Intention Knowledge Graph Construction for User Intention Relation Modeling

Anonymous ACL submission

Abstract

Understanding user intentions is challenging for online platforms. Recent work on intention knowledge graphs addresses this, but often lacks focus on connecting intentions, which is crucial for modeling user behavior and predicting future actions. This paper introduces a framework to automatically generate an intention knowledge graph, capturing connections between user intentions. Using the Amazon m2 dataset, we construct an intention graph with 351 million edges, demonstrating high plausibility and acceptance. Our model effectively predicts new session intentions and enhances product recommendations, outperforming previous state-of-the-art methods and showcasing the approach's practical utility.[1]

1 Introduction

Understanding user intentions is a fundamental challenge for online platforms seeking to optimize user experience. Recent advances in intention knowledge graphs have made significant strides in connecting user behaviors to their underlying intentions. These graphs link various behaviors, such as co-purchasing patterns (Yu et al., 2023) and search queries (Yu et al., 2024), to intentions expressed in natural language, proving valuable for applications ranging from product recommendation to search relevance optimization.

However, a critical limitation of existing approaches lies in their inability to model the relationships between different intentions. Consider a user preparing for Halloween—they may have multiple interconnected intentions: dressing as a vampire, creating spooky decorations, and hosting a party. While their browsing history (Figure 1 (A)) implicitly suggests these intentions, existing knowledge graphs fail to capture how these intentions relate to each other through temporal, causal, and conceptual links.

This gap aligns with Bengio's distinction between System I and System II reasoning, where the latter emphasizes deliberate, logical, and sequential thinking. Research demonstrates that online shopping behaviors frequently reflect System II reasoning (Kleinberg et al., 2022), with users making conscious decisions toward long-term goals rather than engaging in unconscious browsing. By explicitly modeling intention relationships, we can better capture the deliberate nature of user behavior and create more accurate predictive models.

Our research extends beyond merely understanding users' initial intentions to predicting subsequent ones, for instance, inferring that someone who intends to buy a desk may soon need an office chair. This predictive capability has profound implications for user behavior modeling, session understanding, and recommendation systems. However, constructing these intention-to-intention relationships remains unexplored in current research.

We identify two key mechanisms for modeling these relationships. First, commonsense knowledge is essential for understanding how intentions connect—planning a Halloween party naturally entails both costume selection and decoration preparation. We propose leveraging inferential commonsense relations to describe temporal sequences (before/after) and causal connections (because/as a result) between intentions. After identifying intentions from user sessions, we employ a classifier to determine these inferential connections, building a more comprehensive model of user intention dynamics. Figure 1 (B) illustrates this approach, showing the plausible co-occurrence relationship between costume and decoration intentions.

Second, we incorporate abstract conceptualization to improve generalization beyond observed intention pairs. Recent work demonstrates that

[1] We will release all data and code for this paper after this paper is accepted.

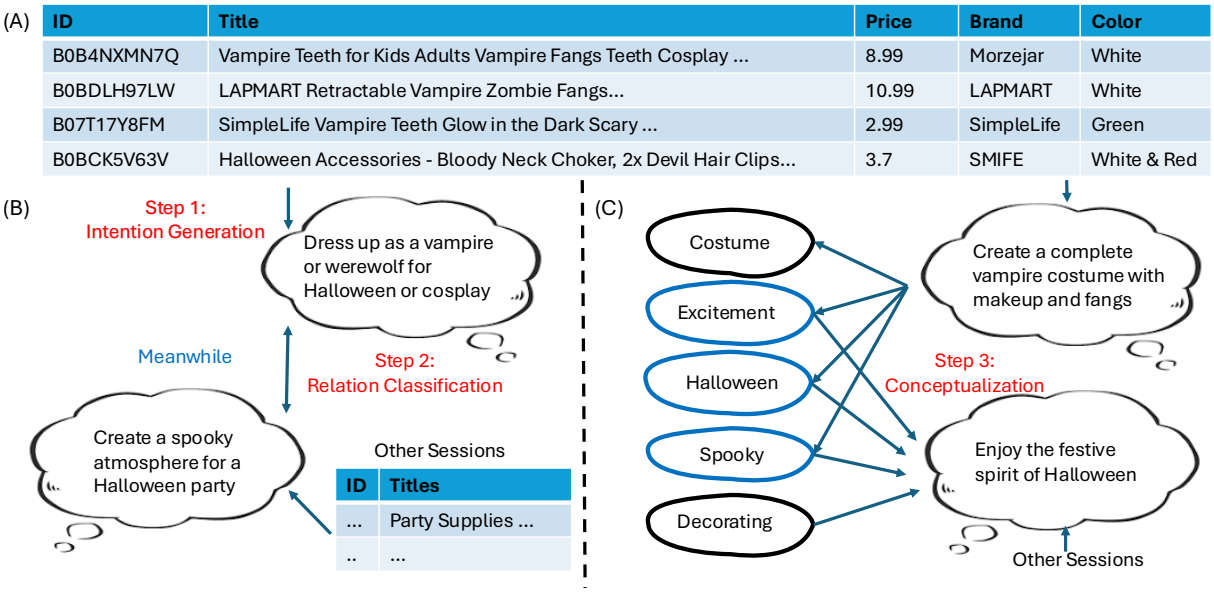

Figure 1: The structure of our knowledge graph. Part (A) shows an example of user behaviors within a session. Our knowledge graph emphasizes building relations between different intentions. In Part (B), we establish commonsense relations, highlighting the temporality and causality of intentions. Part (C) focuses on using conceptualization to connect various intentions.

conceptualization and instantiation enhance commonsense reasoning (Wang et al., 2023b,a). We integrate this insight by using models to abstract intentions into broader concepts, establishing connections between semantically related intentions. For example, conceptualizing a "vampire costume" intention involves higher-level concepts like "costumes" and "Halloween," linking it to other costume-related intentions (Figure 1 (C)).

To address these challenges, we introduce the **Intention Generation, Conceptualization, and Relation Classification (IGC-RC)** framework for constructing a comprehensive commonsense knowledge graph of user intentions. Our approach follows three key steps: **(1) Intention Generation:** We leverage large language models to generate plausible intentions from user session data, capturing the underlying goals driving observable behaviors. **(2) Conceptualization:** We abstract these intentions into higher-level concepts, facilitating connections between semantically related intentions. **(3) Relation Classification:** We generate and verify commonsense statements describing relationships between intentions, creating a structured graph of interconnected user goals.

We apply the IGC-RC framework to the Amazon M2 session-based recommendation dataset, constructing a Relational Intention Graph (RIG) with 351 million edges. This graph captures rich intention-level relationships and demonstrates high plausibility in human evaluations. Our extensive experiments show that RIG enables accurate prediction of intentions in new user sessions and significantly enhances the performance of session-based recommendation models, outperforming previous state-of-the-art approaches.

Our contributions include:

- We pioneer the modeling of connections between users' deliberate mental processes through an intention knowledge graph that captures temporal, causal, and conceptual relationships between intentions.
- We develop the IGC-RC framework, a novel methodology that integrates user behavior data with large language models to automatically construct rich, multi-faceted intention knowledge graphs.
- We construct RIG, a large-scale, high-quality intention knowledge graph from the Amazon M2 dataset, demonstrating superior performance in session understanding and recommendation tasks.

By bridging the gap between discrete user behaviors and their interconnected intentions, our work represents a significant advancement in user modeling for e-commerce and recommendation systems, with broad implications for creating more responsive and intuitive online platforms.

