IntentionQA: A Benchmark for Evaluating Purchase Intention Comprehension Abilities of Language Models in E-commerce

Wenxuan Ding[1*], Weiqi Wang[1*], Sze Heng Douglas Kwok[1], Minghao Liu[1], Tianqing Fang[1], Jiaxin Bai[1], Xin Liu[2], Changlong Yu[2], Zheng Li[2], Chen Luo[2], Qingyu Yin[2], Bing Yin[2], Junxian He[1], Yangqiu Song[1]

[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[2]Amazon.com Inc, Palo Alto, CA, USA

wdingaj@connect.ust.hk, {wwangbw, junxianh, yqsong}@cse.ust.hk
{xliucr, changlyu, amzzhe, cheluo, qingyy, alexbyin}@amazon.com

Abstract

Enhancing Language Models' (LMs) ability to understand purchase intentions in E-commerce scenarios is crucial for their effective assistance in various downstream tasks. However, previous approaches that distill intentions from LMs often fail to generate meaningful and human-centric intentions applicable in real-world E-commerce contexts. This raises concerns about the true comprehension and utilization of purchase intentions by LMs. In this paper, we present IntentionQA, a double-task multiple-choice question answering benchmark to evaluate LMs' comprehension of purchase intentions in E-commerce. Specifically, LMs are tasked to infer intentions based on purchased products and utilize them to predict additional purchases. IntentionQA consists of 4,360 carefully curated problems across three difficulty levels, constructed using an automated pipeline to ensure scalability on large E-commerce services. Human evaluations demonstrate the high quality and low false-negative rate of our benchmark. Extensive experiments across 19 language models show that they still struggle with certain scenarios, such as understanding products and intentions accurately, jointly reasoning with products and intentions, and more, in which they fall far behind human performances. Our code and data are publicly available at https://github.com/HKUST-KnowComp/IntentionQA.

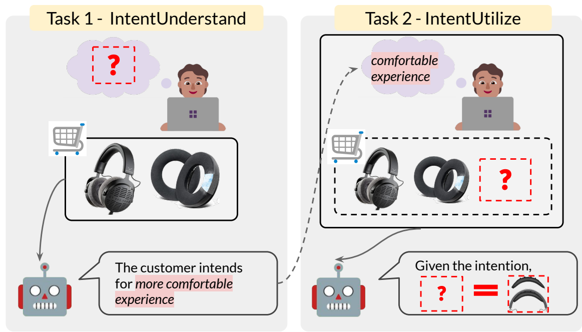

Figure 1: Examples of two tasks in IntentionQA. Task 1 requires the language model to determine the customer's intention in purchasing two products, and Task 2 involves recommending a product that fulfills the customer's intention and matches their currently purchased product.

1 Introduction

Purchase intentions are mental states where agents or humans commit themselves to purchasing the products (Yu et al., 2023). Understanding customers' purchase intentions and making reasonable inferences accordingly are crucial for revolutionizing E-commerce services, whose benefits have been demonstrated in myriads of downstream tasks, such as product recommendation (Grbovic et al., 2015; Zhao et al., 2014; Li et al., 2020) and query answering (Zhao et al., 2019; Hirsch et al., 2020). However, intention comprehension (Fogassi et al., 2005) is a non-trivial task as it involves reasoning with implicit mental states, which are not typically expressed in text or conversations. Thus, in the context of E-commerce, extracting purchase intentions from behaviors without explicit external cues has been challenging (Yang and Tang, 2015).

Recently, Yu et al. (2023, 2024); Lu et al. (2024) proposed to distill purchase intentions from LLMs, such as OPT (Zhang et al., 2022b), by leveraging their inherent advantages of generative and commonsense reasoning abilities, as well as being pretrained on vast textual data including E-commerce knowledge. However, recent analyses by Zhou et al. (2024) show that LMs struggle to generate meaningful and user-centric intentions. Instead, they are biased by over-focusing on similarities among different products' metadata, such as their properties, and often end up regurgitating information from the provided prompts without truly comprehending the underlying purchase intentions.

* Equal Contribution

Findings of the Association for Computational Linguistics: EACL 2024, pages 2247–2266
November 12-16, 2024 ©2024 Association for Computational Linguistics

Thus, an important yet under-explored question arises: *Can LMs comprehend the customers' purchase intention and how effective are they in performing such tasks?* To dive into this, we first break down the comprehension of intention into two key aspects, as shown in Figure 1. First, we have **intention understanding**, which evaluates LMs' capacity to accurately infer customers' purchase intentions based on the products bought. Second, we consider **intention utilization**, which investigates LMs' ability to predict additional purchases based on customer's intentions. Together, they make up the entire process of intention comprehension and play a significant role in enhancing E-commerce search services.

Although LMs have been extensively used in intention knowledge distillation, their actual performances in this area have not been adequately benchmarked. This is because current methods that leverage LMs have been adopting an open-ended generation fashion, which is difficult to consistently evaluate (Gu et al., 2021). Additionally, the extensive and constantly growing number of products on E-commerce stores makes it infeasible and expensive to construct human-curated benchmarks.

To address these challenges and benchmark LMs on purchase intention comprehension in E-commerce, we introduce INTENTIONQA, a double-task multiple-choice question answering (MCQA) dataset, featuring intention understanding and intention utilization respectively. INTENTIONQA contains 4,360 problems for two tasks and covers varying difficulty levels, allowing for fine-grained evaluation. The MCQA setting enables using consistent evaluation metrics to assess the LMs' intention comprehension abilities.

Specifically, we design a pipeline that automatically synthesizes QA pairs by transforming human-annotated intentions from FolkScope (Yu et al., 2023), each involving a pair of co-buy products and the corresponding intention of purchasing them, into questions by masking out the intention or one of the products. To achieve this, we define context-based product similarity and intention similarity metrics. They are computed over ASER (Zhang et al., 2022a), a large-scale eventuality knowledge graph, which we leverage as a reference for our automatic distractor sampling strategy. For each question, we include 3 negative distractors alongside the gold answer through a strict similarity filtering process. We then assign difficulty labels to each QA pair based on the product similarity between the co-buy products in the original intention assertion. These steps are done without human supervision, enabling our benchmark construction pipeline to generalize and accommodate larger-scale product databases and practical applications.

We further conduct human evaluations to demonstrate the high quality and low false-negative rate of INTENTIONQA, followed by extensive experiments across 19 language models with varying sizes and approaches. Results demonstrate that the existing language models still struggle with certain scenarios, such as understanding products and intentions accurately, jointly reasoning with the products and intentions, and more. In the long run, we hope that our benchmark serves as an important cornerstone toward intention-aware E-commerce services that promote integrating intention reasoning abilities into product recommendations.

2 Related Works

2.1 Intention Discovery with Large Language Models in E-commerce

Understanding intentions with language models have been studied in various domains, such as smoothing chatbox conversations (Ouyang et al., 2022), enhancing web search (Zhang et al., 2019), and more. In the E-commerce domain, understanding customers' purchase intentions benefits various downstream tasks (Koo and Ju, 2010; Xu et al., 2024; Wang et al., 2024a), such as automated on-call customer support (Goyal et al., 2022), recommendation systems (Dai et al., 2006; Qian et al., 2023; Jung et al., 2023), product question answering (Deng et al., 2023; Yu and Lam, 2018). While Yu et al. (2023, 2024) proposed leveraging the generation abilities of LLMs to distill purchase intentions from co-buy records, Zhou et al. (2024) showed that LLMs struggle with generating meaningful intentions or understanding user-centric intentions. In this work, we construct INTENTIONQA, a benchmark to evaluate LMs' intention comprehension abilities by selecting highly typical intentions in previously available resources and provide insights for human-centric intention comprehension.

