# 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 behind behaviors poses a significant challenge for online platforms. Recent research has introduced the concept of intention knowledge graphs (Yu et al., 2023, 2024) to address this issue. These graphs connect various user behaviors, such as co-purchasing (Yu et al., 2023) and queries (Yu et al., 2024), to their corresponding intentions expressed in natural language. intention knowledge graphs have proven valuable in applications like product recommendation (Yu et al., 2023) and search relevance (Yu et al., 2024).

Users often have specific intentions, such as preparing for a Halloween party. They may want to dress up as a vampire or werewolf and create a spooky atmosphere with decorations. However, these intentions are often only implicitly suggested by actions, such as the items shown in Figure 1 (A). Previous attempts at constructing intention knowledge graphs have focused on using large language models to generate explanations for behaviors (Yu et al., 2023), utilizing sources like session history and query keywords (Yu et al., 2024). For instance, analyzing the browsing history of items can reveal the intention to dress up for Halloween.

However, previous methods have not explored how intentions interrelate. Explicitly modeling these connections aligns with the shift towards System II [2] reasoning, emphasizing logical and sequential reasoning. Research has shown the importance of this reasoning in online behaviors (Kleinberg et al., 2022), where users focus on long-term rewards rather than unconscious browsing. Therefore, we aim to model explicit intention relations.

Our goal extends beyond understanding initial intentions to predict subsequent ones. For example, intending to buy a desk might lead to buying an office chair. This inference can improve user behavior modeling and recommendations. However, building these connections is unexplored.

Commonsense knowledge is crucial for modeling intention relationships. For example, planning a Halloween party requires understanding that costumes and decorations are likely intentions. We propose using commonsense relations to describe temporal and causal intention connections. After identifying intentions from user sessions, we use a classifier to determine inferential connections, enhancing our understanding of intention relationships. For instance, Figure 1 (B) demonstrates the plausible co-occurrence relationship between dressing up and decorating intentions.

Using inferential commonsense relations can be challenging to generalize to unseen situations. Concepts like conceptualization and instantiation help generalize commonsense reasoning (Wang et al., 2023b,a,c, 2024). We incorporate this into our knowledge graph, using models to conceptualize intentions into broader concepts, connecting related