KG	# Nodes	# Edges	# Rels	Sources	Node Type	Intention Relations	User Behavior
ConceptNet	8M	21M	36	Croudsource	concept	✓	✗
ATOMIC	300K	870K	9	Croudsource	situation, event	✓	✗
AliCoCo	163K	813K	91	Extraction	concept	✗	search logs
AliCG	5M	13.5M	1	Extraction	concept, entity	✗	search logs
FolkScope	1.2M	12M	19	LLM Generation	product, intention	✗	co-buy
COSMO	6.3M	29M	15	LLM Generation	product, query, intention	✗	co-buy search-buy
RIG (Ours)	4.2M	351M	6	LLM Generation	product, session, intention, concept	✓	session item history

Table 1: This table shows the details and differences between different commonsense knowledge graphs. Our graph contains six distinct types of edges, including three types of intention-to-intention relationships: asynchronous (before/after), synchronous (at the same time), and causality (because/as a result) among intentions, and item-to-session, session-to-intention, and intention-to-concept connections, summing up to six edge types.

# Sessions	# Concepts	# Intentions	# Ses.-Int.	# Int.-Con.	# Int.-Int.	# Nodes	# Edges
1,176,296	110,741	2,956,195	5,115,587	5,115,212	341,649,216	4,243,232	351,880,015

Table 2: The overall statistics of our constructed knowledge graph RIG, including sessions, intentions, and concepts. Our knowledge graph includes 351 million edges.

2 Related Work

E-commerce platforms increasingly rely on knowledge graphs (KGs) for personalized experiences. Notable examples include Amazon Product Graph (Zalmout et al., 2021), which aligns with Freebase (Bollacker et al., 2008), and Alibaba's ecosystems: AliCG (Zhang et al., 2021), AliCoCo (Luo et al., 2020), and AliMeKG (Li et al., 2020). While these KGs effectively represent item properties, they typically miss user intentions.

Recent work by FolkScope (Yu et al., 2023) and COSMO (Yu et al., 2024) incorporates user intentions through large language models, but lacks inter-intention relationships crucial for modeling goal-oriented behavior. Commonsense knowledge graphs like ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019; Hwang et al., 2021), Discos (Fang et al., 2021), and WebChild (Tandon et al., 2017) structure general knowledge but don't address e-commerce-specific intentions.

Session-based recommendation has evolved from sequence modeling approaches using RNNs and CNNs (Hidasi et al., 2015; Li et al., 2017; Liu et al., 2018; Tang and Wang, 2018) to Graph Neural Networks (Li et al., 2021a; Guo et al., 2022; Huang et al., 2022). Wu et al. (2019a) pioneered GNNs for modeling session transitions, with subsequent improvements incorporating contextual information (Pan et al., 2020; Xia et al., 2021). However, these methods often simplify user intentions, whereas our approach models them explicitly within their relational context.

3 IGC-RC Framework

We introduce the Intention Generation, Conceptualization, and Relation Classification (IGC-RC) framework for constructing comprehensive intention knowledge graphs from user behavior data. Our approach integrates large language models with structured knowledge representation techniques to build a rich understanding of user intentions and their relationships.

3.1 Intention Generation

We leverage the Amazon M2 dataset (Jin et al., 2023), specifically its English subset containing 1.2 million sessions. For each session, we extract product attributes including titles, descriptions, specifications, and technical details to create a comprehensive representation of user browsing history.

Using GPT-3.5, we generate 4.3 million diverse user intentions from these session histories. Our approach improves upon previous work by employing a more capable language model and removing restrictive relation constraints, resulting in more natural and contextually appropriate intentions. The prompting strategy (detailed in Appendix Figure 3) focuses on extracting concise verb phrases that capture user goals underlying their browsing behavior. Quantitative analysis in Appendix Figure 11 confirms significantly higher n-gram diversity compared to Folkscope because of using a better generation model and removal of conceptnet relation constraints.

3.2 Intention Relation Classification

Creating meaningful connections between intentions requires understanding their complex logical, temporal, and causal relationships. We develop a three-step approach to establish these connections: **(1) Template-based assertion generation:** We transform intention pairs into natural language

Intention	Concepts
Purchase a construction dump truck toy for a 2-year-old boy or girl.	playtime, construction, gift
Enjoy multiplayer gameplay with friends and family	socializing, competition, and fun.
Personalize their drawstring bags.	personalization, gift, accessorizing
Perform sanding and grinding tasks on large surfaces using an orbital sander.	smoothing, surface prep, precision
Improve their precision cutting skills	precision, sharpness, craftsmanship

Table 3: This table maps user intentions to relevant concepts. Each intention is analyzed to highlight the core concepts, showcasing how these insights can inform personalized recommendation systems.

	Precedence			Succession			Simultaneous			Cause			Result			Overall		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Vera	0.73	0.52	0.61	0.78	0.66	0.72	0.75	0.83	0.79	0.75	0.73	0.74	0.76	0.92	0.83	0.75	0.73	0.74
+ Fine-tuning	**0.86**	**0.97**	**0.91**	**0.88**	**0.92**	**0.90**	**0.87**	**0.93**	**0.90**	**0.84**	**0.91**	**0.88**	**0.81**	**0.99**	**0.89**	**0.85**	**0.94**	**0.89**

Table 4: The performance of the VERA classifier on the annotated relation classification task.

assertions using templates that express five relationship types: precedence, succession, simultaneity, cause, and result (examples in Table 12). **(2) Plausibility estimation:** We employ the Vera model (Liu et al., 2023a), fine-tuned on expert-annotated data, to assess the plausibility of these assertions. The annotation process followed the Penn Discourse Treebank 2.0 guidelines (Prasad et al., 2008), with two experts independently evaluating each sample (85% initial agreement). (3) **Threshold-based edge selection:** We retain only high-confidence relationships (Vera score > 0.9) to ensure graph quality. The fine-tuned model achieves 0.89 F1 score overall (Table 4), with particularly strong performance on temporal relationships.

This approach enables us to model diverse intention relationships including temporal sequence (before/after), co-occurrence (simultaneously), and causality (because/results in).

3.3 Intention Conceptualization

To improve generalization and abstraction, we develop techniques to map specific intentions to broader concepts. Building on theoretical foundations from knowledge representation (Gruber, 1993; Himanen et al., 2019) and recent advances in conceptualization (Wang et al., 2023b, 2024), we create a novel approach specifically for intentions.

We employ Meta-Llama-3-8B-Instruct with carefully designed prompts (Appendix Figure 4) to generate concise, non-redundant concept sets for each intention. The prompt emphasizes three key qualities: (1) representativeness: concepts must accurately capture the intention's essence; (2) unambiguity: concepts should have clear, focused meanings; and (3) complementarity: concepts should cover different semantic aspects of the intention.

Table 3 demonstrates the effectiveness of this approach, showing how diverse intentions are mapped to concise concept sets that capture their essential characteristics. These concepts serve as connective tissue in our knowledge graph, enabling us to link semantically related intentions that might otherwise appear unrelated based on surface features.

By integrating these three components, intention generation, relation classification, and conceptualization, our IGC-RC framework creates a rich, structured representation of user intentions and their relationships.

Knowledge Graph	Plausibility	Typicality
FolkScope (Before Filter)	0.6116	0.4491
RIG (Ours)	**0.9552**	**0.6674**
+ ConceptNet Rels	0.9542	0.6430

Table 5: The intention generation quality comparison with FolkScope. The comparison is **only** used to show the quality of intention generation, rather than the overall quality of two knowledge graphs.

4 Intrinsic Evaluation

In this section, we evaluate the quality of the generated knowledge graph using both crowdsourced human annotations and automatic evaluations.