2.2 Benchmarking (Large) Language Models

Since the emergence of (L)LMs, various studies have explored their capabilities in various domains, including temporal reasoning (Tan et al., 2023), causal reasoning (Chan et al., 2024), commonsense

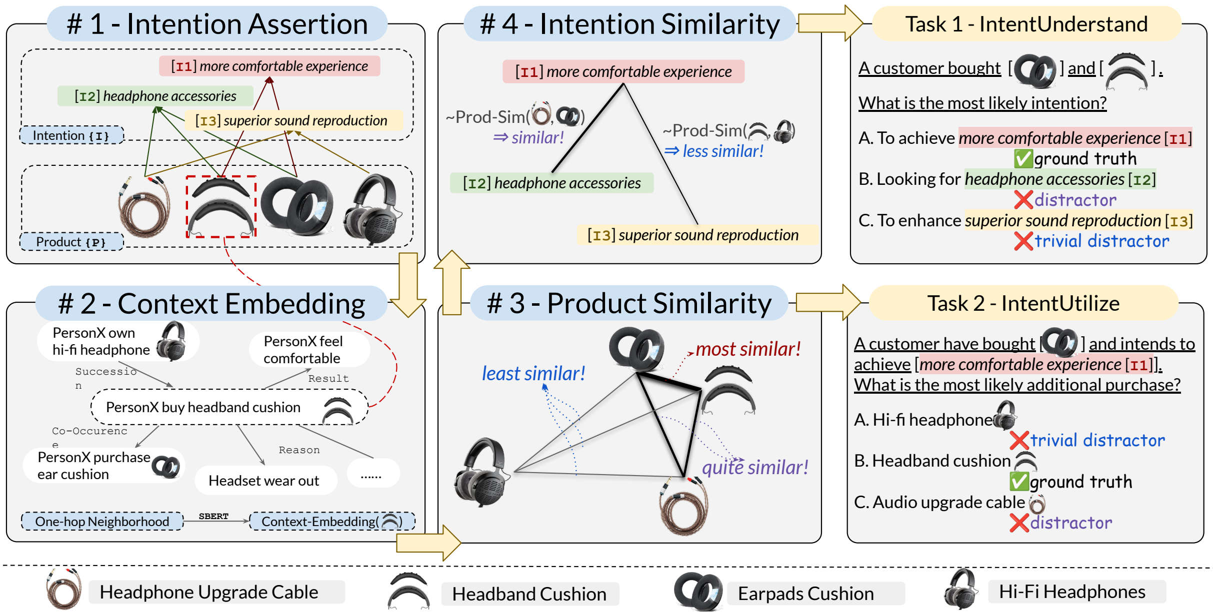

Figure 2: Overview of INTENTIONQA and the construction pipeline. We map products from intention assertions to event nodes in ASER (# 1) and calculate their context embedding with the one-hop neighborhood (# 2). Product (#3) and intention (#4) similarities are then computed accordingly. Products/intentions with higher similarities are represented closer to each other. Negative distractor sampling for Task 1/2 is based on intention/product similarity respectively.

reasoning (Jain et al., 2023), and more (Qin et al., 2023; Ding et al., 2023; Bai et al., 2023; Feng et al., 2024). These benchmarks have made significant contributions to the understanding of large language models, assessing their performance across different parameters and prompting methods. However, in the field of E-commerce, existing works primarily leverage LLMs with explicit instruction-tuning (Li et al., 2024), while neglecting the infeasibility of directly applying LLMs in a generalizable manner. Furthermore, current evaluation benchmarks in E-commerce primarily emphasize product and session comprehension (Jin et al., 2023), which overlooks the important aspect of intention comprehension. In this paper, we step forward by presenting the first benchmark that evaluates the intention comprehension abilities of (L)LMs.

3 INTENTIONQA

In this section, we introduce INTENTIONQA, a Multiple-Choice Question Answering (MCQA) benchmark consisting of two tasks targeting different aspects of purchase intention comprehension and with progressive difficulties, to evaluate the intention understanding and utilization abilities of LMs thoroughly.

3.1 Task Definitions

We begin by formally defining two tasks associated with INTENTIONQA. For the tasks presented, we specifically refer to purchase intention as the intention that drives the customer to buy a pair of products together.

Task 1: INTENTUNDERSTAND The first task examines whether LMs can infer the purchase intentions correctly given a real-world record of the products bought. Formally, given a pair of co-buy products p_1, p_2, LMs are tasked with selecting the most likely purchase intention i^* from a list of candidate options $\mathcal{I} = [i_1, i_2, \ldots, i_{|\mathcal{I}|}]$.

Task 2: INTENTUTILIZE The second task looks further into the capacity of LMs to utilize purchase intention for the product recommendation process. We approach this by examining their abilities to predict the most likely additional purchase based on customer intention. Specifically, given the purchase intention i^* and one product that has been Bought p^B, the LMs are tasked with selecting the most likely Additional purchase p^{A*} from a list of candidate options $\mathcal{P}^{\mathcal{A}} = [p^A_1, p^A_2, \ldots, p^A_{|\mathcal{P}^{\mathcal{A}}|}]$.

3.2 Source Intention Collection and Context Augmentation

We collect co-buy products and intention assertions from FolkScope (Yu et al., 2023) as our source data. FolkScope is an intention knowledge base that is constructed by distilling knowledge from a pre-trained large language model, OPT (Zhang et al., 2022b). It associates customers' co-purchase behaviors with their purchase intentions, as shown in the upper left part (# 1) of Figure 2. Two scores are also assigned to each intention, indicating its plausibility and typicality. To accommodate our tasks, we preprocess FolkScope by filtering and retaining plausible assertions with typicality scores above 0.5. This is to minimize the number of overly-general intentions, which may be plausible for most products but are not specifically related to the given products. Including these intentions in IntentionQA could lead to many false negative distractors, which harms the quality of our QA pairs.

Since we are aiming for automatic QA pair construction, determining the similarity between different intentions and products can serve as powerful hints in selecting appropriate distractors given a correct answer. However, relying solely on product metadata and corresponding purchase behavior falls short of capturing the similarity between intentions, as similar or identical intentions can align with multiple products. To address this limitation and enhance the sampling of distractors while reducing the occurrence of false-negative distractors, we introduce a method to augment customers' purchase behavior. This is achieved by retrieving additional relevant context from ASER (Zhang et al., 2020, 2022a), a large-scale eventuality knowledge graph that covers billions of commonly seen eventualities. We choose ASER for its extensive knowledge coverage, prompt consistency, and cost-effectiveness compared to API-accessed LLMs (Appendix A.1).

Specifically, we first consider the purchasing event as an eventuality and design heuristic rules to align it with nodes in ASER. Formally, we denote ASER as $\mathcal{G} = \{(h, r, t)|h, t \in \mathcal{V}, r \in \mathcal{R}\}$, where $\mathcal{V}$ and $\mathcal{R}$ are the sets of nodes and relations in ASER. In ASER, the nodes are events and the edges are relationships between different events. For example, a node in ASER could be "`PersonX feel comfortable`" and this node is the *result* of the node "`PersonX buy headband cushion`". Tile # 2 in Figure 2 is an example subgraph from ASER.

