate candidate intentions and their branches, and finally enrich each branch with attribute and comparison metadata for the four downstream tasks. Here, A_i and C_i denote the valued attribute and comparison metadata at step i.

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 human-annotated subset used for SESSIONINTENTBENCH. *uni.* denotes unique sessions. Sampling is performed over intention trajectories (and therefore task instances), not directly over raw sessions.

comprehensive comparison <Comp> C_t between the last product P_t and the previous one P_{t-1}, so that it provides logical support for the modeled intention pathways. 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 with emphasis on the following perspectives: (1) the alignment of the proposed intention I_t and session products P_{t+1}; (2) the consistency between the inferred valued attribute A_t and the 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* (corresponding to $\mathcal{S} = 3, 2, 1, 0$). We carried out multiple rounds of annotation worker selection with different criteria to ensure high annotation quality. We further analyze label distributions and worker behavior in Appendix D. Task 2 remains the most subjective task because it requires workers to judge whether a proposed valued attribute actually drives the next interaction, which is one reason we report detailed annotation analyses in the appendix.

5 Evaluations and Analyses

5.1 Intrinsic Evaluations

We present our detailed statistics in Table 1. By filling up the tree with intentions across 10,905 sessions, we obtain more than 1,950,000 intention entries and 1,100,000 intention trajectories. The majority of these sessions contain fewer 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 of three to five. We then sample 2,000 sessions with 2 trajectories per session and later add another disjoint 1,445

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
Meta-Llama-3.2-3B	54.68	63.97	52.02	43.48	33.13	49.48	51.34	36.61
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 different L(V)LMs on the human-annotated SESSIONINTENTBENCH test set. Random samples labels uniformly. Majority always predicts the most frequent label for each task. All "Few-shots" results use 5 demonstrations. Within each block, the best scores are underlined; the best scores overall are **bold**.

sessions with 4 trajectories per session. This gives 9,780 trajectories in total. To grant the model full information availability, we only query the tasks at the end of each session time step, that is, using all the available products and masking the last product when querying TASK 1 and 2.

5.2 Baselines and Model Selections

Evaluation protocol. We report accuracy and Macro-F1. Although the raw annotations are ordinal, we evaluate in a binary setting to reduce neutral-response bias and to align the tasks with practical accept/reject-style decision making. For Tasks 1–3, answers A/B are treated as positive and C/D as negative. For Task 4, answer A (continue exploiting similar products) is treated as positive and B/C as negative. We include two simple baselines: RANDOM, which samples labels uniformly, and MAJORITY, which always predicts the globally most frequent label for a task. The Macro-F1 values for Majority on Tasks 2 and 4 are undefined because the majority label is negative, so the baseline never predicts a positive instance.

Model families. We evaluate four groups of models. **(i) Open zero-shot models:** open LLMs and LVLMs from the Llama (Grattafiori et al., 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), Falcon (Almazrouei et al., 2023), Qwen (Qwen et al., 2025), LLaVA (Liu et al., 2023a), and Qwen-VL (Wang et al., 2024a)

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: Sequential finetuning with MIND followed by SESSIONINTENTBENCH. "MIND + SIB" means that a backbone is first fine-tuned on MIND and then fine-tuned on SESSIONINTENTBENCH (abbreviated as SIB). Within each training setting, the best scores are underlined; the best scores overall are **bold**.

families. **(ii) Few-shot models:** selected open L(V)LMs evaluated with 5 in-context demonstrations. **(iii) Fine-tuned open models:** representative sub-11B backbones from different model families, fine-tuned on SIB with supervised fine-tuning (SFT) and LoRA using LLaMA-Factory. **(iv) Proprietary APIs:** GPT-4o and GPT-4o-mini (OpenAI et al., 2024; OpenAI, 2024) under zero-shot, 5-shot, and Chain-of-Thought prompting (Wei et al., 2023). The detailed split strategy, prompt construction, and fine-tuning setup are described in Appendix A. Table 3 additionally reports **MIND + SIB**, which means sequential fine-tuning on MIND first and then on SIB.