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

[2]http://www.iro.umontreal.ca/~bengioy/AAAI-9feb2020.pdf

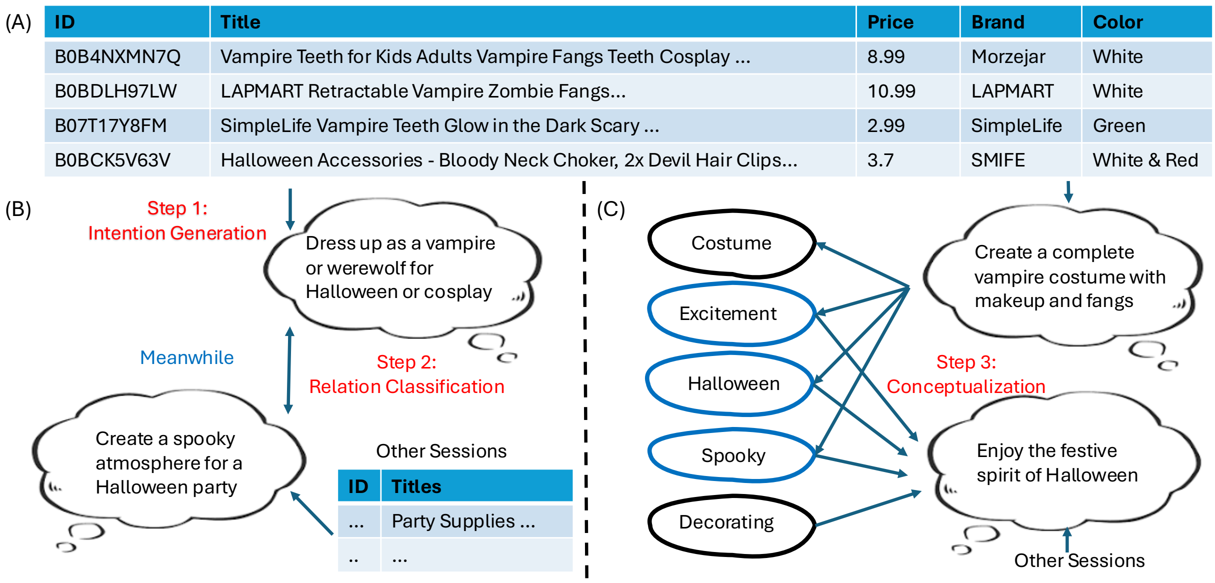



Figure 1: The structure of our knowledge graph. Part (A) shows an example of user behaviors within a session. Our knowledge graph emphasizes building relations between different intentions. In Part (B), we establish commonsense relations, highlighting the temporality and causality of intentions. Part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

intentions. For example, conceptualizing the intention to create a vampire costume can involve concepts like costumes and Halloween, linking it to similar intentions.

We introduce the Intention Generation, Conceptualization, and Relation Classification (IGC-RC) framework for constructing a commonsense knowledge graph from user behaviors to address the challenges. The framework involves three general steps: (1) **Intention Generation:** Using large language models, we generate intentions from user sessions. (2) **Conceptualization:** We conceptualize the generated intentions using conceptualization models. (3) **Relation Classification:** We generate commonsense statements from intention relations and verify their plausibility.

We apply the IGC-RC framework to the Amazon M2 session-based recommendation dataset, constructing a 351-million-edge knowledge graph. This graph captures intention-level relations and shows high acceptance in human evaluations, demonstrating the framework's effectiveness. Our constructed knowledge graph allows accurate prediction of intentions in new user sessions and improves session-based recommendation models, outperforming previous approaches. This underscores the practical utility of our framework.

In summary, our contributions are: (1) We model connections between users' deliberate and logical mental processes using an intention knowledge graph; (2) We develop the Intention Generation, Conceptualization, and Relation Classification (IGC-RC) framework, integrating user data and large language models; (3) We construct a large, high-quality Relational Intention Graph (RIG) from the Amazon M2 dataset, achieving superior performance in session recommendations.

## 2 Related Work

### 2.1 Knowledge Graphs for E-Commerce

Knowledge graphs (KGs) are crucial for enhancing recommendation systems on e-commerce platforms due to their structured and extensible nature. The Amazon Product Graph (Zalmout et al., 2021) exemplifies this by aligning Amazon's catalog with external KGs like Freebase (Bollacker et al., 2008), extracting numerous attributes across product types. This interconnected web of product data enables Amazon to offer personalized shopping experiences. Similarly, Alibaba's developments, such as AliCG (Zhang et al., 2021), AliCoCo (Luo et al., 2020), and AliMeKG (Li et al., 2020), leverage extensive e-commerce data for improved product recommendations and search capabilities. These KGs focus on item properties and categories without directly addressing the user motivations and intentions behind decisions. FolkScope (Yu et al., 2023) and COSMO (Yu et al., 2024) introduce intention knowledge graphs, incorporating user intentions from large language models and query-buy relations. However, they lack direct relations between intentions, such as causality and temporality. Our

| 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.

approach includes these commonsense relations to model intention dynamics more effectively.

Commonsense knowledge graphs (CSKGs) capture the world's commonsense knowledge and enable machines to reason like humans. Notable CSKGs 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) have structured commonsense knowledge but are limited in describing concrete user intentions and motivations, essential for understanding behavior on e-commerce platforms.

### 2.2 Session Understanding

Recent methods have explored using session history to reflect user intentions and enhance recommendation systems. Sequence modeling approaches, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have been employed to model session data (Hidasi et al., 2015; Li et al., 2017; Liu et al., 2018; Tang and Wang, 2018). Advancements in session-based recommendation systems have focused on Graph Neural Networks (GNNs) to capture better session transitions (Li et al., 2021; Guo et al., 2022; Huang et al., 2022). Wu et al. (2019a) first utilized GNNs to capture complex transitions using graph structures. Subsequent research incorporated position, target information, global context, and highway networks to enhance performance (Pan et al., 2020; Xia et al., 2021). However, existing methods often simplify user intentions by focusing on item features or statistical metrics, resulting in an inadequate understanding of user preferences. Our approach aims to address this by explicitly modeling user intentions within sessions.