Inspired by recent works in conceptualization (Wang et al., 2023b,a; Wang and Song, 2024; Wang et al., 2024c; He et al., 2024), we simplify the product name p by instructing ChatGPT to conceptualize it into three plausible categories $\mathcal{C}(p) = [c_1, c_2, c_3]$, using prompts presented in Table 5. For example, *IPhone 14* can be conceptualized as a *phone*, *communication device*, and *Apple product*. This augmentation expands the semantic coverage of the purchasing event, increasing the likelihood of finding relevant nodes in ASER.

Next, we design natural language templates (Appendix D) to convert noun phrases of conceptualized product categories into purchasing events $\mathcal{E}(\mathcal{C}) = [e_1, e_2, \ldots, e_{|\mathcal{E}|}]$. These events are then matched against nodes in ASER to identify overlapping ones through strict string matching. Formally, we denote ASER as $\mathcal{G} = \{(h, r, t)|h, t \in \mathcal{V}, r \in \mathcal{R}\}$, where $\mathcal{V}$ and $\mathcal{R}$ are the sets of nodes and relations in ASER. We denote the set of matched nodes for p in ASER as $\mathcal{VE} = \mathcal{E}(\mathcal{C}) \cap \mathcal{V} = [ve_1, ve_2, \ldots, ve_{|\mathcal{VE}|}]$. Next, we compute the sentence embedding of edges in the one-hop neighborhood of each node in $\mathcal{VE}$ using Sentence-BERT (Reimers and Gurevych, 2019). The context embedding $CE(p)$ is then computed by averaging these embeddings, which serves as the semantic representation of relevant contexts for purchasing the product p.

3.3 Product and Intention Similarity

With the context embeddings of all products computed, they contain valuable background knowledge about purchasing events associated with each product. This includes edges from ASER that capture cause-effect relationships ("Reason" and "Result"), event precedence and succession ("Precedence" and "Succession"), and other relevant information. Intuitively, similar products should have similar contextual information in ASER, and vice versa, as illustrated in Tile # 3 of Figure 2. Thus, we define the similarity between purchasing events of p_1 and p_2 as follows:

$$\mathit{Sim}^{(p)}(p_1, p_2) = \mathit{cos\_sim}(CE(p_1), CE(p_2))$$

where *cos_sim* is the cosine similarity between embeddings from Sentence-BERT.

Similarly, we define the similarity between two intentions (i_1, i_2) in FolkScope by leveraging their corresponding purchase events associated with ASER as follows:

$$Sim^{(i)}(i_1, i_2) = \min_{m=1,2;n=1,2}\{Sim^{(p)}(p_m^{(i_1)}, p_n^{(i_2)})\}$$

where $p_m^{(i)}$ is the $\mathrm{m_{th}}$ product linked to intention i (Tile # 4 of Figure 2).

3.4 Distractor Sampling and QA Construction

Finally, we design INTENTIONQA as a multiple-choice QA benchmark and design specific rules to transform intention assertions in FolkScope into question and gold answer pairs. Each gold answer is accompanied by three distractors, i.e., $|\mathcal{I}|, |\mathcal{P}^{\mathcal{A}}| = 4$. For each task, we propose its unique distractor sampling strategy specifically designed for the task objective, based on the similarity scores defined in §3.3.

Task 1: INTENTUNDERSTAND INTENTUNDERSTAND targets LMs' ability to accurately infer purchase intentions based on the products bought by a customer. We convert the intention assertions from FolkScope to questions by masking out the intentions. These masked intentions are then treated as gold answers, denoted as i^*. To obtain the distractor intentions $\mathcal{I}^- = [i_1^-, i_2^-, i_3^-]$, we randomly select intentions from FolkScope whose intention-similarity score with i^* ($Sim^{(i)}(i^*, i^-)$) fall within $[0.6, 0.9]$. The lower bound of the range filters out trivial distractors, while the upper bound minimizes the false negative rate in the resulting benchmark.

Task 2: INTENTUTILIZE INTENTUTILIZE evaluates the LMs' ability to utilize intentions to predict future purchase behavior. Specifically, we formulate the task as providing LMs with one product that the customer has bought and the corresponding intention, and task LMs with predicting the most likely purchase accompanied by the purchased product. Questions for INTENTUTILIZE are obtained by masking out one of the products (p^{A*}) in each intention assertion of FolkScope. The distractor products ($p_i^-, i = 1, 2, 3$) are randomly selected from those products whose product-similarity score $Sim^p(p^{A*}, p^{A-})$ falls within $[0.7, 0.9]$. Threshold values for both tasks are determined through observations of the distribution and preliminary experiments.

Difficulty Labeling To allow for fine-grained evaluation, we categorize each question into three difficulty levels. Intuitively, intention assertions with high product-similarity scores among co-buy products result in relatively easy problems. This is based on the assumption that understanding just one product is sufficient for comprehending the corresponding intention, without necessitating reasoning about the relationship between the products. Conversely, intention assertions with low product-similarity scores contribute to harder problems as they require comprehending both products and their corresponding intentions, as well as reasoning about the potentially complementary relationship between the products.

Therefore, we categorize the problems based on the product-similarity scores of co-buy products in the original intention assertion. Specifically, problems with a product-similarity score within the range of $[0.85, 1]$ are classified as easy problems, those within the range of $[0.6, 0.85)$ are considered medium, and those within the range of $[0, 0.6)$ are classified as hard problems. These thresholds are determined based on distributions and human observations of problem difficulty.

Quality Control After a preliminary human evaluation of the resulting QA pairs for both tasks, we observe that for the hard subset problems in TASK 1 the correctness rate is relatively low and the false-negative rate is relatively high. Therefore, we manually review every problem in this subset and discard those with incorrect gold answers or false-negative options. (Details in Appendix A.3)

4 Benchmark Evaluations

4.1 Statistics

We initially construct INTENTIONQA by using 2,315 intention assertions sourced from FolkScope. They are selected by filtering those with high plausibility and typicality scores and whose both products can be aligned with purchasing event nodes of ASER. We then construct 4,360 problems for both tasks in INTENTIONQA, with each problem labeled with difficulty accordingly. The benchmark statistics are reported in Table 1.

4.2 Human Evaluations

To evaluate the effectiveness of our benchmark construction pipeline and assess the quality of our constructed QA benchmark, we conduct human annotation to evaluate (1) the correctness of product conceptualization by ChatGPT and (2) the quality of the QA pairs in both tasks.

Subset	Task 1		Task 2	
	#Q	Avg. Sim^p	#Q	Avg. Sim^p
easy	1703	0.972	1625	0.971
medium	424	0.740	385	0.744
hard	90	0.530	133	0.514
Average	2217	0.905	2143	0.902

Table 1: Statistics of the IntentionQA. We report the number of questions (#Q) and the average product-similarity scores between the co-buy products among all intentions (Avg. Sim^p) within each difficulty subset.

Subset	Task 1		Task 2	
	Correct	F-Neg	Correct	F-Neg
easy	96.07	2.77	98.20	1.20
medium	94.00	2.67	92.59	4.32
hard	100.00	0.00	100.00	0.00
Average	96.00	2.56	97.33	1.67

Table 2: Annotated correctness (Correct; %) and false-negative rate (F-Neg; %) of 600 randomly sampled QA pairs from two tasks.