4.1 Human Evaluation

Intention Generation Annotators evaluate two aspects of the generated intentions: plausibility and typicality. **Plausibility** refers to the likelihood that an assertion is valid based on its properties, usages, and functions. **Typicality** measures how well an assertion reflects specific features influencing user behavior, such as informativeness and causality. For example, "they are used for Halloween parties" is more informative than "they are used for the same purpose." We collected annotations

Knowledge Graph	Acceptance	Size
Atomic2020	**86.8**	0.6M
Atomic10x	78.5	6.5M
Atomic NOVA	-	2.1M
RIG (ours)	81.2	**341.6M**
- Asynchronous Relation	80.6	100.2M
- Synchronous Relation	82.8	112.6M
- Causality Relation	80.4	128.7M

Table 6: This table presents a comparative analysis of our proposed knowledge graph with existing commonsense knowledge graphs in terms of relation plausibility acceptance and scale of relation edges. As evaluated by human annotators, the acceptance rate represents the percentage of plausible relations in each knowledge graph. The size of each knowledge graph is measured in millions (M) of relation edges.

for 3,000 session-intention pairs, each evaluated by three annotators. The inter-annotator agreement scores were 0.91 for plausibility and 0.74 for typicality. We use *the same* annotation guidelines and criteria from the FolkScope paper, and this can ensure the same standard of annotations on plausibility and typicality. As shown in Table 5, our RIG surpasses FolkScope (Yu et al., 2023) in plausibility and typically. Moreover, as shown in Table 12, the model successfully captures particular, long-tail intentions such as "Relieve discomfort and soothe itching caused by hemorrhoids" and "Purchase unscented baby wipes for sensitive skin." These examples illustrate the model's ability to understand and articulate context-specific user needs. Meanwhile, adding ConceptNet relations as constraints is not useful for plausibility also negatively impacts the intention typicality.

Intention Relation Classification Annotators rate the plausibility of predicted intention relations on a four-point scale: Plausible, Somewhat plausible, Not plausible, and Not applicable. The first two are deemed acceptable. For intention relation classification, three annotators independently evaluated 1,000 intention-intention discourse pairs, achieving an overall inter-annotator agreement of 0.69. As shown in Table 6, our graph achieves an acceptance rate of 81.2% with a significantly larger scale than previous graphs. Atomic2020 has a higher acceptance rate due to manual creation and annotation.

Intention Conceptualization In this task, annotators evaluate the correctness of conceptualized intentions derived from user sessions. We sample 1,000 intention-concept pairs and use Amazon Mechanical Turk for annotation. The conceptualization performance is 86.60%, with an inter-annotator agreement of 77.19%.

	MRR	Hit@1	Hit@3	Hit@10	Inf. Time
Llama3-8B	0.4680	0.4062	0.4480	0.5879	4,102.92ms
Mistral-7B	**0.5544**	**0.4954**	0.5425	0.6763	3,625.63ms
Flan-T5	0.1575	0.0528	0.1295	0.3642	2,021.36ms
RIG (ours)	0.5377	0.4483	**0.5470**	**0.7260**	**3.01ms**

Table 7: The performance on intention prediction.

4.2 Automatic Evaluation

In this section, we systematically evaluate three key aspects: (1) Intention Prediction; (2) Conceptualization of New Intentions; and (3) Item Recovery. These experiments are termed "automatic evaluation" as they can uniformly be regarded as inductive knowledge graph completion tasks.

Intention Prediction This study evaluates the task of linking unseen sessions to an existing intention graph. The dataset, comprising all session-intention edges, was split into training, validation, and test sets (8:1:1) based on sessions, with each session linked to 2-4 positive intentions. For evaluation, we applied random negative sampling to generate 30 candidate intentions per session, ranking the positive intentions among them. We compared large language models (Mistral-7B-Instruct-v0.1, Meta-Llama-3-8B-Instruct, and flan-t5-large) to our RIG model. Baselines use perplexity scores derived via proper prompting to rank intentions, while RIG employs an embedding-based approach. Specifically, we generate embeddings for sessions and intentions using a sentence model and use SASRec as a session encoder to compute session representations. As shown in Table 7, our RIG model achieves competitive accuracy (e.g., best Hit@10) while significantly outperforming LLMs in inference speed (3.01ms vs. 2,021–4,102ms), demonstrating its practicality for real-world applications requiring fast decision-making.

Conceptualization Prediction To evaluate the performance of conceptualization prediction, we constructed a dataset of intention-concept pairs derived from our intention knowledge graph. This dataset was split into training, validation, and test sets using an 8:1:1 ratio. The test set consisted of 147,801 intention-concept pairs, which were used to perform ranking tasks with various methods. For the baseline models, we employed large language models (LLMs) to perform

	MRR	Hit@1	Hit@3	Hit@10	Inf. Time
Llama3-8B	0.3224	0.3023	0.3192	0.3449	2,069.96ms
Mistral-7B	0.1110	0.0894	0.1001	0.1359	3,005.63ms
Flan-T5	0.0294	0.0058	0.0175	0.0564	1,790.91ms
RIG (ours)	**0.4259**	**0.2476**	**0.5170**	**0.7906**	**181.82ms**

Table 8: The ranking performance and inference time on the task of conceptualization prediction.

ranking in a generative manner. Specifically, we experimented with `Mistral-7B-Instruct-v0.3`, `Meta-Llama-3-8B-Instruct`, and `flan-t5-xl`. For each intention, a candidate pool containing both true and false concepts was constructed, and the concepts were ranked based on their generation order by the LLMs. Similar to the intention prediction task, we applied negative sampling for each intention to rank positive concepts among a pool of 500 candidates. Our proposed method relied on an embedding-based approach, leveraging a fine-tuned embedding model to generate embeddings for intentions and concepts. This fine-tuning ensured that intention embeddings were closer to their corresponding positive concept embeddings.

As shown in Table 8, the results demonstrate that our conceptualization prediction method, based on the intention knowledge graph, achieves superior performance compared to LLMs regarding both ranking accuracy and inference time.

Product Recovery We constructed a session-intention pair dataset in previous sessions, including product IDs, descriptions, and user intentions. Here, we construct a new benchmarking dataset of the product-intention pairs from all session-intention pairs by assuming the items within the session share the same intentions as the sessions. These product-intention pairs were randomly split into training, validation, and test sets with an 8:1:1 ratio. This dataset served as the basis for evaluating the ability of different methods to recover relevant intentions for given products. To ensure a fair comparison, we focused on 1,203 overlapping products between the test set of the product-intention pairs in RIG and Folkscope. For these overlapped products, we compared the intentions generated by Folkscope and RIG. Rankings and evaluations were conducted on this same set of overlapping products, enabling a direct and balanced comparison of intention quality between the two systems. This methodology ensured the evaluation results reflected each method's capability to generate relevant and accurate intentions. Table 9 provides the detailed evaluation results. We used pre-computed

	# Intentions	(Ovlp.)	# Products	(Ovlp.)
FolkScope	1,846,715	67,789	211,372	1,203
RIG	295,620	4,829	453,124	1,203
	MRR	Hit@1	Hit@3	Hit@10
FolkScope(overlap)	0.2808	0.0977	0.2816	0.9096
RIG (overlap)	**0.3161**	**0.1257**	**0.3453**	**0.9263**
FolkScope (full)	0.2779	0.0947	0.2782	0.9071
RIG (full)	**0.3025**	**0.1188**	**0.3147**	**0.9072**

Table 9: The performance on product recovery. Ovlp. stands for overlap.

intention and product embeddings as input to a Multi-Layer Perceptron (MLP) scoring model to train our model. The MLP was trained using Noise Contrastive Estimation (NCE) loss to distinguish between relevant and irrelevant intentions. We assessed the model's performance on the test set during the evaluation phase by analyzing one positive sample against ten negative samples. Cosine similarity scores were computed for ranking intentions, enabling precise comparisons across methods. Finally, we compared our method with Folkscope under identical experimental settings. The results, presented in Table 9, demonstrate that our knowledge graph significantly outperforms Folkscope in recovering relevant products from intentions, highlighting the superior efficacy of our approach.