## 3 IGC-RC Framework

### 3.1 Intention Generation

We utilize the Amazon M2 dataset (Jin et al., 2023), focusing on the English subset with 1.2 million sessions, to generate user intentions. The session history, comprised of sequences of user-browsed products, is processed to extract attributes such as titles, descriptions, models, sizes, and colors. These details are compiled into a JSON file, serving as the input for our intention generation module.

Using GPT-3.5, we generate concise, informative, and diverse user intentions, totaling 4.3 million from the dataset. Our approach surpasses FolkScope by using an advanced language model and eliminating ConceptNet relation constraints, fostering a more diverse output. See Figure 3 in the appendix for our prompt designs. We also include an evaluation on intention diversity in the Appendix Figure 11.

### 3.2 Intention Relation Classification

A significant challenge is linking user intentions meaningfully. These intentions, often event-based, rely on commonsense knowledge to establish connections, such as understanding that purchasing a gift usually precedes celebrating Christmas (Speer et al., 2017; Sap et al., 2019; Hwang et al., 2021; Fang et al., 2021; Tandon et al., 2017).

As shown in Table 12, we first convert candidate edges into assertions using a template-based approach to achieve this. Then, the plausibility estimation model Vera (Liu et al., 2023a) evaluates these assertions. Finally, we set a plausibility threshold to add high-plausibility edges to the knowledge graph. We conducted binary classifica-

| 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.

tion on relational edges. We conduct expert annotation to improve the performance of `VERA-T5-Base` model. Two experts independently annotated the data used to train the VERA model. Each intention pair received at least two independent annotations, with relationships classified according to the Penn Discourse Treebank 2.0 Annotation Manual (Prasad et al., 2008). The annotated relationships included temporal sequence, causality, and synchronization. The initial inter-annotator agreement was 85%, and disagreements were resolved through discussion between the experts. Overall, they annotated 1000 positive examples and randomly sampled another 1000 negative samples to construct a binary classification dataset, keeping 1600 for training and 200 for validation and testing. Table 4 shows the performance of the fine-tuned vera model. We applied this model to randomly drawn intention pairs, retaining those with a Vera score over 0.9.

### 3.3 Intention Conceptualization

A conceptualization provides an abstract, simplified view of a selected part of the world, encompassing objects, concepts, and relationships (Gruber, 1993; Himanen et al., 2019). Intention conceptualization extracts abstract concepts from user intentions to represent them in a knowledge graph, aiding e-commerce recommendation systems by modeling user intention dynamics.

Recent studies (Wang et al., 2023b,a,c, 2024) primarily address entity and event conceptualization, leaving user intention conceptualization as an open challenge. Two main approaches exist: crowdsourced and large language model (LLM) annotations (Wang et al., 2024). While crowd-sourcing offers high-quality results, it is costly and limited in coverage. Thus, we utilize a large language model for intention conceptualization.

| Knowledge Graph | Plausibility | Typicality |
|---|---|---|
| FolkScope | 0.6116 | 0.4491 |
| RIG (Ours) | **0.9552** | **0.6674** |

Table 5: The intention generation quality comparison with FolkScope.

To ensure diverse, unambiguous conceptualizations for machine integration, we design specific prompts to guide the LLM. Using the `Meta-Llama-3-8B-Instruct` model, we generate conceptualizations for each intention. The prompts used are detailed in Figure 4, and examples of the generated conceptualizations are shown in Table 3.

## 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 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 plausi-

| Knowledge Graph | Acceptance | Size |
|---|---|---|
| Atomic2020 | **86.8** | 0.6M |
| Atomic10x | 78.5 | 6.5M |
| Atomic NOVA | - | 2.1M |
| RIG (ours) | <u>81.2</u> | **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.

bility 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.