4.2.1 Annotation Details

We recruit human annotators from the Amazon Mechanical Turk service for human evaluation.

For product conceptualization, results show that 89.4% of products are reasonably conceptualized, demonstrating the strong product understanding ability of ChatGPT and validating the feasibility of leveraging its generative power to aid our benchmark construction process.

For the quality of resulting QA pairs, we randomly sample 300 QA pairs and ask the annotators to assess the quality of these problems, including the correctness of ground truth options (*Correct*) and assess the false-negativeness of the distractor options by determining whether a distractor option is superior to or equally plausible as the ground truth option (*F-Neg*).

4.2.2 Annotation Results

We report the annotation results in Table 2. We find that IntentionQA exhibits high correctness rates among ground truth options. Meanwhile, the low false-negative rates demonstrate the high quality of both tasks. Both statistics validate the reliability of our automatic QA construction pipeline and the quality of the resulting IntentionQA benchmark.

5 Experiments and Analysis

5.1 Baseline Selection and Setup

Evaluation Metric We use accuracy as the evaluation metric, which is quantified by the percentage of QA pairs that a language model answers correctly in IntentionQA.

Model Selection We evaluate a wide range of (L)LMs in four categories: **(1) PTLM:** We evaluate several pre-trained language models, including RoBERTa (Liu et al., 2019), DeBERTa-v3 (He et al., 2023), T0 (Sanh et al., 2022), T5 (Raffel et al., 2020), and Flan-T5 (Chung et al., 2022). **(2) Commonsense:** We also evaluate PTLMs with commonsense knowledge injected, including HyKAS (Ma et al., 2021), CAR (Wang et al., 2023a), VERA (Liu et al., 2023b), CANDLE (Wang et al., 2024b), and VERA-CANDLE (Wang et al., 2024b). **(3) Open LLM:** We then evaluate representative open-sourced LLMs of varying sizes and versions in zero-shot settings as well as after fine-tuning on intention knowledge (Open LLM + Mind, details in §5.5). These models cover LLaMA2 (Touvron et al., 2023), Gemma (Mesnard et al., 2024), Mistral (Jiang et al., 2023), Falcon (Almazrouei et al., 2023), Vicuna (Zheng et al., 2023), Phi-2 (Gunasekar et al., 2023), and Alpaca (Taori et al., 2023; Wang et al., 2023d). **(4) LLM API:** Finally, we adopt Chain-of-Thought prompting (CoT; Wei et al., 2022) and CoT with Self-Consistency (CoT-SC; Wang et al., 2023c) together with zero-shot prompting to assess ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023). The sampling temperature τ is set to 0.1 by default. 5 CoT responses are sampled with τ set to 0.7 under CoT-SC. Random and Majority voting are also added as baselines to demonstrate the characteristic of IntentionQA. Human performance is calculated based on annotation results of 600 randomly selected QA pairs from both tasks. (See prompts in Table 6 and 7)

5.2 Results

The results of all models are presented in Table 3. From the results, we observe that:

Commonsense knowledge does help in intention comprehension. Models injected with commonsense knowledge showcase comparable performance to significantly larger models. Specifically, CAR and CANDLE (435M) achieve 96.64% of the performance of Flan-T5-xxl (11B) in Inten-

Methods	Backbone	IntentUnderstand				IntentUtilize			
		Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.
Random	-	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Majority Vote	-	26.37	25.24	26.27	25.00	26.09	28.57	28.57	26.60
PTLM	RoBERTa-Large *214M*	41.46	41.98	38.98	41.43	54.95	35.06	30.08	49.84
	DeBERTa-v3-Large *435M*	36.40	38.72	37.62	36.90	26.52	29.35	32.33	27.39
	T5-v1.1-xxl *11B*	24.84	25.47	23.91	24.92	26.71	26.23	25.56	26.55
	Flan-T5-xxl *11B*	75.98	73.58	75.00	75.48	79.26	81.82	81.95	79.89
	T0-pp *11B*	71.70	68.87	69.57	71.07	77.11	76.10	78.20	76.99
Commonsense	HyKAS *435M*	71.81	67.17	46.69	69.61	47.02	45.97	48.12	46.90
	CAR *435M*	73.69	71.46	54.38	72.20	36.18	43.12	44.36	37.94
	CANDLE *435M*	74.34	70.75	52.54	72.52	35.94	43.90	43.61	37.84
	VERA *11B*	69.82	70.52	61.02	69.49	59.20	58.18	64.66	59.36
	VERA-CANDLE *11B*	70.59	71.33	63.41	70.02	62.18	60.13	66.13	61.81
Open LLM	LLaMA2-7B	22.47	26.24	22.37	23.18	26.42	27.87	29.03	26.84
	LLaMA2-7B-chat	64.98	66.54	57.38	64.93	59.90	54.86	47.37	58.04
	LLaMA2-13B	24.21	27.70	25.00	24.91	27.92	30.59	28.03	28.40
	LLaMA2-13B-chat	69.63	63.96	62.50	68.21	45.53	41.95	39.71	44.52
	Gemma-2B	21.73	23.87	19.51	22.08	30.66	30.63	30.99	30.67
	Gemma-2B-instruct	48.77	47.23	53.41	48.67	39.45	39.15	38.17	39.32
	Gemma-7B	50.94	50.86	47.19	50.77	26.75	30.19	31.20	27.65
	Gemma-7B-instruct	65.55	64.31	61.04	65.13	33.18	36.01	41.51	34.20
	Mistral-7B-instruct-v0.1	53.49	55.04	53.64	53.80	26.18	28.27	28.57	26.70
	Mistral-7B-instruct-v0.2	76.57	74.53	72.83	76.03	59.78	62.60	65.41	60.64
	Falcon-7B	24.19	20.52	25.00	23.52	25.40	25.45	27.82	25.56
	Falcon-7B-instruct	24.54	22.17	28.26	24.25	26.15	28.05	26.32	26.50
	Vicuna-7B-v1.5	57.13	57.08	55.43	57.05	27.88	30.13	23.31	28.00
	Phi-2 *3B*	33.24	37.97	33.70	34.16	26.71	28.57	28.57	27.16
	Alpaca-LLaMA-7B	48.97	46.93	44.57	48.40	50.15	46.49	37.59	48.72
Open LLM + MIND	LLaMA2-7B-chat	65.78	64.61	55.75	66.15	59.43	57.13	60.03	59.04
	Mistral-7B-instruct-v0.2	78.57	74.31	80.89	76.97	61.14	65.42	62.16	62.02
LLM API	ChatGPT	75.06	73.76	77.17	74.90	80.74	76.62	68.42	79.23
	ChatGPT (CoT)	76.07	74.53	72.83	75.64	78.89	75.32	78.20	78.21
	ChatGPT (CoT-SC)	76.51	73.82	71.74	75.80	85.72	77.14	82.71	83.99
	GPT-4	78.12	**75.41**	73.91	77.43	**86.03**	**82.34**	**84.96**	**85.30**
	GPT-4 (CoT)	77.43	73.11	**80.43**	76.73	83.57	79.74	82.71	82.83
	GPT-4 (CoT-SC)	**78.80**	72.88	75.00	**77.51**	84.00	80.78	**84.96**	83.48
Human	-	89.96	90.00	100.00	90.67	95.50	85.19	100.0	94.00

Table 3: Evaluation results (Accuracy%) of various language models on both tasks of the IntentionQA benchmark. The best performances within each category are underlined and the best among all baselines are **bold-faced**.

tUnderstand, despite being 25 times smaller. This demonstrates the effectiveness of incorporating commonsense knowledge in improving intention comprehension in the E-commerce domain.