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*), on which models scored 49.63%, there is a large 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 noting that GPT-4o with Chain-of-Thought prompting is able to achieve the highest rate of 58.42% among all models and methods. This might be because the larger model size and the technique of enabling reasoning at run time could help the model better mimic the thinking process of a real-life customer. This result shows that more work needs to be done to improve the model's capability to capture long-term user intention trends.

Fine-tuning can greatly improve poorly performing models, but struggles to help mediocre ones. Poorly performing models, which we refer to as those that receive a low score compared to models under the same category in some evaluation tasks, can quickly acquire 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 performance increase of 36.5% and demonstrates outcomes comparable with other larger 7B or 8B models. Mediocre performing models, which we refer to as those that score near the highest among the models but still struggle to surpass the top accuracy records, benefit less from fine-tuning. Among the proposed tasks, the largest maximum accuracy increase from zero-shot to fine-tuned occurs in 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 Appendix 6.2 for additional discussions on fine-tuning.

LVLMs struggle to make good use of visual signals. In comparison to LLMs, which only use

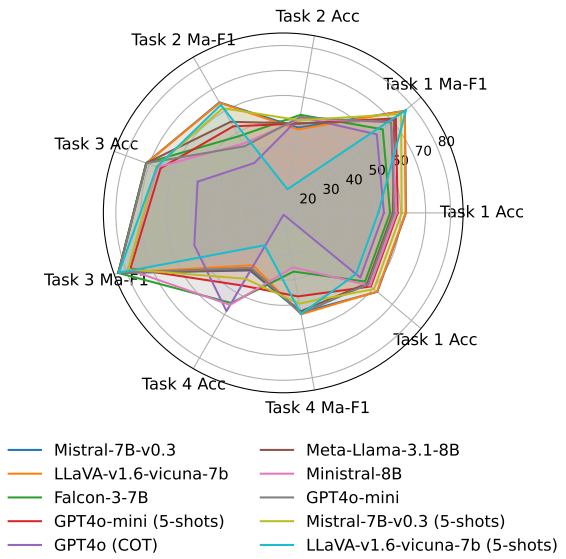

Figure 3: Radar chart for representative best-performing models from different evaluation settings. No single model dominates all four tasks.

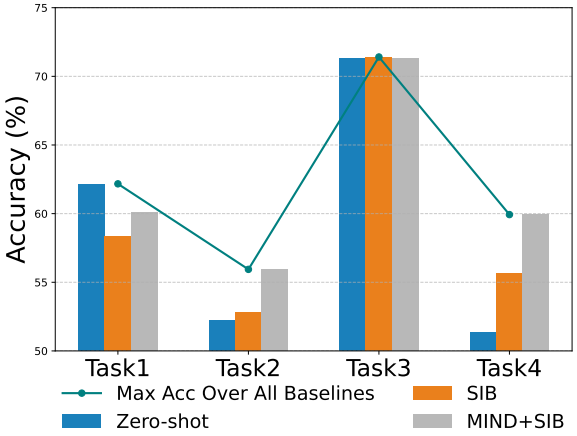

Figure 4: Comparison of the best performance across methods and tasks. The maximum accuracy achieved by the baselines is consistent with the maximum accuracy observed among open models across all methods (i.e., zero-shot, fine-tuning with SESSIONINTENTBENCH (SIB), and sequential fine-tuning with MIND followed by SIB). In this setting, using proprietary APIs does not provide additional performance gains.

textual signals as input, LVLMs can refer to image information to facilitate their question answering and inference reasoning. However, as shown in Table 2, the highest accuracy scores of LVLMs still lag behind those of LLMs. When evaluated on TASK 4 using direct zero-shot prompting, the best LVLM outcome is even behind the best LLM by a large gap of 11.27%. Possible reasons include the low signal-to-noise ratio of the collected images, and the fact that sellers usually include more comprehensive and concise product features in text format.

No model dominates across all tasks. Figure 3 visualizes representative top models from different settings. Mistral-7B-v0.3 is strongest among open zero-shot LLMs, LLaVA-v1.6-vicuna-7b is competitive among zero-shot LVLMs, and Falcon-3-7B performs best overall after SIB fine-tuning on several tasks. Yet none of them is uniformly best. This reinforces that SESSIONINTENTBENCH probes multiple distinct capabilities rather than a single dominant skill.