**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%.

### 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.

| | 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.

**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), 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 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

| | 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.

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 Benchmark** 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 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 intentions for products, highlighting the superior efficacy of our approach.

| | # 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.

## 5 Extrinsic Evaluation

Since our datasets are from the real-world Amazon M2 recommendation datasets, it is natural and necessary to validate the effectiveness of our work on the corresponding downstream tasks to show its usefulness. With this section, we demonstrate that our KG can provide extra information gain for session recommendation, which is also one of the major goals of the current KG.

### 5.1 Data Preparation

The whole English subset of the M2 dataset with our generated knowledge graph is utilized in this step for the session recommendation task. We split all sessions into train, validation, and test sets at a ratio of 8:1:1. Since the original dataset is preposessed before release, we do not employ additional filtering operations to avoid the risk of session loss or false connections among items.

Since all of the current session recommendation paradigms are difficult to integrate a million-level commonsense knowledge graph as we constructed in an end-to-end manner, we build upon an item relation graph based on it with meta-path methods to describe product relations. Specifically, we first gather all 1-hop session pairs where concepts or temporal relations directly connect their intentions. To alleviate redundant connection and noise, we only keep session pairs if (1) they have no less than six distinct meta paths passed through commonsense relation nodes or (2) one can reach the other from all of its 1-hop concept nodes. Then, we construct a weight item graph $G = (V, E)$, where

| Datasets | Metric | FPMC | GRU4Rec | BERT4Rec | SASRec | SASRecF | CORE | SR-GNN | GCE-GNN | DIF-SR | FEARec | RIGRec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M2 (UK) | Recall@5 | 0.2523 | 0.2792 | 0.1899 | 0.3075 | 0.2957 | 0.2990 | 0.2928 | 0.3130 | 0.3128 | 0.3088 | **0.3342***$^{*}$ |
| | Recall@10 | 0.3121 | 0.3469 | 0.2641 | 0.3964 | 0.3713 | 0.3949 | 0.3678 | 0.4001 | 0.3990 | 0.3941 | **0.4229**$^{*}$ |
| | Recall@20 | 0.3696 | 0.4108 | 0.3349 | 0.4723 | 0.4406 | 0.4768 | 0.4381 | 0.4726 | 0.4739 | 0.4691 | **0.5003**$^{*}$ |
| | Recall@50 | 0.4389 | 0.4865 | 0.4197 | 0.5621 | 0.5245 | 0.5697 | 0.5171 | 0.5542 | 0.5598 | 0.5552 | **0.5863**$^{*}$ |
| | Recall@100 | 0.4841 | 0.5346 | 0.4744 | 0.6159 | 0.5771 | 0.6223 | 0.5662 | 0.6032 | 0.6171 | 0.6100 | **0.6398**$^{*}$ |
| | NDCG@5 | 0.1933 | 0.2118 | 0.1260 | 0.2121 | 0.2208 | 0.1673 | 0.2195 | 0.2214 | 0.2171 | 0.2138 | **0.2214** |
| | NDCG@10 | 0.2126 | 0.2327 | 0.1501 | 0.2406 | 0.2432 | 0.1985 | 0.2418 | 0.2441 | 0.2451 | 0.2415 | **0.2503**$^{*}$ |
| | NDCG@20 | 0.2272 | 0.2499 | 0.1682 | 0.2598 | 0.2634 | 0.2193 | 0.2616 | 0.2626 | 0.2641 | 0.2605 | **0.2703**$^{*}$ |
| | NDCG@50 | 0.2411 | 0.2648 | 0.1848 | 0.2679 | 0.2801 | 0.2379 | 0.2784 | 0.2807 | 0.2812 | 0.2777 | **0.2877**$^{*}$ |
| | NDCG@100 | 0.2484 | 0.2718 | 0.1937 | 0.2862 | 0.2887 | 0.2472 | 0.2865 | 0.2871 | 0.2891 | 0.2866 | **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.