IntentUtilize is more challenging. For approximately all models, excluding ChatGPT and GPT-4, that exhibit above Random performances in IntentUnderstand, their performances drop significantly when evaluated on IntentUtilize, with an average accuracy gap of 14.20%. While IntentUnderstand involves understanding the purchase intention behind a single pair of products, IntentUtilize requires product understanding of all candidate options as well as reasoning with potential intentions behind four pairs of products. This expanded reasoning scope and higher demand for product understanding pose challenges for these models, as their training data may be limited in terms of the variety and quantity of products included. However, ChatGPT and GPT-4 excelled in both tasks, presumably due to their stronger product reasoning abilities.

Intention comprehension abilities of current models are still far from perfect. Although various models perform considerably better than Random guessing, there remains a substantial gap between their performance and that of humans.

5.3 Performances Across Intention Types

To further investigate the reasons why language models fail in intention comprehension, we conduct a more fine-grained analysis by delving into

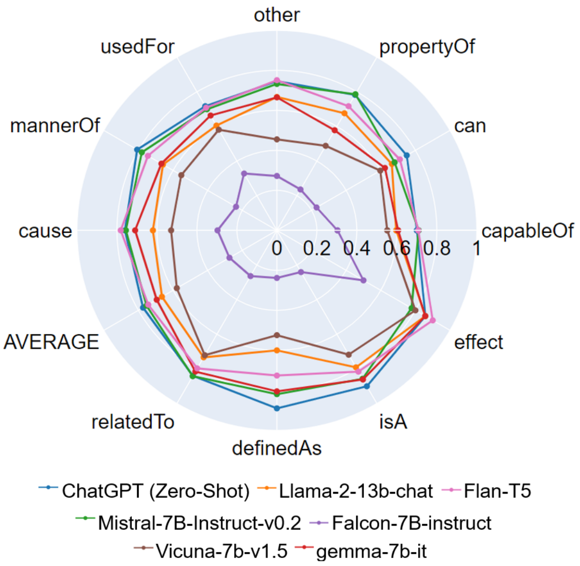

Figure 3: Performances of various language models in comprehending intentions with different relations.

intentions with different commonsense relations grounded in ConceptNet (Speer et al., 2017). Specifically, we construct a sibling QA set using our proposed pipeline, with the only additional constraint being that the distractor options share the same relation type as the ground truth option. From the results presented in Figure 3, all the evaluated language models are more effective in understanding the product definition, with an average of 70.47% across relations `isA`, `definedAs`, and `relatedTo`. However, a performance decline of 6.69% is observed in relations that require a deeper understanding of the cause and effect behind the purchasing event, such as `capableOf` and `cause`.

5.4 Error Analysis

In this section, we randomly sample 120 questions that GPT-4 answers incorrectly from INTENTIONQA and categorize the errors by asking experts to annotate them manually. (Details in Appendix A.2)

Among 60 annotated error samples from INTENTUNDERSTAND, we found:

- 56.7% errors are caused by failing to identify the most typical intention, e.g., choosing "because the product is of good quality" instead of "because the person wants to build a water cooling system."
- 18.3% errors are due to overarching inference. The selected options, while seemingly plausible, cannot be deduced from the products provided.

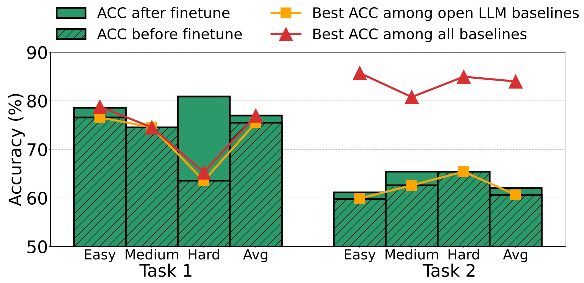

Figure 4: Comparisons between models fine-tuned on intentions from MIND and baseline models achieving top performances.

- 11.7% errors are due to selecting implausible options. The model selects an option that is irrelevant to the given products or implausible. Cases where the rationales in COT responses are irrelevant to selected options are also observed.
- 13.3% errors are due to inaccurate understanding of the given products.

Among 60 annotated error samples from INTENTUTILIZE, we found:

- 40% errors are due to inaccurate understanding of the given intention. For example, the model chooses "iPod" under the intention "because the customer wanted to use them *with* his/her iPod".
- 38.3% errors are due to inaccurate understanding of the given products. The reasoning in their response demonstrates inaccurate understanding of the purchased products or those in the options. Or, when the intention is not typical enough to filter out distractors, they fail to rely more on the purchased product to select the best option.
- 21.7% errors are due to false-negative distractors or incorrect ground truth answers.

5.5 Transferring from Other Sources

In this section, we explore whether transferring intention knowledge from other sources can aid the model's performance via fine-tuning. Specifically, we use MIND, a knowledge base constructed by Xu et al. (2024), besides FolkScope, as a rich source of purchase intentions. MIND is a multi-modal intention knowledge base distilled from LLaVA (Liu et al., 2023a), which includes product images in the knowledge generation process. To ensure the quality of generated intentions, a human-centric intention filtering module is developed to eliminate implausible and atypical intentions.

To incorporate MIND's intention knowledge, we convert 4,059 sets of co-buy records and their corresponding intentions into an instruction-tuning format. We then fine-tune the LLaMA2-7B-chat and Mistral-7B-instruct-v0.2 models on this data using LoRA (Hu et al., 2022). The results, reported in OPEN LLM + MIND of Table 3, reveal an average performance gain of 1.51% and 1.19% for two tasks respectively.

Next, we compare the performance of the fine-tuned Mistral-7B-instruct-v0.2 model with the highest accuracy achieved by all OPEN LLMs and all baselines. The trends are shown in Figure 4. Notably, fine-tuning enables Mistral-7B-instruct-v0.2 to achieve performance comparable to that of GPT-4 in INTENTUNDERSTAND. However, INTENTUTILIZE remains a challenging task even after fine-tuning. This disparity suggests that fine-tuning with intention knowledge facilitates the acquisition of intention understanding abilities, while improving INTENTUTILIZE performance requires more effort. One possible reason is that INTENTUTILIZE places a higher demand on product understanding and reasoning abilities compared to INTENTUNDERSTAND, which cannot be easily improved by simple knowledge injection.

6 Conclusions

In conclusion, this paper presents INTENTIONQA, a double-task MCQA dataset designed to assess the intention comprehension capabilities of LMs. Extensive experiments and analyses demonstrate that LMs face significant challenges in certain scenarios, trailing far behind human performance levels, while fine-tuning on external resources brings considerable performance gains. We hope our work sheds light on the limitations of current LMs in E-commerce intention understanding tasks and opens up a new paradigm of leveraging LM in E-commerce services.

Limitations

We base the negative distractor sampling on similarity filtering with manually selected thresholds. While these thresholds are decided after multiple rounds of parameter searches and observation of the resulting data quality and have been validated by the human annotation we conduct, automated threshold tuning methods (Xu et al., 2021) could be implemented to facilitate this process.

As we build the dataset based on FolkScope, the quality of the latter is upper-bounded by the former. Nevertheless, the construction pipeline introduced in this work can be generalized to expand the dataset by incorporating other intention knowledge bases. Meanwhile, more advanced LLMs have the potential of curating intention knowledge bases with high quality, further boosting the quality of our QA benchmark.

