

FolkScope: Intention Knowledge Graph Construction for E-commerce Commonsense Discovery

Changlong Yu^{1*}, Weiqi Wang¹, Xin Liu^{1*}, Jiaxin Bai^{1*}, Yangqiu Song^{1†}
Zheng Li², Yifan Gao², Tianyu Cao², Bing Yin²

¹The Hong Kong University of Science and Technology, Hong Kong SAR, China

²Amazon.com Inc, Palo Alto, USA

{cyuaq, wwangbw, xliucr, jbai, yqsong}@cse.ust.hk

{amzzhe, yifangao, caoty, alexbyin}@amazon.com

Abstract

Understanding users’ intentions in e-commerce platforms requires commonsense knowledge. In this paper, we present FolkScope, an intention knowledge graph construction framework to reveal the structure of humans’ minds about purchasing items. As commonsense knowledge is usually ineffable and not expressed explicitly, it is challenging to perform information extraction. Thus, we propose a new approach that leverages the generation power of large language models (LLMs) and human-in-the-loop annotation to semi-automatically construct the knowledge graph. LLMs first generate intention assertions via e-commerce-specific prompts to explain shopping behaviors, where the intention can be an open reason or a predicate falling into one of 18 categories aligning with ConceptNet, e.g., *IsA*, *MadeOf*, *UsedFor*, etc. Then we annotate plausibility and typicality labels of sampled intentions as training data in order to populate human judgments to all automatic generations. Last, to structure the assertions, we propose pattern mining and conceptualization to form more condensed and abstract knowledge. Extensive evaluations and studies demonstrate that our constructed knowledge graph can well model e-commerce knowledge and have many potential applications. Our codes and datasets are publicly available at <https://github.com/HKUST-KnowComp/FolkScope>.

1 Introduction

In e-commerce platforms, understanding users’ searching or purchasing intentions can benefit and motivate a lot of recommendation tasks (Dai et al., 2006; Zhang et al., 2016; Hao et al., 2022b). Intentions are mental states where agents or humans commit themselves to actions. Understanding others’ behaviors and mental states requires rationalizing intentional actions (Hutto and Ravenscroft,

2021), where we need commonsense, or, in other words, good judgements (Liu and Singh, 2004). For example, “at a birthday party, we usually need a birthday cake.” Meanwhile, commonsense knowledge can be *factoid* (Gordon et al., 2010), which is not invariably true, and is usually ineffable and not expressed explicitly. Existing intention-based studies on recommendation are either of limited numbers of intention categories (Dai et al., 2006; Zhang et al., 2016) or using models to implicitly model the intention memberships (Hao et al., 2022b). Thus, it is very challenging to acquire fine-grained intention knowledge in a scalable way.

Existing related knowledge graphs (KGs) can be categorized into two folds. First, some general situational commonsense KGs deal with everyday social situations (Rashkin et al., 2018; Sap et al., 2019; Zhang et al., 2020b), but they are not directly related to massive products on e-commerce platforms and thus not generalized well on users’ behavior data even for generative models, e.g., COMET (Bosselut et al., 2019). Second, most e-commerce KGs leverage existing KGs, such as ConceptNet (Liu and Singh, 2004; Speer et al., 2017) and Freebase (Bollacker et al., 2008), to integrate them into the e-commerce catalog data (Li et al., 2020a; Luo et al., 2020; Zalmout et al., 2021; Luo et al., 2021; Deng et al., 2022). However, such integration is still based on factual knowledge, such as *IsA* and *DirectorOf* relations, and does not truly model the commonsense knowledge for purchase intentions. Although some of these KGs may include information related to space, crowd, time, function, and event, they still fall short of modeling true commonsense knowledge (Luo et al., 2021).

Existing KGs constructed for e-commerce platforms can be evaluated for their factual knowledge in terms of *plausibility*. However, when it comes to purchasing intentions, a person’s beliefs and desires (Kashima et al., 1998) are mediated by their intentions, which can be reflected by the *typicality*

* Work done during internship at Amazon.

† Visiting academic scholar at Amazon.