5.4 The Impact of Intention Injection

From Table 2, we observe that L(V)LMs struggle to directly leverage intention for next-product inference (*Intent-Based Inference*) and to capture long-term shifts in intention from session history (*Intention Evolution Modeling*). Table 3 further examines whether generic intention knowledge transfers to, or assists with, answering questions in SESSIONINTENTBENCH. We use MIND (Xu et al., 2024), a co-buy intention resource, as an external source of intention supervision and sequentially fine-tune models on MIND and then SIB. This *MIND + SIB* setting improves the best scores on Task 1 by 1.75 points, Task 2 by 3.09 points, and Task 4 by 4.24 points over SIB-only fine-tuning.

The gains are not uniform across tasks. Tasks 1, 2, and 4 depend directly on modeling latent intent and therefore benefit the most from intention injection. By contrast, Task 3 is already relatively strong in zero-shot and focuses more on local comparison consistency between adjacent products; it therefore has less room for improvement. We also evaluated models fine-tuned *only* on MIND. Those models achieved less than 10% accuracy on SIB because they did not reliably follow SIB's option-selection format, so we do not include them in the main table; details and examples are given in Appendix F.

6 Model Performance Insights

6.1 Evaluation Task Performance Metrics

We display the confusion-matrix statistics for GPT-4o with Chain-of-Thought prompting in Table 4. Task 4 has the smallest true-positive count and the largest true-negative share, which is consistent with its overall difficulty: models often default to broader exploration judgments rather than correctly identifying when a session should remain in

exploitation mode.

	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 4: Task performance metrics for error analyses of GPT-4o with Chain-of-Thought on SESSIONINTENT-BENCH. TP, FN, TN, and FP denote true positive, false negative, true negative, and false positive predictions, respectively.

6.2 Finetuning

SIB-only fine-tuning is helpful for learning task format, but it does not always improve generalization. One reason is that SIB contains a broad and heterogeneous distribution of product categories, attributes, and intention trajectories. This makes the train–test gap relatively large even when the question format is fixed (Wang et al., 2023b,a, 2024b; Wang and Song, 2025; Wang et al., 2025c). External intention supervision partially mitigates this issue: for example, Llama-3.1-8B drops from 56.87% on Task 1 in zero-shot to 52.82% after SIB-only fine-tuning, but improves to 60.10% under sequential MIND + SIB fine-tuning.

6.3 Error Analyses

We randomly sample 200 error cases from GPT-4o with Chain-of-Thought prompting and ask three NLP PhD researchers to analyze them. The most common failure mode (47.5%) is incorrect use of the provided metadata, especially failure to integrate earlier session context. Another 24% of the sampled errors arise from annotation-task mismatches. We also observe failures to capture decisive product features (7%), irrelevant or hallucinatory reasoning (6.5%), and broader difficulty inferring the session's overall intent when the provided metadata is vague or weakly decisive (15%). These percentages are computed within the error-analysis subset only; therefore, they should not be interpreted as estimates of the benchmark's overall noise rate.

7 Conclusions

In conclusion, we propose an automated pipeline to construct a large-scale knowledge base and further construct a sample dataset SESSIONINTENT-BENCH for L(V)LM evaluations. Extensive experiments show that current models struggle to understand and infer customers' intentions, while injecting intention information from other knowledge bases can improve 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 also hope this framework can benefit the community by enabling better services with future models.

Limitations

Our benchmark inherits limitations from both the source data and the curation process. First, a product interaction inside a session is only an imperfect proxy for user intention: users may click out of curiosity, because of presentation bias, or for reasons that are not reflected in the available metadata. Second, our intention tree is generated with GPT-4o-mini, so the benchmark may reflect generator bias even though we validate a subset through human annotation. Third, our current formulation does not use personalized signals such as long-term purchase history, demographics, or social context. Finally, Task 2 is inherently more subjective than the other tasks because which valued attribute best explains the observed transition is not always uniquely determined.