$V$ is the node-set, $E$ is the edge set generated by linked items in all sampled session pairs, and the co-occurrence frequency is as weight. Here, we focus on evaluating the effectiveness of relations derived from our session-intention KG, leaving the exploration of optimal sub-graph sampling strategy to be explored in future work.

### 5.2 Graph Representations for Session Recommendation

Given the relation $G$, one can obtain item representations with various graph representation methods. We adopt the simple yet effective graph convolution operation to learn informative item representations, which is formulated as

$$\boldsymbol{E}^{l+1} = \boldsymbol{A}\boldsymbol{E}^{l},$$

where $\boldsymbol{A}$ is the adjacency matrix of $G$, $\boldsymbol{E} \in \mathbb{R}^{N\times d}$ is the $d$-dimensional embedding dictionary of all items. Canonical methods are utilized during graph representation learning, including light-weight convolution and sum pooling. After $L$-th convolution, the knowledge graph-based representations $\boldsymbol{E}^{L}$ can be directly utilized in the recommendation tasks for enhancing performance.

To achieve end-to-end recommendation, we designed a Relational Intention Knowledge Graph-based recommender named RIGRec, which seamlessly integrates the graph representations learning module into the session recommendation framework. Concretely, the widely used attention-based method SASRec is adopted as our session encoder, which aggregates the learned knowledge graph-based item representations within each session for user preference estimation.

### 5.3 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.

We employ two standard evaluation 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 $k \in \{5, 10, 20, 50, 100\}$. The implementation details are in the Appendix A.5.

### 5.4 Performance Comparison

Table 10 presents the recommendation performance of all models on the M2 dataset. An unpaired T-test with a p-value of 0.05 is conducted to prove the improvement is statistically significant. From these results, We have the following observations.

Our method consistently surpasses all baselines by a considerable margin regarding most metrics. Performance improvement should be attributed to

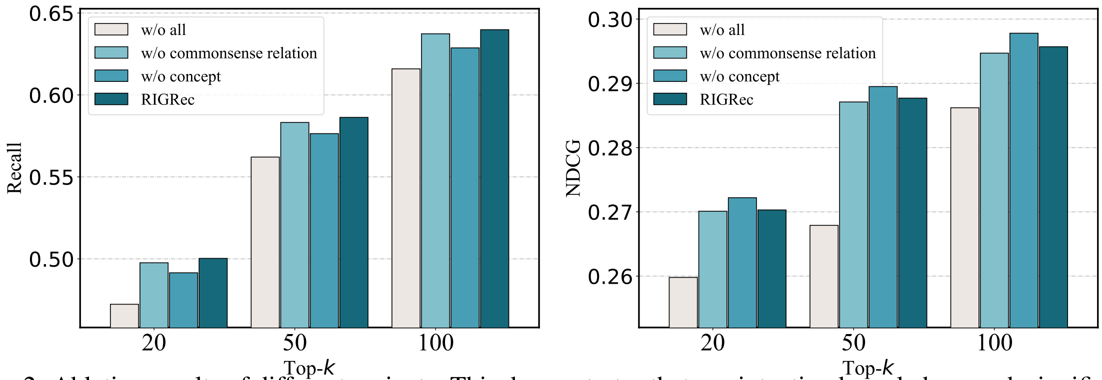



Figure 2: Ablation results of different variants. This demonstrates that our intention knowledge graph significantly enhances recommendation performance compared to SASRec. Both intention conceptualization and concept relations effectively improve results, with each type of relation contributing uniquely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

the major characteristic of our model, i.e., refining useful information from the intention knowledge graph to enhance recommendations. More persuasive evidence is the performance discrepancy between our RIGRec and SASRec, which serves as the backbone of our method, further verifying the effectiveness of our intention knowledge graph.

Graph-based methods SR-GNN and GCE-GNN present inferior effectiveness compared to our model, which is not a surprise since they only capture transition relation among items, quickly leading to unsatisfactory modeling. In contrast, we endow item representations with adequate complex dependencies acquired from our intention knowledge graph and thus achieve improvement. This justifies the necessity of enriching recommendation tasks with amply modeling user intentions.