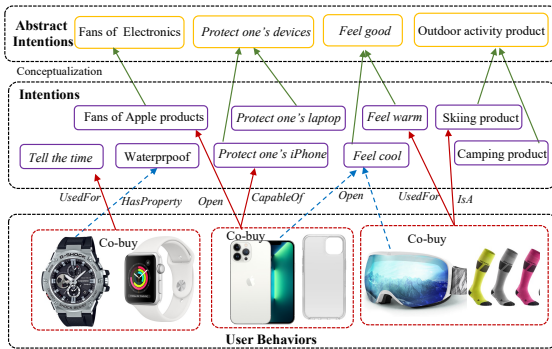


Figure 1: An overview of FolkScope. It starts from users’ purchasing or co-purchasing behaviors and links them to intentions. Then more abstract intentions are formed to condense the representation of intentions. The intentions can be noun phrases or verb phrases (*italics*).

of commonsense (Chalier et al., 2020; Wilhelm, 2022). For example, in Figure 1, a user bought an Apple watch because “Apple watches can be used for telling the time” where the reason is highly plausible (but other watches can also serve similar functions), whereas a more typical reason would be “apple watches are able to track running,” or “the user is simply a fan of Apple products.” Thus, no matter what kind of factual knowledge a KG contains, if it is not directly linked to rationalization, it cannot be regarded as typical commonsense. In addition, the task of explaining a user’s rating of an item has been proposed as a means of providing recommendations. To achieve this, researchers have suggested using online reviews as a natural source of explanation (Ni et al., 2019; Li et al., 2020b). However, online reviews are often noisy and diverse and may not directly reflect the user’s intention behind their purchase or rating. Instead, they may reflect the consequences of the purchase or the reasons behind the user’s rating. Existing sources of information, such as question-answering pairs, reviews, or product descriptions, do not explicitly mention the user’s intentions behind their purchases, making it a challenge to extract intentional commonsense knowledge for e-commerce. As a result, constructing an intention KG for e-commerce requires sophisticated information extraction techniques and thus remains challenging.

In this paper, we propose a new framework, FolkScope, to acquire intention knowledge in e-commerce. Instead of performing information extraction, we start from enormous user behaviors that entail sustainable intentions, such as *co-buy* behaviors, and leverage the generation power of large

language models (LLMs), e.g., GPT (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022), to generate possible intentions of the purchasing behaviors as candidates. LLMs have shown the capability of memorizing factual and commonsense knowledge (Petroni et al., 2019; West et al., 2022), and “sometimes infer approximate, partial representations of the beliefs, desires, and intentions possessed by the agent that produced the context” (Andreas, 2022). As open prompts in the above example can be arbitrary and loosely constrained, we also align our prompts with 18 ConceptNet relations, such as *IsA*, *HasPropertyOf*, *CapableOf*, *UsedFor*, etc. In addition, as the generated knowledge by LLMs can be noisy and may not be able to reflect human’s rationalization of a purchasing action, we also perform human annotation for *plausibility* and *typicality*.

Given generated candidates and annotations to construct the KG, we first perform pattern mining to remove irregular generations. Then we train classifiers to populate the prediction scores to all generated data. Finally, for each of the generated intentions, we perform conceptualization to map the key entities or concepts in the intention to more high-level concepts so that we can build a denser and more abstract KG for future generalization. An illustration of our KG is shown in Figure 1. To assess the overall quality of our KG, we randomly sample populated assertions and estimate their quality. Furthermore, we demonstrate the quality and usefulness of our KG by using it in a downstream task, CF-based (collaborative filtering) recommendation. The contributions of our work can be summarized as follows.

- We propose a new framework, FolkScope, to construct large-scale intention KG for discovering e-commerce commonsense knowledge.
- We leverage LLMs to generate candidates and perform two-step efficient annotation on Amazon data with two popular domains, and the process can be well generalized to other domains.
- We define the schema of the intention KG aligning with famous commonsense KG, ConceptNet, and populate a large KG based on our generation and annotation with 184,146 items, 217,108 intentions, 857,972 abstract intentions, and 12,755,525 edges (assertions).
- We perform a comprehensive study to verify the validity and usefulness of our KG.

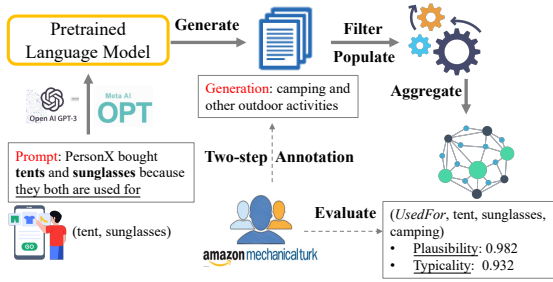


Figure 2: The overall framework of FolkScope. It includes the generation, population, and conceptualization to semi-automatically construct the e-commerce intention commonsense KG with the help of human-in-the-loop annotations and evaluation.