Ethics Statement

Offensive Content Inspection We use publicly available e-commerce resources and model-generated metadata to build the benchmark. The generated metadata are grounded in product information and constrained by the session context. We do not ask models to produce free-form harmful content; the downstream evaluation tasks are structured classification problems.

Annotation Wage Annotators were recruited through Amazon Mechanical Turk. Workers participated voluntarily and were paid at an average hourly rate of approximately USD 15, in accordance with local requirements.

Licenses The Amazon-M2 dataset is released under the Apache 2.0 license. This grants us free

access to the dataset. Our code and data will be shared under the MIT license. It allows the free distribution of the assets we propose and curate. All associated licenses permit user access for research purposes, and we agree to follow all terms of use.

Acknowledgments

The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from the Innovation and Technology Commission of Hong Kong SAR, China, and by the AoE (AoE/E-601/24-N), RIF (R6021-20), and GRF (16205322) from the Research Grants Council of Hong Kong SAR, China. We also thank the Amazon Stores Foundational AI team for their support and for providing valuable insights on data curation and evaluation.

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.

Jiaxin Bai, Zhaobo Wang, Junfei Cheng, Dan Yu, Zerui Huang, Weiqi Wang, Xin Liu, Chen Luo, Yanming Zhu, Bo Li, and Yangqiu Song. 2026. Intention knowledge graph construction for user intention relation modeling. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 466–484. Association for Computational Linguistics.

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. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *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. 2025a. 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.

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Chen Luo, Sheikh Muhammad Sarwar, Yang Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, and Yangqiu Song. 2025b. Ecomscriptbench: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1–22. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song. 2024b. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2351–2374. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Haochen Shi, Baixuan Xu, Wenxuan Ding, Liyu Zhang, Wei Fan, Jiaxin Bai, Haoran Li, Xin Liu, and Yangqiu Song. 2025c. On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 2260–2281. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Weiqi Wang, Xin Liu, Binxuan Huang, Hejie Cui, Rongzhi Zhang, Changlong Yu, Shuowei Jin, Jingfeng Yang, Qingyu Yin, Zhengyang Wang, Zheng Li, Yifan Gao, Priyanka Nigam, Bing Yin, Lihong Li, and Yangqiu Song. 2026a. Heapa: Difficulty-aware heap sampling and on-policy query augmentation for LLM reinforcement learning. *CoRR*, abs/2601.22448.

Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme, and Daniel Khashabi. 2026b. arxiv2table: Toward realistic benchmarking and evaluation for llm-based literature-review table generation. *Preprint*, arXiv:2504.10284.

Weiqi Wang and Yangqiu Song. 2025. MARS: benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1568–1596. Association for Computational Linguistics.

Yu Wang, Amin Javari, Janani Balaji, Walid Shalaby, Tyler Derr, and Xiquan Cui. 2024c. 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. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 1173–1191. Association for Computational Linguistics.

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *Preprint*, arXiv:2403.13372.

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 product attributes with GPT-4o-mini, we use the following 3-shot prompt 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 Appliance
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 and populate it with intentions, valued attributes, and supporting comparisons, we use the following 5-shot 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 factors, we simply reduce the number of requested outputs. For larger branching factors, we extend the prompt with additional examples as needed.

A.3 Intention Generator Model Selection

We first tested open LVLMs such as Mantis and LLaVA for metadata generation, but found them unreliable for large-scale intention-tree construction. The main issue was not raw semantic quality, but format control: many open models failed to follow the required output schema consistently when given long textual metadata together with images, occasionally degenerating into repeated tokens or malformed structures. GPT-4o-mini was substantially more stable while remaining cost-effective, so we use it as the primary generator for the intention tree.

A.4 Fine-tuning Model Selection

We select fine-tuning backbones using three criteria. *(1) Organizational diversity:* we choose models from different model families to avoid overfitting conclusions to a single ecosystem. *(2) Size constraint:* models must be smaller than 11B parameters so they can be fine-tuned on our hardware. *(3) Representative strength:* within each family, we choose a strong and practically relevant checkpoint. For example, we use Llama-3.1-8B as Meta's representative 7B/8B backbone because later Llama-3.2 releases focus on smaller models and Llama-3.3 targets much larger ones.