Compared to most baselines, DIF-SR integrates side information for recommendation with an elaborately designed structure to achieve competitive performance. In contrast, another side information fusion method, SASRecF, exhibits performance deterioration compared to SASRec in terms of Recall@$k$. Such an observation is consistent with previous works (Xie et al., 2022; Liu et al., 2021), suggesting that inappropriate information fusion strategies may degenerate the model efficiency. Besides, noise and the inevitable increase of defect value with the expansion of the data scale also limit their performance. On the contrary, we resort to inferring user intentions with the aid of LLM, alleviating the aforementioned problems by reducing ineffectual information from raw features of items.

### 5.5 Ablation Study

To research the effectiveness of information provided by conceptualization and concepts relation of intentions, we conduct ablation studies by comparing our RIGRec with three variants generated by removing different types of edges as follows: (1) **w/o all**: This variant removes the whole item graph, equal to SASRec. (2) **w/o concept**: This variant removes edges acquired by the meta path through concepts. (3) **w/o commonsense relation**: This variant removes edges acquired by meta path through commonsense relation nodes.

Figure 2 presents the results of the ablation studies; we interpret the results with the following discoveries. First, our constructed intention knowledge graph is of vital use to produce helpful item representations for downstream recommendation tasks. As we can see, the performance of our model exceeds that of SASRec by a large margin. Second, both types of information contained in intention conceptualization and concepts relation prove efficient. It can be confirmed that the two variants solely with corresponding edges can outperform w/o all variants. Third, distinct relations provide dissimilar contributions to each metric, w/o commonsense relation presenting high Recall metrics. At the same time, w/o conceptualization performs better in NDCG metrics, which reaffirms the significance of constructing multi-type nodes and relations in our intention knowledge graph.

## 6 Conclusion

We present IGC-RC, a framework for automatically constructing intentional commonsense knowledge graphs from user behaviors. Using this framework, we built the Relational Intention Knowledge Graph (RIG) and validated its quality through extensive evaluations. Results show that RIG significantly enhances the performance of state-of-the-art session recommendation models.

## 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.

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. 2021. 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.

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.

Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*, volume 35, pages 4249–4256.

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.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2023c. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. *CoRR*, abs/2311.09174.

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.

# 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 |
|---|---|---|
| 2-gram | 0.0157 | **0.0570** |
| 3-gram | 0.0370 | **0.1851** |
| 4-gram | 0.0638 | **0.3461** |
| 5-gram | 0.0949 | **0.4896** |
| 6-gram | 0.1329 | **0.6061** |

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.

```
Below is a user 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 session intentions quality annotation, the intention relation classification annotation, and the 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:

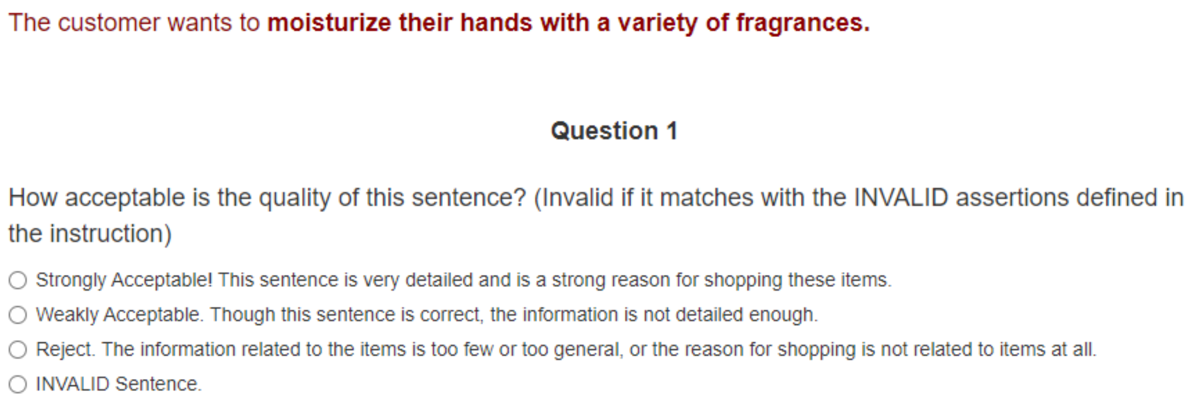
$$\text{Diversity}(D, n) = \frac{\#\text{ unique } n\text{-grams in } D^{\oplus}}{\#\text{ n-grams in } D^{\oplus}} \tag{1}$$

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. 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 generates more diverse intentions.