5 Extrinsic Evaluation

To demonstrate the practical utility of our relational intention knowledge graph, we evaluate its effectiveness in enhancing session-based recommendation, a key application domain for e-commerce platforms. Using the Amazon M2 dataset, we show how the rich semantic relationships captured in our knowledge graph can improve recommendation quality compared to state-of-the-art methods.

Data Preparation We use the complete English subset of the Amazon M2 dataset integrated with our constructed knowledge graph. Following standard practice, we partition all sessions into training, validation, and test sets at an 8:1:1 ratio. To preserve the integrity of the dataset, we use the original preprocessing without filtering, thus avoiding potential session loss or artificial item connections.

The integration of a large-scale commonsense knowledge graph (containing over 351 million edges) with conventional recommendation systems presents significant challenges. To address this, we develop a meta-path approach to extract actionable item relationships from our knowledge graph. First, we identify session pairs connected

Datasets	Metric	FPMC	GRU4Rec	BERT4Rec	SASRec	SASRecF	CORE	SR-GNN	GCE-GNN	DIF-SR	FEARec	DGNN	ISRec	Satori	KA-MemNN	RIGRec
M2 (UK)	Recall@5	0.2523	0.2792	0.1899	0.3075	0.2957	0.2990	0.2928	<u>0.3130</u>	0.3128	0.3088	0.3021	0.3073	0.2973	0.2932	**0.3342***
	Recall@10	0.3121	0.3469	0.2641	0.3964	0.3713	0.3949	0.3678	<u>0.4001</u>	0.3990	0.3941	0.3882	0.3981	0.3821	0.3781	**0.4229***
	Recall@20	0.3696	0.4108	0.3349	0.4723	0.4406	<u>0.4768</u>	0.4381	0.4726	0.4739	0.4691	0.4563	0.4754	0.4436	0.4419	**0.5003***
	Recall@50	0.4389	0.4865	0.4197	0.5621	0.5245	<u>0.5697</u>	0.5171	0.5542	0.5598	0.5552	0.5584	0.5676	0.5348	0.5295	**0.5863***
	Recall@100	0.4841	0.5346	0.4744	0.6159	0.5771	<u>0.6223</u>	0.5662	0.6032	0.6171	0.6100	0.6072	0.6201	0.5931	0.5847	**0.6398***
	NDCG@5	0.1933	0.2118	0.1260	0.2121	0.2208	0.1673	0.2195	<u>0.2214</u>	0.2171	0.2138	0.2201	0.2207	0.2189	0.2163	**0.2214**
	NDCG@10	0.2126	0.2327	0.1501	0.2406	0.2432	0.1985	0.2418	0.2441	<u>0.2451</u>	0.2415	0.2431	0.2438	0.2425	0.2386	**0.2503***
	NDCG@20	0.2272	0.2499	0.1682	0.2598	0.2634	0.2193	0.2616	0.2626	<u>0.2641</u>	0.2605	0.2637	0.2633	0.2623	0.2591	**0.2703***
	NDCG@50	0.2411	0.2648	0.1848	0.2679	0.2801	0.2379	0.2784	0.2807	<u>0.2812</u>	0.2777	0.2797	0.2795	0.2793	0.2737	**0.2877***
	NDCG@100	0.2484	0.2718	0.1937	0.2862	0.2887	0.2472	0.2865	0.2871	<u>0.2891</u>	0.2866	0.2867	0.2881	0.2864	0.2843	**0.2957***

Table 10: Performance comparison with baselines. The best and second-best results are shown in bold and underlined fonts. "$*$" represents the significant improvement over the best baseline with p-value < 0.05.

through intention-level relationships (either conceptual or temporal). To ensure relationship quality, we retain only session pairs that either (1) share at least six distinct meta-paths through commonsense relation nodes or (2) demonstrate complete reachability from one session's concept nodes to the other's. From these high-quality session pairs, we construct a weighted item graph $G = (V, E)$, where nodes represent products and edges represent their co-occurrence relationships within semantically connected sessions. Edge weights correspond to co-occurrence frequency, providing a measure of relationship strength. This approach effectively distills our comprehensive knowledge graph into a focused item relationship network tailored for recommendation tasks.

RIGRec: Intention-Enhanced Recommendation Model We develop RIGRec, a novel recommendation model that leverages the rich semantic relationships captured in our intention knowledge graph. The model employs graph convolution to derive informative item representations:$\boldsymbol{E}^{l+1} = \boldsymbol{A}\boldsymbol{E}^{l}$, where $\boldsymbol{A}$ represents the adjacency matrix of graph G, and $\boldsymbol{E} \in \mathbb{R}^{N\times d}$ stores the d-dimensional embeddings for all N items. We implement lightweight convolution operations and sum pooling to enhance computational efficiency. After L convolution layers, the resulting representations $\boldsymbol{E}^{L}$ capture both item characteristics and their intention-based relationships. For session modeling, we adopt SASRec's self-attention architecture, which effectively aggregates item representations within a session to model user preferences. This creates an end-to-end recommendation system that seamlessly integrates intention-level reasoning with sequential pattern recognition, enabling more contextually appropriate recommendations.

Baselines and Evaluation Metrics We compare our model with following ten representative and state-of-the-art methods, covering (1) the classical method FPMC (Rendle et al., 2010), (2) the RNN-based method GRU4Rec (Hidasi et al., 2016), (3) the predominant attention-based methods including BERT4Rec (Sun et al., 2019), SASRec (Kang and McAuley, 2018), CORE (Hou et al., 2022) and FEARec (Du et al., 2023), (4) graph-based methods including SR-GNN (Wu et al., 2019b) and GCE-GNN (Wang et al., 2020), (5) side information fusion methods including SASRecF (Kang and McAuley, 2018) and DIF-SR (Xie et al., 2022). We exclude some state-of-the-art methods like FAPAT (Liu et al., 2023b) due to the need for massive support resources or exponential computation complexity. Additionally, we compare our approach with state-of-the-art intention-aware recommendation models: DGNN (Li et al., 2023), ISRec (Li et al., 2021b), Satori (Chen et al., 2022), and KA-MemNN (Zhu et al., 2021).

We employ two standard metrics in the field of recommender systems, including Recall at a cutoff top k (Recall@k) and Normalized Discounted Cumulative Gain at a cutoff top k (NDCG@k). We rank the ground-truth item alongside all candidates to ensure an unbiased evaluation rather than adopting the negative sampling strategy. We report the averaged metrics over 5 runs with the commonly utilized $k \in \{5, 10, 20, 50, 100\}$. To ensure unbiased evaluation, we rank the ground-truth item alongside all candidate items rather than using negative sampling. Results represent averages across five independent runs, with statistical significance determined through unpaired t-tests (p < 0.05). The implementation details are in the Appendix A.5.