A.5 Training-Test Splits

The detailed process is outlined as follows: *(1) Indexing:* We created an index for all annotated questions. Each questionnaire contained four questions corresponding to Tasks 1–4, ensuring an equal number of samples per task. Therefore, the proportions of indices for each question are equal. *(2) Index Set Creation:* A unified set of indices was constructed, where each index uniquely corresponds to a specific Task and Session number for traceability. *(3) Splitting:* We adopted a 4:1 train-test split. Indices were randomly sampled to create the training and test sets. Although the number of samples for Tasks 1–4 may vary slightly due to random sampling, the distribution remains largely balanced. *(4)* The resulting training and test sets were used across different models and training schemes (e.g., zero-shot, fine-tuning with SIB, and sequential fine-tuning with MIND followed by SIB)

A.6 Few-shot Example Curation

When curating few-shot demonstrations, we prioritize clarity over distributional coverage. Each demonstration should be concise, internally consistent, and easy for a human annotator to judge. We first generate a larger candidate pool with GPT-4o across several product categories, then manually filter, revise, and validate the final examples. We also test prompt variants with different example categories and example counts; on GPT-4o, these changes lead to only minor accuracy fluctuations (within 1%), so we use a single 5-shot configuration throughout the paper.

A.7 Model Evaluation

We evaluate models with zero-shot prompts (Table 8), 5-shot prompts (Tables 9–12), and Chain-of-Thought prompts (Table 13). For consistency, every result described as "Few-shots" in Table 2 uses exactly five demonstrations.

A.8 Finetuning Methods

All fine-tuning experiments use supervised fine-tuning (SFT) with LoRA. We implement them

with the open-source LLaMA-Factory framework (Zheng et al., 2024).

A.9 Role of GPT-4o-mini

GPT-4o-mini is used in both dataset construction and proprietary-model evaluation. We include it in evaluation because it is the generator used to produce the intention tree, and we want to test whether generation-time familiarity translates into an unfair advantage. It does not: GPT-4o-mini is competitive, but not dominant, on the benchmark. To avoid confounding the error analysis with the generation model, we analyze GPT-4o Chain-of-Thought errors rather than GPT-4o-mini errors in Section 6.3.

A.10 Expert Selection for Error Analysis

The error-analysis experts are three Computer Science PhD researchers from our institution, each with multiple publications in NLP or closely related areas. We randomly sample 200 incorrect GPT-4o-CoT predictions across Tasks 1–4. Each expert independently reviews the sampled cases using a shared error taxonomy, and the final labels are determined by consensus discussion.

B Theoretical Framework

B.1 Intention Tree

The intention tree $\mathbf{T}$ is defined inductively over the session timeline. At each time step $t \in \mathbb{N}^+$, we extend $\mathbf{T}_{1,\dots,t-1}$ to $\mathbf{T}_{1,\dots,t}$ by attaching one or more candidate intentions for the newly observed product. The information accumulated up to time step t is denoted by $\mathcal{H}_t$.

Traditional formulations often predict the next interaction directly from the most recent product or intention state, for example through $\mathbb{P}(P_{t+1} \mid P_t)$ or $\mathbb{P}(I_{t+1} \mid I_t)$. Our formulation instead factorizes the problem through an intermediate session-level latent state:

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

where $\mathcal{M}_t^{\phi}$ is the model's approximation of the latent session state $\mathcal{M}_t$.

Rather than representing this latent state as a monolithic variable, we decompose it into three explicit components:

- I_t^{ζ}: inferred intention,
- A_t^{ζ}: valued attribute, and
- C_t^{ζ}: comparison metadata.

Together they form

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

so the next-step reasoning problem becomes

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

The superscript ζ indicates that these components are branch-specific approximations generated by the intention-tree construction model.

This decomposition also motivates the four tasks. For example, in Valued Attribute Regularization we sample a subset of candidate attributes $\mathcal{A}^{\pi} \subseteq \mathcal{A}$ and ask the model to determine whether the transition to P_{t+1} is plausible under the assumption that a valued attribute lies in $\mathcal{A}^{\pi}$. The task can therefore be seen as constraining the branch state so that $A_t^{\zeta} \in \mathcal{A}^{\pi}$ and then evaluating whether the resulting branch remains compatible with the observed session history.