Since (L)LMs demonstrate strong generative capabilities and commonsense reasoning, it is potentially feasible to leverage models such as ChatGPT to generate contextual information for purchase events. However, we rely on the eventuality knowledge graph, ASER, to facilitate the calculation of context embeddings. This offers advantages in terms of cost control and the potential to scale up. Additionally, the human annotation results of our dataset confirm the effectiveness of leveraging ASER for this purpose.

Our work mainly focuses on intention comprehension in E-commerce, which specifically involves product understanding, purchase intention reasoning, and mental state sharing, as well as commonsense reasoning within the context of intention comprehension in broader domains. While our work studies intention comprehension in a specific domain and holds potential for real-world applications, we believe its findings can offer general insights for broader research.

Ethics Statement

6.1 Offensive Content Elimination

While we adopt LMs in a generative setting, generating harmful or biased content from them is limited as INTENTIONQA is evaluated in multiple-choice question form. In most cases, the language models generate a single letter representing the option. In COT, the LLMs generate a short rationale and then output the final answer, where the rationale is closely related to the question itself.

6.2 Annotations

The annotators are paid a wage higher than our local law, and the expert annotators are graduate students specializing in natural language processing. They have all agreed to participate voluntarily and are well-instructed about the tasks.

6.3 Licenses

FolkScope and ASER are released under the MIT license, which grants us access to the datasets for free. Assets including models and tokenizers obtained from Huggingface Hub [1] are shared via li-

[1] https://huggingface.co/

censes that support research purposes. We will share our code and data under the MIT license, which allows free distribution of our curated assets. All associated licenses permit user access for research purposes, and we have agreed to follow all terms of use.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the general support from Amazon.com Inc.

References

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. *CoRR*, abs/2311.16867.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023. Complex query answering on eventuality knowledge graph with implicit logical constraints. 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*.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

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 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 829–837. ACM.

Yang Deng, Wenxuan Zhang, Qian Yu, and Wai Lam. 2023. Product question answering in e-commerce: A survey. 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 11951–11964. Association for Computational Linguistics.

Wenxuan Ding, Shangbin Feng, Yuhan Liu, Zhaoxuan Tan, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Knowledge crosswords: Geometric reasoning over structured knowledge with large language models. *arXiv preprint arXiv:2310.01290*.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Leonardo Fogassi, Pier Francesco Ferrari, Benno Gesierich, Stefano Rozzi, Fabian Chersi, and Giacomo Rizzolatti. 2005. Parietal lobe: from action organization to intention understanding. *Science*, 308(5722):662–667.

Abhinav Goyal, Anupam Singh, and Nikesh Garera. 2022. End-to-end speech to intent prediction to improve e-commerce customer support voicebot in hindi and english. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, pages 579–586. Association for Computational Linguistics.

Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1809–1818. ACM.

Jing Gu, Qingyang Wu, and Zhou Yu. 2021. Perception score: A learned metric for open-ended text generation evaluation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12902–12910. AAAI Press.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2024. Acquiring and modeling abstract commonsense knowledge via conceptualization. *Artif. Intell.*, 333:104149.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query reformulation in e-commerce search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1319–1328. ACM.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6750–6774. Association for Computational Linguistics.

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. *CoRR*, abs/2310.06825.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, and Xianfeng Tang. 2023. 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*.

Yeongseo Jung, Eunseo Jung, and Lei Chen. 2023. Towards a unified conversational recommendation system: Multi-task learning via contextualized knowledge distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13625–13637. Association for Computational Linguistics.

Dong-Mo Koo and Seon-Hee Ju. 2010. The interactional effects of atmospherics and perceptual curiosity on emotions and online shopping intention. *Computers in Human Behavior*, 26(3):377–388.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 755–764. ACM.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Haitao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18582–18590. AAAI Press.

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

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023b. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1264–1287. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Feihong Lu, Weiqi Wang, Yangyifei Luo, Ziqin Zhu, Qingyun Sun, Baixuan Xu, Haochen Shi, Shiqi Gao, Qian Li, Yangqiu Song, and Jianxin Li. 2024. MIKO: multimodal intention knowledge distillation from large language models for social-media commonsense discovery. *CoRR*, abs/2402.18169.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Mingjie Qian, Yongsen Zheng, Jinghui Qin, and Liang Lin. 2023. Hutcrs: Hierarchical user-interest tracking for conversational recommender system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10281–10290. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1339–1384. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. QADYNAMICS: training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15329–15341. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. 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 14820–14835. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Sheikh Sarwar, Chen Luo, Yang Laurence Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, and Yangqiu Song. 2024a. EcomScript: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association. *CoRR*.

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, Jiayang Cheng, 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*. 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. 2024c. On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions. *CoRR*, abs/2406.10885.

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 and Yangqiu Song. 2024. MARS: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *CoRR*, abs/2406.02106.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language models with self-generated instructions. 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 13484–13508. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Bing Yin, Long Chen, and Yangqiu Song. 2024. MIND: multimodal shopping intention distillation from large vision-language models for e-commerce purchase understanding. *CoRR*, abs/2406.10701.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11525–11536. PMLR.

Yang Yang and Jie Tang. 2015. Beyond query: Interactive user intention understanding. In *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pages 519–528. IEEE Computer Society.

Changlong Yu, Xin Liu, Jefferson Maia, Tianyu Cao, Laurence Yang Li, Yifan Gao, Yangqiu Song, Rahul Goutam, Haiyang Zhang, Bing Yin, et al. 2024. Cosmo: A large-scale e-commerce common sense knowledge generation and serving system at amazon. In *Proceedings of the 2024 International Conference on Management of Data, SIGMOD 2024*.

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*, pages 1173–1191. Association for Computational Linguistics.

Qian Yu and Wai Lam. 2018. Product question intent detection using indicative clause attention and adversarial learning. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China, September 14-17, 2018*, pages 75–82. ACM.

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 65–74. ACM.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022a. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artif. Intell.*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Jiashu Zhao, Hongshen Chen, and Dawei Yin. 2019. A dynamic product-aware learning model for e-commerce query intent understanding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1843–1852. ACM.

Wayne Xin Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1935–1944. ACM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. 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*.

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

Wendi Zhou, Tianyi Li, Pavlos Vougiouklis, Mark Steedman, and Jeff Z. Pan. 2024. A usage-centric take on intent understanding in e-commerce. *CoRR*, abs/2402.14901.

Appendices

A Discussion

A.1 Plausibility of leveraging ASER for context augmentation

We decide to use ASER after considering the following factors. First, LLMs accessed via API such as ChatGPT or GPT-4 will incur high costs to generate the context information of each purchase event and therefore inhibit the potential to scale up. Besides, since LLMs are sensitive to prompts, the similarity computed based on their generated context information may be inconsistent. Secondly, we observe that open-sourced LMs are relatively weak in understanding the purchase events and generating meaningful context information consistently. Such weakness is further confirmed by their low performance in both tasks.

However, since ASER is a unified large-scale eventuality knowledge graph, it enjoys a large knowledge coverage and consistency, the similarity computed based on the context information extracted from ASER is fair and consistent. The human annotation also verifies the effectiveness of using ASER.