Performance Comparison As shown in Table 10, RIGRec consistently outperforms all baseline methods across nearly all evaluation metrics, with statistically significant improvements over the strongest competitors. These results demonstrate the substantial benefits of incorporating intention-

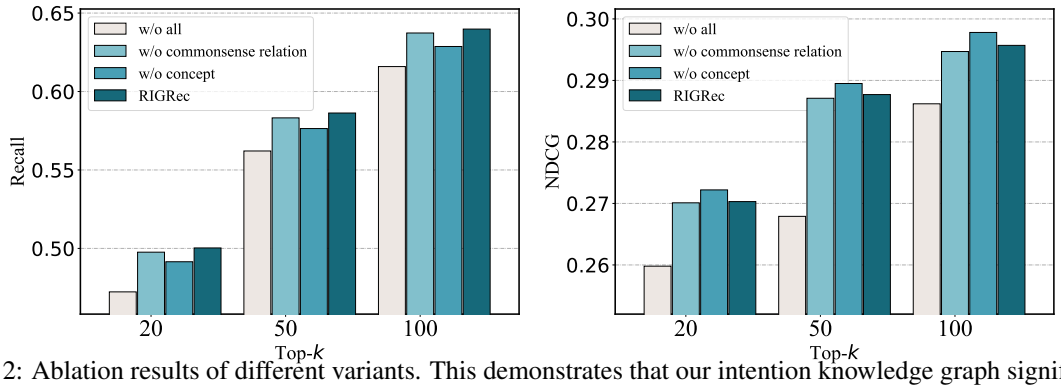

Figure 2: Ablation results of different variants. This demonstrates that our intention knowledge graph significantly enhances recommendation performance compared to SASRec. Both intention conceptualization and concept relations effectively improve results, with each type of relation contributing uniquely to different metrics. This highlights the importance of incorporating diverse nodes and relations in the knowledge graph.

level relationships.

The performance gap between RIGRec and SASRec—which serves as RIGRec's backbone architecture—is particularly notable. This difference directly quantifies the value added by our intention knowledge graph, confirming that modeling user intentions and their relationships provides critical information beyond what can be captured by sequential patterns alone.

Graph-based methods like SR-GNN and GCE-GNN, while theoretically capable of modeling complex item relationships, show inferior performance compared to our approach. This stems from their focus on direct item transitions without consideration of the underlying user intentions driving these transitions. By explicitly modeling intention-level relationships, RIGRec captures deeper semantic connections between items, resulting in more contextually appropriate recommendations.

The mixed performance of side information fusion methods (DIF-SR and SASRecF) highlights the challenges of effectively integrating auxiliary information into recommendation systems. While DIF-SR achieves competitive results through careful information fusion, SASRecF shows performance degradation compared to the base SASRec model in terms of Recall@k. This aligns with findings that inappropriate information integration strategies can degrade model effectiveness. Our approach, by contrast, leverages large language models to extract high-quality intention information, filtering out noise and irrelevant features that might otherwise compromise recommendation quality.

Ablation Study To isolate the contributions of different components in our knowledge graph, we conduct ablation studies comparing RIGRec against three variants: (1) **w/o all**: Removes the entire item graph derived from our knowledge graph (equivalent to standard SASRec). (2) **w/o concept**: Removes edges derived from concept-mediated paths. (3) **w/o commonsense relation**: Removes edges derived from commonsense relation paths.

The ablation study reveals three insights: (1) The intention knowledge graph substantially improves recommendation quality beyond what sequential patterns alone can achieve; (2) Both conceptualization and commonsense relations provide complementary signals that enhance performance; and (3) These relation types contribute differently to evaluation metrics—conceptualization improves recall by identifying broadly relevant items, while commonsense relations enhance NDCG by capturing fine-grained semantic connections for better ranking. These findings confirm that modeling diverse intention relationships leads to more contextually appropriate recommendations by providing a more comprehensive understanding of user behavior.

6 Conclusion

We introduce IGC-RC, a framework for automatically constructing knowledge graphs that model relationships between user intentions in e-commerce contexts. Our Relational Intention Graph (RIG) captures temporal, causal, and conceptual connections between user goals, demonstrating exceptional quality in both intrinsic evaluations and recommendation tasks. By bridging the gap between observable behaviors and underlying intentions, RIG enables more accurate prediction of user needs and significantly enhances recommendation performance. This advancement represents an important step toward more human-like understanding in intelligent e-commerce systems.

Limitations

The intention generation process relies on GPT-3.5, which may introduce additional computational overhead. Future work could explore more efficient language models to streamline this component. Our framework is evaluated using the Amazon M2 dataset, which is specific to e-commerce. The applicability of the proposed method to other domains remains to be tested. The current implementation focuses on the English subset of the dataset. Extending the framework to support multiple languages could enhance its versatility. While we incorporate commonsense relations such as temporality and causality, the scope of relation types is limited. Incorporating a broader range of relational categories may improve the knowledge graph's comprehensiveness. We would like to provide more validation if more suitable datasets are available. However, most public datasets are desensitized and anonymized, making generating intention based on the anonymized ID features hard. Besides, our utilized M2 dataset is a mixed-type dataset, which already contains multi-typed items (sports, beauty, baby, etc.). The item and session sizes also exceed general research works.

Ethics Statement

This study ensures the responsible use of data and technology by utilizing the anonymized Amazon M2 dataset, which safeguards user privacy and complies with data protection regulations. We have implemented measures to prevent the inclusion of any personally identifiable information (PII). Additionally, we acknowledge potential biases in the dataset and have taken steps to mitigate them through standard preprocessing techniques. Our use of large language models focuses on enhancing user experience without manipulating behavior, and we advocate for transparency in deploying intention knowledge graphs within e-commerce platforms.

References

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*, pages 1247–1250.

Jinpeng Chen, Yuan Cao, Fan Zhang, Pengfei Sun, and Kaimin Wei. 2022. Sequential intention-aware recommender based on user interaction graph. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 118–126.

Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guanfeng Liu, Yanchi Liu, and Victor S Sheng. 2023. Frequency enhanced hybrid attention network for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023*, pages 78–88.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2021*, pages 2648–2659.

Thomas R Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

Jiayan Guo, Peiyan Zhang, Chaozhuo Li, Xing Xie, Yan Zhang, and Sunghun Kim. 2022. Evolutionary preference learning via graph nested GRU ODE for session-based recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM 2022*, pages 624–634.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *The 4th International Conference on Learning Representations, ICLR 2016*.

Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. 2019. Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, 6(21):1900808.

Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. 2022. Core: simple and effective session-based recommendation within consistent representation space. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022*, pages 1796–1801.

Zhongyu Huang, Yingheng Wang, Chaozhuo Li, and Huiguang He. 2022. Going deeper into permutation-sensitive graph neural networks. In *International Conference on Machine Learning, ICML 2022*, volume 162, pages 9377–9409.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 6384–6392.

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. 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*.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining, ICDM 2018*, pages 197–206.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2022. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. In *Proceedings of the 23rd ACM Conference on Economics and Computation, EC 2022*, page 29.

Chaozhuo Li, Bochen Pang, Yuming Liu, Hao Sun, Zheng Liu, Xing Xie, Tianqi Yang, Yanling Cui, Liangjie Zhang, and Qi Zhang. 2021a. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021*, pages 223–232.

Feng-Lin Li, Hehong Chen, Guohai Xu, Tian Qiu, Feng Ji, Ji Zhang, and Haiqing Chen. 2020. Alimekg: Domain knowledge graph construction and application in e-commerce. In *In proceedings of The 29th ACM International Conference on Information and Knowledge Management, CIKM 2020*, pages 2581–2588.

Haoyang Li, Xin Wang, Ziwei Zhang, Jianxin Ma, Peng Cui, and Wenwu Zhu. 2021b. Intention-aware sequential recommendation with structured intent transition. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5403–5414.

Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 26th ACM on Conference on Information and Knowledge Management, CIKM 2017*, pages 1419–1428.