B.2 Intuition

The branching process is easiest to understand with a short session. Suppose a session contains two products and uses a 5-branching scheme. We first infer one intention for the first product. After the second product is observed, the model proposes five plausible follow-up intentions. When a third product arrives, each of those intentions can in turn branch into additional candidate continuations. Every intention node is paired with one attribute, one rationale, and one comparison.

C Task and Evaluation Design

C.1 Design Criteria for Choice Options

The 0–3 scores in the task definitions are symbolic encodings of concrete answer choices. For Tasks 1–3, score 3 corresponds to a strong positive judgment, score 2 to a weak positive judgment, score 1 to a weak negative judgment, and score 0 to a strong negative judgment. Table 8 provides the exact formulations used in prompting and annotation.

For evaluation, we merge adjacent options into binary labels. Specifically, for Tasks 1–3 we group A/B together and C/D together. This reduces neutral-response bias and makes the final prediction problem better aligned with practical decision settings. For Task 4, the positive class is option A

(continue recommending similar products), while B/C are grouped as the negative class because they both indicate a need for broader exploration.

C.2 Why Not Use Open-Ended Question Answering

We considered an open-ended formulation but ultimately chose structured multiple-choice evaluation for three reasons. First, open-ended intention proposals are difficult to standardize and cluster at scale, which makes both benchmarking and human evaluation much less reliable. Second, open-ended annotation is substantially more expensive because it requires more expert labor per example. Third, our worker-selection process already shows that reliability is a challenge even for structured questions: among 300 initial candidates, only 11 passed the full quality-control pipeline. In our setting, a structured formulation offers the best trade-off between scalability, annotation quality, and reproducibility.

D Annotation Process

D.1 Worker Selection Protocol

We apply a multi-stage quality-control pipeline to obtain reliable annotations. Qualification invitations are sent only to AMT workers with more than 2,000 approved HITs and an approval rate above 90%. We then administer a qualification test built from sampled sessions with author-validated gold labels. Workers must score above 75% while completing at least 20 questions to move forward.

After the qualification stage, we further remove workers who exhibit obvious low-effort behavior, especially those who choose the same side of the label space almost all the time. A second screening round reveals multiple such one-sided annotators, who are excluded before the main round. In the end, 11 workers remain out of 300 initial candidates, corresponding to a 3.67% retention rate.

D.2 Annotation Instructions

We present the annotation interface in non-technical language while keeping it closely aligned with the task definitions in Section 3. For the first three questions, workers assign a score on a four-point plausibility scale from 0 to 3. For the fourth question, workers choose among three exploration options, also mapped to an ordered scale. We explicitly explain the session-product list, the proposed intention metadata, and the meaning of each answer choice to reduce ambiguity.

E Annotation Result Analysis

E.1 Raw Label Result

The label distribution summarized in Table 5 corresponds to the Majority baseline in Table 2. We report merged binary labels, for Tasks 1–3, A/B and C/D are grouped together, and for Task 4, A is contrasted with B/C. We grouped them to mitigate individual annotator biases observed during the annotation process, where some annotators consistently favored extreme responses while others tended to choose intermediate options.

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 5: Merged label counts and percentages after binarization. *Task_Ind* denotes the task index. For Tasks 1–3, A_B denotes positive labels and C_D denotes negative labels; for Task 4, A denotes the exploitative recommendation choice and B_C denotes the exploratory choices.

E.2 Consistency

We establish the final ground truth by majority vote over three annotators. In Table 6, "3:0" means full agreement and "2:1" means one annotator disagrees with the other two after labels are mapped to the binary evaluation space. More than half of the questions receive full agreement in their binary labels, indicating that the benchmark is difficult but not annotation-random.

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

Table 6: Consistency analysis of binary answer label distribution. *Task_Ind* denotes the task index, ranging from 1 to 4.

E.3 Annotation Quality Filter

Beyond dataset-level statistics, we also inspect individual annotator behavior. Table 7 illustrates a

failure mode observed during worker screening: some workers overwhelmingly favor one side of the label space regardless of the example. We use this pattern as one of the exclusion criteria in the qualification pipeline, because it suggests low engagement rather than a principled annotation strategy.