A.2 Elaboration on Error Analysis

Overly-general intentions vs. typical intentions Overly general intentions are the intentions applicable to almost all products, such as "because the products are similar to each other", "because they are high quality products", "because they both are a type of product that he needed". These intentions are too general and thus lack meaningful specificity.

On the other hand, examples such as "because they both are capable of producing high quality sound", "because they are both used for his laptop", "because they both are capable of cooling his CPU" are more typical intentions and could be more helpful in real-world applications.

Selection of implausible intentions We observe from the errors that the rationale generated with CoT is neither relevant to the options provided nor to the answer the model provides. This may indicate that LLM is distracted by some other information that shares high correlation with the option and thus fails to reason within the provided scenario. We provide an example of such cases in Tabel 4.

A.3 Quality Control Details

The preliminary human annotation over 300 randomly sampled QA pairs shows that the Correctness rate of the **hard** subset in TASK 1 is 85.71% and the false-negative rate is 4.76%. Therefore, we manually correct this subset by discarding problematic QA pairs. The relatively low correctness rate and high false-negative rate in the hard subset may be because the intention assertions in FolkScope where the two products are less similar suffer from lower quality.

B Implementation Details

B.1 Hyperparameter Settings

For models in the category of LLM API, we set the `max_tokens` to 10 and 200 respectively for ZERO-SHOT and COT (COT+SC) prompting. The temperature τ is set to 0.1 for non-Self-Consistency decoding and 0.7 otherwise.

For models in the category of OPEN LLM, we use the default setting as presented in Hugging Face.

For fine-tuning LLMs, we use the open-sourced library LLaMA-Factory[2] (Zheng et al., 2024) to train all models. All hyperparameters follow the default settings, and a LoRA rank of $\alpha = 64$ is used. We conduct all experiments on a Linux machine with eight NVIDIA V100 GPUs.

C Annotation Setup

We follow annotation setup of Fang et al. (2021a,b); Shi et al. (2023) and recruit human annotators from the Amazon Mechanical Turk service for human evaluation. For strict quality control, we only invite workers satisfying the following requirements: 1) at least 1K HITs approved, and 2) at least 95% approval rate. We then host two rounds of qualification rounds using questions sampled from our curated benchmark, with expert-annotated answers. 400 workers are invited in total and around 60 (15%) of them are selected.

For product conceptualization, we randomly sample 2,000 products and task each annotator to label the plausibility of the generated categories. Specifically, we ask the annotators to assess whether all three generated product categories are reasonable according to the original products. Each product is annotated by three annotators and the majority vote is taken as the final label.

[2]https://github.com/hiyouga/LLaMA-Factory

Task	Prompt & Response
Task 1	[Prompt] A customer buys 14 Karat Yellow Gold Round Bead Ball Stud Earrings, 2mm and 14 Karat White Gold Round Bead Ball Stud Earrings, 3mm. What is the most likely intention for buying them? A. because they both are capable of being worn for a long time. B. because they both are a type of USB extension cable. C. because the person wants to use them together. D. because they both have a property of anti-glare/anti-fingerprint. Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. Step 2: Answer A or B or C or D only without any other word. [Response from GPT-4] Step 1: The customer is buying two different types of gold earrings, which indicates they are likely purchasing them for personal use or as a gift. The other options do not relate to earrings or their intended use. Step 2: C [Correct answer: A]

Table 4: An example of selection of implausible intention by GPT-4 with CoT prompting. The LLM may be distracted by the property of earrings as being gifts and provide a rationale that is irrelevant to the given options. Although it is hard to use two pairs of earrings together, the LLM fails to select the most likely and plausible intention probably due to its irrelevant rationale.

We then evaluate the resulting QA pairs from INTENTUNDERSTAND and INTENTUTILIZE, we randomly sample 300 QA pairs and ask the annotators to assess the quality of these problems, including the correctness of ground truth options (*Correct*) and assess the false-negativeness of the distractor options by determining whether a distractor option is superior to or equally plausible as the ground truth option (*F-Neg*). Again, we collect three votes for each QA pair and take the majority of them.

D Prompts

We report the prompt used for product simplification with ChatGPT in Tabel 5.

To transfer a conceptualized product into ASER nodes' format, we utilize natural language templates. These templates consist of assembling each product category with a subject and a verb that carry semantic meanings related to purchasing. Specifically, the subjects we use include: `PersonX`,`PersonY`,`PeopleX`,`PeopleY`; Similarly, the verbs we employ are: `buy`, `shop`, `purchase`, `get`, `obtain`, `have`, in simple present tense, original form, simple perfect tense, or past tense, with optional articles (`a`, `an`,`the`,`1`,`2`) added before the conceptualized product name. As a result, when a product such as "iPhone 14" occurs, we transform it into a list of concise yet semantically complete events that can potentially be matched in ASER. For example, one of the transformed events could be "PersonX bought a phone."

We report the prompts used for INTENTUNDERSTAND and INTENTUTILIZE in Table 6 and Table 7 respectively.

E Case Study

We present example questions that GPT-4 successfully answer or fail with COT for both tasks in Table 8.

F Error Analysis examples

We present examples of erroneous responses by GPT-4 with CoT prompting on both tasks for each error type in Table 9 and Table 10.

Method	Prompt for Product Name Simplification
ZERO-SHOT	Product name: <product>; What is the category of the product? Generate three possible categories, each in 2 words, separated by a comma.

Table 5: Prompt used to instruct ChatGPT to conceptualize the product name.

Method	Prompt for INTENTUNDERSTAND
ZERO-SHOT	A customer buys <product 1> and <product 2>. What is the most likely intention for buying them? A. because <intention 1> B. because <intention 2> C. because <intention 3> D. because <intention 4> Answer A or B or C or D only without any other word.
COT	A customer buys <product 1> and <product 2>. What is the most likely intention for buying them? A. because <intention 1> B. because <intention 2> C. because <intention 3> D. because <intention 4> Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. Step 2: Answer A or B or C or D only without any other word.

Table 6: Prompts for INTENTUNDERSTAND with ZERO-SHOT prompting and COT respectively.

Method	Prompt for INTENTUTILIZE
ZERO-SHOT	A customer buys <product>, because <intention>. What is the customer's most probable additional purchase? A. <product 1> B. <product 2> C. <product 3> D. <product 4> Answer A or B or C or D only without any other word.
COT	A customer buys <product>, because <intention>. What is the customer's most probable additional purchase? A. <product 1> B. <product 2> C. <product 3> D. <product 4> Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. Step 2: Answer A or B or C or D only without any other word.

Table 7: Prompts for INTENTUTILIZE with ZERO-SHOT prompting and COT respectively.