Zihao Li, Xianzhi Wang, Chao Yang, Lina Yao, Julian McAuley, and Guandong Xu. 2023. Exploiting explicit and implicit item relationships for session-based recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pages 553–561.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023a. 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*, pages 1264–1287.

Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 1831–1839.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinghang Wu, Qiang Li, Keping Yang, and Kenny Q. Zhu. 2020. Alicoco: Alibaba e-commerce cognitive concept net. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD 2020*, pages 313–327.

Zhiqiang Pan, Fei Cai, Wanyu Chen, Honghui Chen, and Maarten de Rijke. 2020. Star graph neural networks for session-based recommendation. In *In proceedings of the 29th ACM International Conference on Information and Knowledge Management, CIKM 2020*, pages 1195–1204.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 811–820.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 3027–3035.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 4444–4451.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pages 1441–1450.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2017. Webchild 2.0 : Fine-grained commonsense knowledge distillation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 115–120.

Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018*, pages 565–573.

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*, pages 13520–13545.

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. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *CoRR*, abs/2401.07286.

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, ACL 2023*, pages 13111–13140.

Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. 2020. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020*, pages 169–178.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019a. Session-based recommendation with graph neural networks. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 346–353.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019b. Session-based recommendation with graph neural networks. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, volume 33, pages 346–353.

Xin Xia, Hongzhi Yin, Junliang Yu, Qinyong Wang, Lizhen Cui, and Xiangliang Zhang. 2021. Self-supervised hypergraph convolutional networks for session-based recommendation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 4503–4511.

Yueqi Xie, Peilin Zhou, and Sunghun Kim. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022*, pages 1611–1621.

Changlong Yu, Xin Liu, Jefferson Maia, Yang Li, Tianyu Cao, Yifan Gao, Yangqiu Song, Rahul Goutam, Haiyang Zhang, Bing Yin, and Zheng Li. 2024. Cosmo: A large-scale e-commerce common sense knowledge generation and serving system at amazon. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, page 148–160.

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*, pages 1173–1191.

Nasser Zalmout, Chenwei Zhang, Xian Li, Yan Liang, and Xin Luna Dong. 2021. All you need to know to build a product knowledge graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021*, pages 4090–4091.

Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. 2021. Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021*, pages 3895–3905.

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.

Nengjun Zhu, Jian Cao, Xinjiang Lu, and Hui Xiong. 2021. Learning a hierarchical intent model for next-item recommendation. *ACM Transactions on Information Systems (TOIS)*, 40(2):1–28.

A Further Details on Experiments

A.1 Human Annotation

We collected annotations for 3,000 session-intention pairs, each evaluated by three annotators. The inter-annotator agreement scores were 0.91 for plausibility and 0.74 for typicality. Three annotators independently evaluated each of the 1,000 intention-intention discourse pairs for intention relation classification, achieving an overall inter-annotator agreement of 0.69. All raw annotation data will be made publicly available to support future research in this area. We use exactly the same annotation guidelines and criteria from the FolkScope paper, which can ensure the same standard of annotations on plausibility and typicality.

A.2 Concept Prediction

We conducted experiments to compare different methods for predicting conceptualized intentions, where the goal is to predict corresponding concepts for a given purpose.

The first method employed a generative approach using LLMs, specifically Mistral-7B-Instruct-v0.3, Meta-Llama-3-8B-Instruct, and flan-t5-xl. This approach generated concepts in the same way as our RIG construction process. During testing, we created a candidate pool containing both true and false concepts for each intention. The LLMs generated ten concepts per intention, and matching concepts were ranked first while maintaining their generation order. We used negative sampling to create a candidate pool of 500 concepts. Importantly, we did not fine-tune the LLMs to maintain consistency with the approach used in RIG's construction.

Our proposed approach's second method utilized an embedding-based model (bge-base-en-v1.5) to transform intentions and concepts into embeddings and compute their cosine similarities. We fine-tuned the embedding model using contrastive learning, incorporating cross-entropy loss to improve matching performance. During testing, we created a candidate pool of 500 concepts and ranked them based on their cosine similarities with the given intention.

We constructed a dataset of intention-concept pairs from RIG for our experimental setup. We split it into training, validation, and testing sets with an 8:1:1 ratio, resulting in 147,801 intention-concept pairs in the test set. We conducted our experiments using an Nvidia RTX A6000 GPU. The models were evaluated using standard metrics, including MRR, Hit@1, Hit@3, Hit@10, and inference speed. The results demonstrated that our embedding-based method achieved superior prediction accuracy and computational efficiency performance compared to the LLM approach.

A.3 Intention Prediction

The input of this task is a session, and the output of this task is the ranking of the ground-truth intention over a pool of negative intentions. The goal of this task is to rank the correct intention higher than the incorrect intentions. Because the input of this task is a session, so we use the SASRec as the backbone model. We used the all-MiniLM-L6-v2 model to generate embeddings for session items and intentions. These item embeddings initialized the item embedding matrix in SASRec. During training, the parameters of session encoders and item embeddings are tuned. TripletMarginLoss minimized the distance between session and positive intention embeddings while maximizing the distance to negative intentions. Random sampling of positive and negative samples, followed by backpropagation and early stopping, was used to optimize the model.

A.4 Product Recovery Benchmark

We use the same evaluation method to ensure the fairness of the evaluation. Table 9 shows the evaluation based on 1,203 overlapping products between the two graphs. For these identical products, we compared the intentions generated by Folkscope and RIG, respectively. The rankings and evaluations were conducted using this same set of overlapping products, allowing for a direct comparison of intention quality between the two systems. This methodology ensures a fair and balanced assessment of each method's ability to generate relevant intentions.

Using the same sentence embedding model (BGE), we acquired pre-computed intention and product embeddings. Then, we used the intention embeddings as input and the product embeddings as output to train a Multi-Layer Perceptron (MLP) scoring model using Noise Contrastive Estimation (NCE) loss. In the evaluation phase, for each intention, we analyzed one positive sample against 10 negative samples by calculating the cosine similarity scores of their embeddings to the target embedding and subsequent rankings, where the rank was determined by the count of negative samples

scoring higher than the positive sample plus one.

We have identified and addressed key limitations in prior work on intention generation using large language models. (Zhou et al., 2024) highlighted two significant issues with FolkScope's intentions: property-ambiguity and category-rigidity. These issues primarily stem from two factors. First, using a relatively weak language model (OPT-30B) for intention generation limited its ability to produce high-quality outputs. Second, the reliance on ConceptNet relations for prompt construction introduced constraints, as some relations (e.g., "made of") were not well-suited for generating diverse and meaningful user intentions.

Our work tackles these limitations through two key improvements. First, we employ a more capable language model, Llama3-8B-instruction, to enhance the quality of intention generation. Second, we remove the reliance on ConceptNet relations in prompts. Instead, we leverage the advanced capabilities of modern language models to capture open-ended intentions, enabling the generation of more natural and diverse user intentions without being restricted by predefined relation types.

We have achieved improved intention quality by addressing these issues from the outset. To validate these improvements, we followed the evaluation methodology outlined in (Zhou et al., 2024), using product recovery benchmarks. As shown in Table 9, our approach demonstrates superior performance compared to the baseline model.

A.5 Implementation Details of Session Recommendation

For a fair comparison, the dimension of item embedding is set to 64 for all methods. Grid search strategy is applied to determine the optimal configuration of standard parameters, involving the learning rate in $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$, the dropout rate in $\{0, 0.1, 0.2, 0.3, 0.4\}$, the loss function in {BPR loss, Binary Cross Entropy loss, Cross Entropy loss} and the coefficient of $L2$ regularization in $\{0, 1e^{-2}, 1e^{-3}, 1e^{-4}\}$.