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 7: Annotators excluded during screening because they overwhelmingly favored the same option across many questions.

E.4 Benchmark and Data Quality Validation

Some sessions are inherently ambiguous. Users may make abrupt jumps between products, and the generated intention metadata can occasionally expose those inconsistencies rather than resolve them. These cases are part of the difficulty of the benchmark, although additional preprocessing could reduce them further in future releases.

E.5 Clarifications on the Majority Vote Score

The Majority baseline in Table 2 is a task-level prediction baseline, *not a measure of human performance*. For each task, we examine the final binary labels over the whole dataset and *always predict the globally most frequent class*. For example, on Task 3 the majority class is the positive side (A/B), while on Tasks 2 and 4 it is the negative side. This baseline is useful because it reveals *label skew*, but it should not be interpreted as evidence that a fixed answer is correct for every question.

E.6 Missing F1 Score for Tasks 2 and 4

The NaN Macro-F1 values for the Majority baseline on Tasks 2 and 4 are a consequence of the prediction distribution. Because the baseline always predicts the negative class on those tasks, it never produces a positive prediction; precision for the positive class is therefore undefined, which propagates to the F1 score. Accuracy remains well defined and is still reported.

F Model Performance Insights

F.1 Imbalanced Task Performance Gain with Intention Injection

The gains from intention injection are largest on tasks that require the model to use latent intent directly. Task 1 asks whether a new product is compatible with a proposed intent, Task 2 asks whether a valued attribute explains the transition, and Task 4 asks whether the session has become exploratory. All three benefit from generic intention supervision. Task 3 is different: it primarily tests whether a generated comparison is locally coherent with adjacent products, so it already has a strong zero-shot signal and shows less room for improvement.

F.2 Solely Fine-Tuning with MIND

Models fine-tuned only on MIND perform poorly on SIB (below 10% accuracy). The main issue is an output format mismatch. MIND is open-ended and encourages descriptive intent generation, whereas SIB requires selecting discrete options. As a result, MIND-only models tend to produce fluent but verbose reasoning without clearly stated answers, leading to uniformly low performance across models (<10%). This issue is further amplified by model capacity: small models used in our experiments (those under 11B parameters) have limited ability to rapidly adapt to a different answer format at inference time without additional task-specific fine-tuning. We intentionally keep MIND in the transfer study because its open-ended intent supervision is precisely what makes it a useful source of generic intention knowledge, but it must be followed by SIB fine-tuning to adapt to the benchmark format (Wang et al., 2026b,a).

["instruction": "Act as a customer who bought these two product: ", "input": "Product A: Lincoln Stain Wax Shoe Polish 3 Fl Oz (Selection of Colors); Product B: Angelus Shoe Wax Polish 3fl Oz (Color Variety); What is your possible co-buy intention for these two products?", "output": "The potential co-buy intention could be that the person wants to purchase both shoe polish products to have a variety of colors to choose from when polishing their shoes" , "instruction": "Act as a

```
customer who bought these two product: ",
"input": "Product A: BMC Mens 6 pc Mixed
Design Self Tie Bowtie Pocket Square
Suit Accessories; Product B: Tenby
Living 2-Pack Black Tie Rack, Organizer,
Hanger, Holder - Affordable Ti.; What
is your possible co-buy intention
for these two products?", "output":
"The potential co-buy intention for
people purchasing these two products
simultaneously could be to enhance their
wardrobe and maintain an organized and
stylish appearance" , ...]
```


F.3 The BERT-based Models

We also test strong pretrained encoder baselines such as RoBERTa-large-355M and DeBERTa-v3-large, but do not include them in the main comparison because they fail to follow the task format reliably.