Task	Prompt & Response
Task 1	[Prompt] A customer buys Belkin USB A/A Extension Cable, USB Type-A Female and USB Type-A Male (6 Feet) and Cable Matters 3-Pack USB Extension Cable (USB Extender Cable/USB 2.0 Extension Cable) - 3 Feet - Available 3FT - 10FT in Length. What is the most likely intention for buying them? A. because they both are a type of USB extension cable. B. because they both are a manner of "Audio & Video Accessories". C. because they both are related to the topic of "Army Style". D. because they could both be worn as a bracelet. Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. Step 2: Answer A or B or C or D only without any other word. [Response from GPT-4] Step 1: The customer is buying two different USB extension cables, which indicates they need to extend the length of their USB connections. Step 2: A [Correct!]
Task 1	[Prompt] A customer buys Barbra Collection Hawaiian Style Kukui Nut Lei Hand Painted Multi-color Turtle 32 Inches and Barbra Collection Hawaiian Style Kukui Nut Lei Hibiscus Hand Painted Turquoise Flower 32 Inches. What is the most likely intention for buying them? A. because the person wants to replace the LCD screen of the laptop. B. because they both are capable of giving him a lot of benefits. C. because they both are a type of toys. D. because they could both be found on Aliexpress.com. Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. Step 2: Answer A or B or C or D only without any other word. [Response from GPT-4] Step 1: The customer is buying two different styles of Kukui Nut Lei, which are Hawaiian-style accessories. The intention is most likely related to their use as accessories or decorations. Step 2: D [Correct answer: B]
Task 2	[Prompt] A customer buys StarTech.com 60x10mm Replacement Ball Bearing Computer Case Fan w/ TX3 Connector - 3 pin case Fan - TX3 Fan - 60mm Fan, because they both are defined as "Computer Accessories". What is the customer's most probable additional purchase? A. StarTech 8-Inch 24 Pin ATX 2.01 Power Extension Cable (ATX24POWEXT) B. Genuine Power On-Off On/Off On Off Flex Cable Ribbon FOR Apple iPad 2 iPad2 C. WiFi Bluetooth Antenna Flex Cable Replacement Part for IPad 3 and Ipad 4 D. Headphone Audio Jack Flex Cable Ribbon Replacment for Ipad Air (Black) Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. Step 2: Answer A or B or C or D only without any other word. [Response from GPT-4] Step 1: The customer has purchased a computer case fan, which indicates that they are likely working on a desktop computer or building one. Therefore, the most probable additional purchase would be related to desktop computer components or accessories. Step 2: A [Correct!]
Task 2	[Prompt] A customer buys Nakamichi NK2000 Over-the-Head Headphones | White, because he wanted to use them with his iPod. What is the customer's most probable additional purchase? A. Nakamichi Amplified Stereo Headphones NK2000 Turbulence B. Apple iPod nano 8 GB Silver (4th Generation) (Discontinued by Manufacturer) C. Iwoo Bluetooth Headphone, Blue D. Portable CD Cassette Boombox with AM/FM Radio Formulate your answer in this way. Step 1: Give a short and brief rationale by thinking step by step. [Response from GPT-4] Step 2: Answer A or B or C or D only without any other word. Step 1: The customer already bought headphones for his iPod, so he is unlikely to buy another pair of headphones or a different audio device. He might need an accessory that is compatible with his iPod. Step 2: B [Correct answer: A]

Table 8: Example prompts and responses from GPT-4 with CoT prompting methods.

Error Type	Description	Example
Atypical	The selected option is not the most typical intention given the products.	[Question] A customer buys Invicta Men's 1515 I Force Collection 18k Gold Ion-Plated Watch with Black Cloth-Covered Band and Invicta Men's 16011 S1 Rally; 18k Gold Ion-Plated Watch with Brown Leather Strap. What is the most likely intention for buying them? A. because they both are a manner of the same brand. B. because they both are related to watches. C. because they could both be used to tell time. D. because they both are capable of providing him with the comfort and support he needs. [Response] Step 1: The customer is likely buying both watches because they are from the same brand and can be used to tell time, but they have different styles (one with a black cloth-covered band and the other with a brown leather strap). Step 2: A [Correct answer: C]
Overrarching Inference	The selected options, while seemingly plausible, cannot be deduced from the products provided.	[Question] A customer buys KabelDirekt (20 feet) CL3 HDMI Cable (1080p 4K 3D High Speed with Ethernet ARC rated for in wall installation) - TOP Series and BlueRigger In-Wall High Speed HDMI Cable - 15 Feet - CL3 Rated - Supports 4K, Ultra HD, 3D, 1080p, Ethernet and Audio Return (Latest Standard). What is the most likely intention for buying them? A. because the person wants his home theater system to be better. B. because the person wants to record his voice and video. C. because the person wants to connect the device to the TV. D. because they both are defined as "Network Cable" and "Network Adapter" in the same category. [Response] Step 1: The customer is buying two different HDMI cables, both of which are rated for in-wall installation and support high-quality video and audio transmission. This suggests that they are looking to improve their home theater system by connecting devices to their TV. Step 2: A [Correct answer: C]
Implausible	The model selects an option that is irrelevant to the given products or implausible. Cases where the rationales in CoT responses are irrelevant to selected options are also observed.	[Question] A customer buys 14 Karat Yellow Gold Round Bead Ball Stud Earrings, 2mm and 14 Karat White Gold Round Bead Ball Stud Earrings, 3mm. What is the most likely intention for buying them? A. because they both are capable of being worn for a long time. B. because they both are a type of USB extension cable. C. because the person wants to use them together. D. because they both have a property of anti-glare/anti-fingerprint. [Response] Step 1: The customer is buying two different types of gold earrings, which indicates they are likely purchasing them for personal use or as a gift. The other options do not relate to earrings or their intended use. Step 2: C [Correct answer: A]
Inaccurate Product Understanding	The selected options do not align well with the product.	[Question] A customer buys Sony Clip-on Stereo Headphones With Double Retractable Cord | MDR-Q68LW L Blue (Japanese Imports) and Sony Clip-on Stereo Headphones With Double Retractable Cord | MDR-Q68LW T Brown (Japanese Imports). What is the most likely intention for buying them? A. because they both are defined as ""Headset"" in the product name. B. because the person wants to listen to music or watch a movie. C. because they both are capable of being worn with a variety of outfits. D. because they both are capable of blocking out noise. [Response] Step 1: The customer is buying two different colors of the same product, which indicates they might want to match them with different outfits or have options for style. Step 2: C [Correct answer: B]

Table 9: Examples of erroneous responses by GPT-4 with CoT prompting methods on INTENTUNDERSTAND.

Error Type	Description	Example
Inaccurate Intention Understanding	The selected options do not align with the given intention.	[Question] A customer buys Nakamichi NK2000 Over-the-Head Headphones | White, because he wanted to use them with his iPod. What is the customer's most probable additional purchase? A. Nakamichi Amplified Stereo Headphones NK2000 Turbulence B. Apple iPod nano 8 GB Silver (4th Generation) (Discontinued by Manufacturer) C. Iwoo Bluetooth Headphone, Blue D. Portable CD Cassette Boombox with AM/FM Radio [Response] Step 1: The customer already bought headphones for his iPod, so he is unlikely to buy another pair of headphones or a different audio device. He might need an accessory that is compatible with his iPod. Step 2: B [Correct answer: A]
Inaccurate Product Understanding	The reasoning in their response demonstrates inaccurate understanding of the purchased products or those in the options. Or, when the intention is not typical enough to filter out distractors, they fail to rely more on the purchased product to select the best option.	[Question] A customer buys Lapis ID Necklace, because they could both be used as a necklace. What is the customer's most probable additional purchase? A. The Lord Fashion Jewelry Arwen Evenstar Twilight Star Pendant Necklace B. Sterling Silver Antiqued Tennessee State Charm C. Juvel Jewelry Women Fashion 18K Gold Plated Stud Earrings Elegant Cubic Zircon Ring Style With Diamond For Party D. Ms. Sophisticated ID Necklace Formulate your answer in this way. [Response] Step 1: The customer is most likely to purchase another item that can also be used as a necklace and complements their initial purchase. Step 2: A [Correct answer: D]

Table 10: Examples of erroneous responses by GPT-4 with CoT prompting methods on INTENTUTILIZE.