```
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.



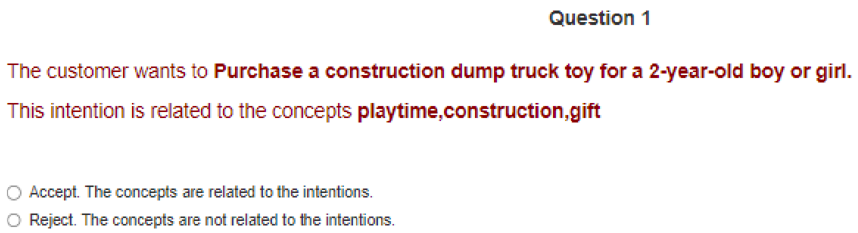
The customer wants to **moisturize their hands with a variety of fragrances.**

**Question 1**

How acceptable is the quality of this sentence? (Invalid if it matches with the INVALID assertions defined in the instruction)

- ○ Strongly Acceptable! This sentence is very detailed and is a strong reason for shopping these items.
- ○ Weakly Acceptable. Though this sentence is correct, the information is not detailed enough.
- ○ Reject. The information related to the items is too few or too general, or the reason for shopping is not related to items at all.
- ○ INVALID Sentence.

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



**Question 1**

The customer wants to **Purchase a construction dump truck toy for a 2-year-old boy or girl.**

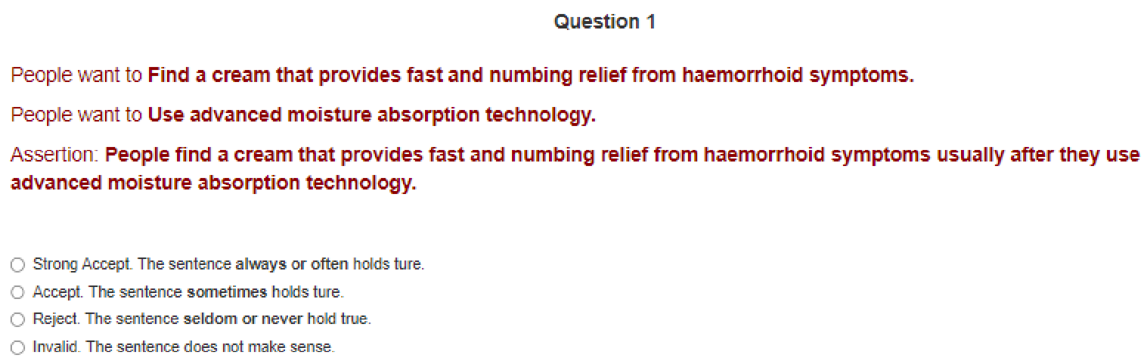
This intention is related to the concepts **playtime,construction,gift**

- ○ Accept. The concepts are related to the intentions.
- ○ Reject. The concepts are not related to the intentions.

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.



**Question 1**

People want to **Find a cream that provides fast and numbing relief from haemorrhoid symptoms.**

People want to **Use advanced moisture absorption technology.**

Assertion: **People find a cream that provides fast and numbing relief from haemorrhoid symptoms usually after they use advanced moisture absorption technology.**

- Strong Accept. The sentence **always or often** holds ture.
- Accept. The sentence **sometimes** holds ture.
- Reject. The sentence **seldom or never** hold true.
- Invalid. The sentence does not make sense.

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