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**", ...]
```


RoBERTa often defaults to generic answers such as "A" or "Yes" regardless of the question, suggesting that it is not grounding its prediction in the full session context.

For DeBERTa-v3-large, the raw output often consists of malformed strings such as "IBILITY" and "Measurement" rather than valid options:


```
["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**", ...]
```


These failures suggest that standard encoder-only models are not well suited to the instruction-following setting required by SESSIONINTENT-BENCH.

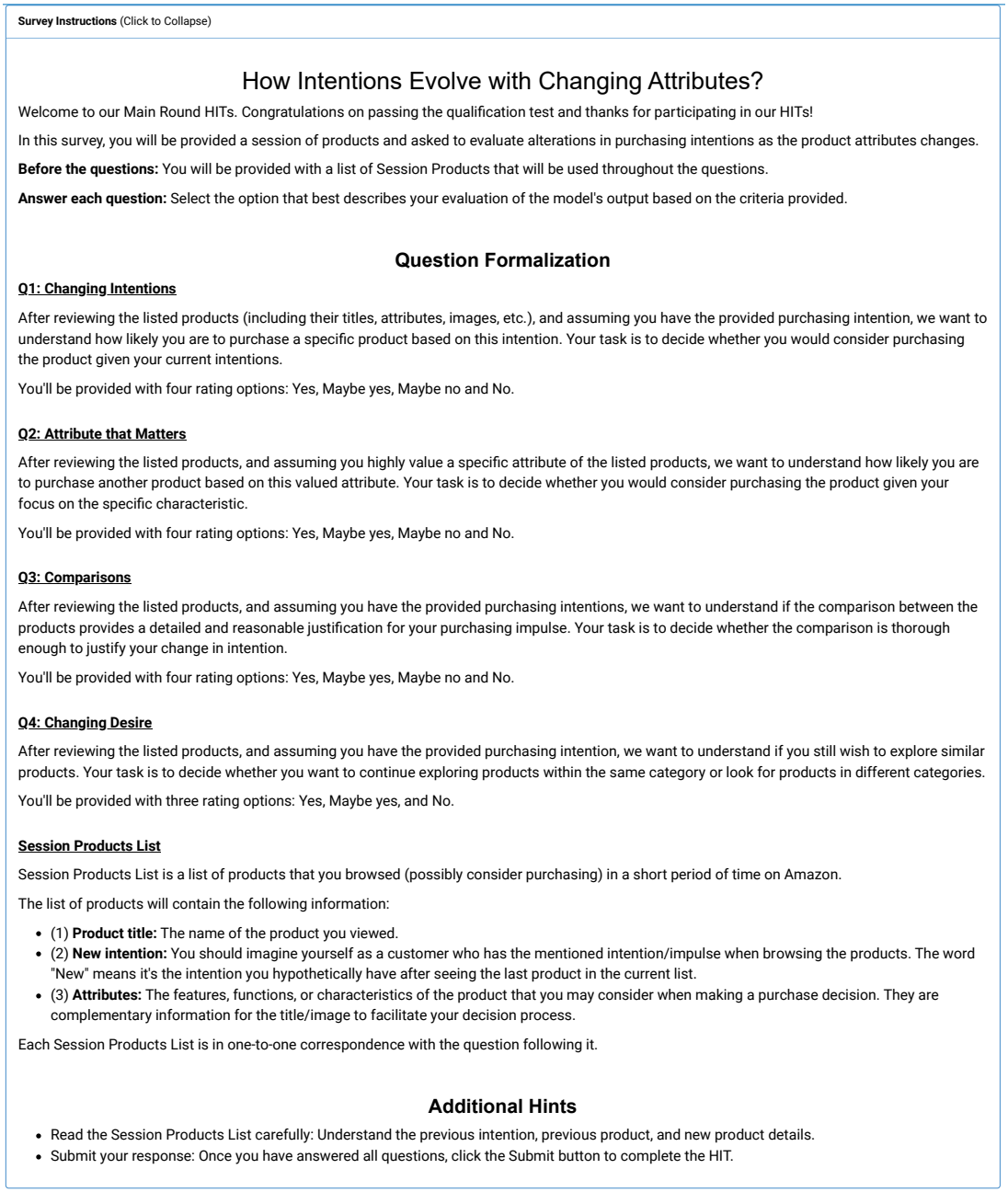

Figure 5: Annotation instructions shown to workers. The interface explains the task definitions in plain language and previews the information available in the session-product list.

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***.