B Large Language Model Generation Prompts

The following Figure 3 and Figure 4 denote the prompts that we used to generate the intentions and concepts.

N-gram	FolkScope	RIG	w/ConceptNet Rel
2-gram	0.0307	**0.5274**	0.3759
3-gram	0.0480	**0.7931**	0.5806
4-gram	0.0648	**0.9100**	0.6997
5-gram	0.0837	**0.9623**	0.7751
6-gram	0.1046	**0.9833**	0.8264

Table 11: N-gram diversity scores of intentions extracted from FolkScope and RIG. The results validate that RIG generates more diverse intentions by removing ConceptNet relation constraints.

Session Intention Generation

Below is a user's chronological record list: [SESSION]
Explain the basic intentions of this user exactly. Output several different intentions one by one to answer the following question: Users buy these items because they want to:
intention 1: {a simple verb phrase within 10 words}
intention 2: {a simple verb phrase within 10 words}
...

Figure 3: This figure shows the prompts we use to make LLM understand and generate intentions from user sessions.

C Annotation Questions

Here, we give three examples of annotation questions from our questionnaire for the Amazon Mechanical Turk. They are for the intentions quality annotation, intention relation classification annotation, and intention conceptualization classification.

D Evaluating Diversity of Intention

As shown in Figure 11, we further compare the diversity of the generated intentions by using the n-gram diversity. It is defined as the ratio of the unique n-gram counts to all n-gram counts:

$$\text{Diversity}(D, n) = \frac{\#\text{ unique } n\text{-grams in } D^{\oplus}}{\#\text{ n-grams in } D^{\oplus}} \quad (1)$$

Where $D^{\oplus}$ denotes the dataset D concatenated into a single string. We use six as the maximum n-gram length. This method captures repeated sequences in addition to single-token diversity.

We measured the diversity of the corpus formed by the intentions extracted from ForkScope and RIG. For the intention generation with conceptnet relation, we randomly sample 100 intentions

Abstract Intention Generator

I will give you an INTENTION. You need to give several phrases containing 1-3 words for the ABSTRACT INTENTION of this INTENTION. You must return your answer in the following format: phrases1,phrases2,phrases3,...., which means you can't return anything other than answers. These abstract intention words should fulfill the following requirements:
1. The ABSTRACT INTENTION phrases can well represent the INTENTION.
2. The ABSTRACT INTENTION phrases don't have a lot of less relevant word meanings. For example, "spring" is not a good abstract intention word because it can represent both a coiled metal device and the season of the year.
3. The ABSTRACT INTENTION phrases of the same INTENTION cannot be semantically similar to each other. For example, health and wellness are two close synonyms, so they can't be together.
INTENTION: Moisturize dry skin while enjoying a special effect bath.
Your answer: hydration, skincare
INTENTION: Create a festive atmosphere for a Christmas party.
Your answer: party planning, celebration, decorations, holiday spirit
INTENTION: [INTENTION].
Your answer:

Figure 4: This figure shows our prompts to make LLM conceptualize the user intentions.

derived from each of 17 conceptnet relations, totaling 1700 intentions. For RIG and Folkscope, we sample 1700 intentions from each of them to compute the n-gram diversity. The results demonstrate that RIG's N-gram diversity of intentions is significantly higher than ForkScope's. These findings validate our claim that removing ConceptNet relation constraints and using a better generation model generates more diverse intentions.

E Analysis of Different Backbones for RIGRec

The choice of backbone architecture is an important consideration when implementing our approach. In our main experiments, we primarily utilized SASRec as the backbone for RIGRec due to its simplicity, robustness, and widespread adoption as a baseline in session-based recommendation research. This choice provides clear interpretability of the improvements brought by our intention knowledge graph.

However, to demonstrate the generalizability of our approach, we conducted additional experiments using GRU4Rec as an alternative backbone. Table 13 presents the performance comparison between the original backbones (GRU4Rec and SASRec) and their RIG-enhanced versions (RIG-GRU4Rec and RIGRec).

The results demonstrate that our approach consistently improves performance regardless of the backbone architecture. When applied to GRU4Rec, our intention knowledge graph enhances performance across all metrics, with particularly notable improvements in Recall@20 (13.6% improvement) and Recall@50 (14.3% improvement). This confirms that the benefits of our relational intention knowledge graph are not limited to a specific recommendation architecture.

Nevertheless, the RIGRec model (based on SASRec) still achieves the best overall performance. This can be attributed to SASRec's inherent advantages in capturing long-range dependencies through its self-attention mechanism, which may better complement the high-level intention information provided by our knowledge graph. The results suggest that while our approach generalizes well to different backbones, the choice of backbone can still influence the absolute performance levels achieved.

F Case Studies

To illustrate the practical effectiveness of our RIG approach, we present three representative case studies in Table 14. These examples highlight how our model leverages intention relationships to generate more contextually relevant recommendations compared to baseline methods.

F.1 Gaming Setup Completion

In the first case, the user has purchased basic gaming peripherals (mouse, keyboard, and headset). The baseline model recommends similar gaming peripherals and accessories without considering the user's broader intention. In contrast, our RIGRec model identifies the user's goal of setting up a complete gaming system and recommends complementary products that enhance the overall gaming environment, such as a desk with cable management and RGB lighting solutions. This demonstrates how our model captures the conceptual relationships between products through the user's intentions.

F.2 Baby Care Essentials

The second case shows a user purchasing basic newborn care items. While the baseline recommends additional hygiene products that are similar

to those already in the cart, our model recognizes the broader intention of preparing for a newborn's feeding needs. By identifying the connection between diaper changing and feeding essentials as part of comprehensive baby care, RIGRec suggests complementary products like anti-colic bottles and a bottle warmer that the baseline misses entirely.

F.3 Halloween Preparation

In the third case, the user is collecting Halloween decorations. The baseline model focuses on recommending additional decorative items and costumes. Our model, however, identifies the potential intention of hosting a Halloween party and recommends party-specific items like themed serving trays and snack bowls. This demonstrates how RIGRec's understanding of intention relationships (decorating for Halloween $\rightarrow$ hosting a Halloween party) enables it to anticipate future user needs that aren't explicitly indicated in the current session.

F.4 Analysis of Intention Relations

Table 15 shows examples of how our model captures key intention relationships that help generate better recommendations. These intention pairs demonstrate the temporal and causal connections our knowledge graph identifies.

These cases illustrate how our intention knowledge graph helps bridge the gap between observed behaviors and underlying user goals, resulting in recommendations that better address the user's complete needs rather than simply suggesting similar products. The ability to identify conceptual relationships between different intentions enables our model to make more contextually appropriate and diverse recommendations.

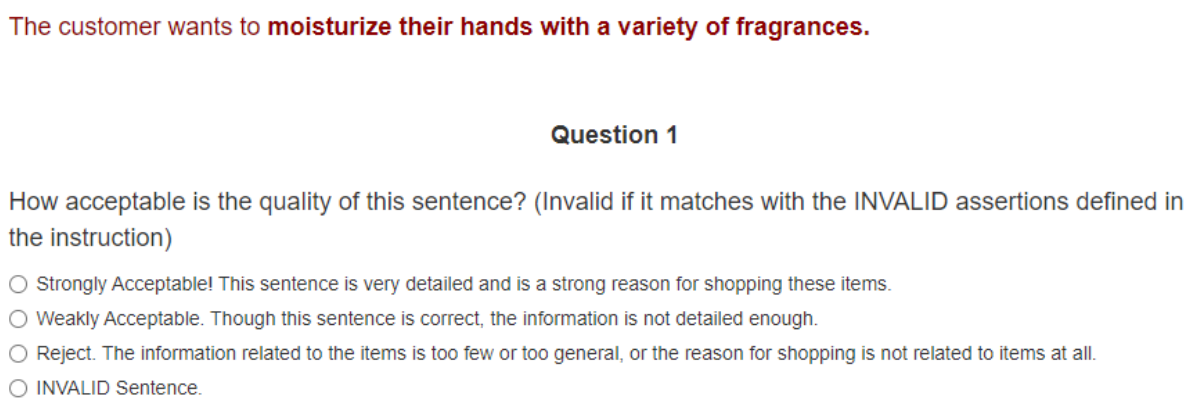

Figure 5: This figure shows an example annotation question for the quality of session intention generation.

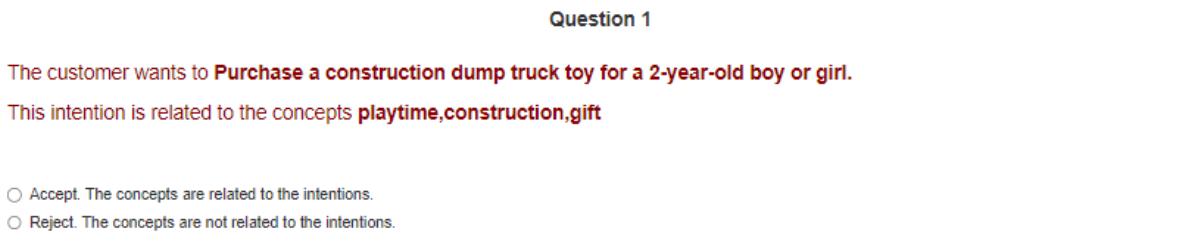

Figure 6: This figure shows an example annotation question for the quality of session intention conceptualization.

Intention 1	Intention 2	Assertion
Relieve discomfort and soothe itching caused by haemorrhoids.	Purchase unscented baby wipes for sensitive skin.	People relieve discomfort and soothe itching caused by haemorrhoids, and simultaneously, they purchase unscented baby wipes for sensitive skin.
Make coffee at home.	Enjoy a variety of coffee flavors at home.	People make coffee at home usually after they enjoy a variety of coffee flavors at home.
Dress up as Lara Croft for a costume party or event.	Have fun with Halloween-themed party games.	People dress up as Lara Croft for a costume party or event, and simultaneously, they have fun with Halloween-themed party games.
Find a cream that provides fast and numbing relief from haemorrhoid symptoms.	Use advanced moisture absorption technology.	People find a cream that provides fast and numbing relief from haemorrhoid symptoms because they use advanced moisture absorption technology.
Maintain personal hygiene and cleanliness.	Purchase a razor handle and blade refills for men's shaving.	People maintain personal hygiene and cleanliness, as a result, they purchase a razor handle and blade refills for men's shaving.

Table 12: This table presents two candidate intentions and related assertions. The assertions provide an interpretive summary of the relationship between the paired intentions. The templates mapping from triples to assertions are marked in green.

Model	R@5	R@10	R@20	R@50	R@100	N@5	N@10	N@20	N@50	N@100
GRU4Rec	0.2792	0.3469	0.4108	0.4865	0.5346	0.2118	0.2327	0.2499	0.2648	0.2718
RIG-GRU4Rec	0.2923	0.3852	0.4665	0.5562	0.6028	0.2126	0.2367	0.2474	0.2693	0.2746
SASRec	0.3075	0.3964	0.4723	0.5621	0.6159	0.2121	0.2406	0.2598	0.2679	0.2862
RIGRec	**0.3342**	**0.4229**	**0.5003**	**0.5863**	**0.6398**	**0.2214**	**0.2503**	**0.2703**	**0.2877**	**0.2957**

Table 13: Performance comparison of different backbone architectures with and without our knowledge graph.

Session History	Generated Intentions	Baseline Recommendations (SASRec)	Our Model Recommendations (RIGRec)
Wireless Gaming Mouse; RGB Gaming Keyboard; Gaming Headset with Microphone	Set up a complete gaming peripherals system; Enhance gaming experience with RGB accessories; Improve communication during multiplayer games	Ergonomic Gaming Chair; Gaming Mouse Pad; USB Hub; Mechanical Gaming Keyboard; RGB Gaming Mouse	**Gaming Mouse Pad with RGB**; **Gaming Desk with Cable Management**; **Dual Monitor Stand**; **RGB LED Strip for Desk**; **Gaming Controller**
Baby Diapers Size 1 (240 Count); Baby Wipes Unscented; Diaper Rash Cream; Baby Bottle Sterilizer	Prepare essential baby care items for a newborn; Maintain baby hygiene; Prevent and treat diaper rash; Ensure safe feeding equipment	Baby Powder; Baby Wipes Sensitive; Baby Lotion; Baby Shampoo; Diaper Bag	**Baby Bottles Anti-Colic**; **Bottle Drying Rack**; **Changing Pad**; **Baby Bottle Brush Set**; **Bottle Warmer**
Pumpkin Carving Kit; Halloween String Lights; Artificial Spider Web; Halloween Doorbell with Spooky Sounds	Decorate home for Halloween; Create a spooky atmosphere; Prepare for trick-or-treaters; Host a Halloween party	Halloween Window Decorations; Halloween Candy Bowl; Halloween Mask; Halloween Costume; Fog Machine	**Halloween Door Wreath**; **Motion-Activated Halloween Props**; **Outdoor Halloween Projector Lights**; **Halloween Themed Serving Tray**; **Halloween Party Snack Bowls**

Table 14: Case Study: Example Sessions and Recommendations

Initial Intention	Related Intention	Relation Type
Set up a gaming peripherals system	Create an immersive gaming environment	Causality
Decorate home for Halloween	Host a Halloween party	Temporal (Before→After)
Ensure safe feeding equipment	Prepare bottles for infant feeding	Synchronous
Upgrade PC components	Enhance gaming performance	Causality
Purchase cooking utensils	Prepare homemade meals	Temporal (Before→After)

Table 15: Examples of Captured Intention Relations in RIG

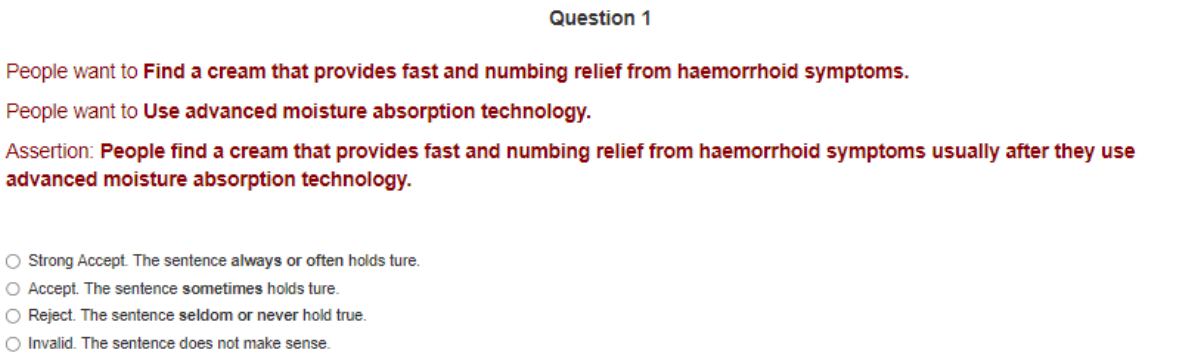

Figure 7: This figure shows an example annotation question for the quality of session intention relation classification.