2 Methodology

2.1 Overview of FolkScope Framework

We call our framework FolkScope as we are the first attempt to reveal the structure of e-commerce intentional commonsense to rationalize purchasing behaviors. As shown in Figure 2, FolkScope is a human-in-the-loop approach for the semi-automatic construction of the KG. We first leverage the LLMs to generate candidate assertions of intentions for purchasing or co-purchasing behaviors based on *co-buy* data from the released Amazon dataset. Then we employ two-step annotations to annotate the plausibility and typicality of the generated intentions, where the corresponding definitions of scores are as follows.

- *Plausibility*: how possible the assertion is valid regarding their properties, usages, functions, etc.
- *Typicality*: how well the assertion reflects a specific feature that causes the user behavior. Typical intentional assertions should satisfy the following criteria. 1) Informativeness: contains key information about the shopping context rather than a general one, e.g., “they are used for Halloween parties.” v.s. “they are used for the same purpose.” 2) Causality: captures the typical intention of user behaviors, e.g., “they have a property of water resistance.” Some specific attributes or features might largely affect the users’ purchase decisions.

After the annotation, we design classifiers to populate prediction scores to all generated candidates. Then the high-quality ones will be further structured using pattern mining on their dependency parses to aggregate similar assertions. Then, we also perform conceptualization (Song et al., 2011; Zhang et al., 2022a) to further aggregate assertions to form more abstract intentions.

	Clothing	Electronics	Total
# Item Pairs	199,560	93,889	293,449
# Unique Items	151,509	64,244	211,349
# Assertions	11,358,637	5,282,273	16,640,910
# Uniq. Assertions	2,865,118	1,280,259	4,063,764
Avg. # Tokens	6.66	5.25	6.21

Table 1: Statistics of sampled co-buy pairs and generated candidate assertions. Note that the prompts in the generation are not included in the calculations of assertion lengths.

2.2 Knowledge Generation

User Behavior Data Sampling. We extract the users’ behavior datasets from open-sourced Amazon Review Data (2018)¹ (Ni et al., 2019) with 15.5M items from Amazon.com. In our work, we mainly consider *co-buy* pairs, which might indicate stronger shopping intent signals than *co-view* pairs. After the pre-processing and removing duplicated items, the resulting co-buy graph covers 3.5M nodes and 31.4M edges. The items are organized into 25 top-level categories from the Amazon website, and among them, we choose two frequent categories: “*Clothing, Shoes & Jewelry*” and “*Electronics*” to sample *co-buy* pairs because those items substantially appear in situations requiring commonsense knowledge to understand, while other categories such as “*Movie*” or “*Music*” are more relevant to factual knowledge between entities. We uniformly sample *co-buy* pairs from the two categories, and the statistics are shown in Table 1.

Prompted Generation. As shown in Table 2, we verbalize the prompt templates using the titles of co-buy pairs. Besides the general prompt (i.e., “open”), we also align our prompts with 18 relations in ConceptNet highly related to commonsense. For example, for the relation *HasA*, we can design a prompt “A user bought ‘item 1’ and ‘item 2’ because they both have [GEN]” where [GEN] is a special token indicating generation. Since the long item titles might contain noise besides useful attributes, we use heuristic rules to filter out items whose titles potentially affect the conditional generation, like repeated words. We use the OPT model (Zhang et al., 2022b) of 30B parameters² with two NVIDIA A100 GPUs based on the HuggingFace library (Wolf et al., 2020) to generate assertion candidates³. For each relation of the co-

¹<https://nijianmo.github.io/amazon/>

²<https://huggingface.co/facebook/opt-30b>

³As we will further annotate the plausibility and typicality of candidates, larger models will reduce annotation cost. However, the generation is also constrained by API or compu-

Type	Relation	Prompt
Open	/	/
Item	<i>HasA</i>	they both have
	<i>HasProperty</i>	they both have a property of
	<i>RelatedTo</i>	they both are related to
	<i>SimilarTo</i>	they both are similar to
	<i>PartOf</i>	they both are a part of
	<i>IsA</i>	they both are a type of
	<i>MadeOf</i>	they both are made of
	<i>CreatedBy</i>	they are created by
Function	<i>DistinctFrom</i>	they are distinct from
	<i>DerivedFrom</i>	they are derived from
	<i>UsedFor</i>	they are both used for
	<i>CapableOf</i>	they both are capable of
	<i>SymbolOf</i>	they both are symbols of
	<i>MannerOf</i>	they both are a manner of
Human	<i>DefinedAs</i>	they both are defined as
	<i>Result</i>	as a result, the person
	<i>Cause</i>	the person wants the
	<i>CauseDesire</i>	person wants his

Table 2: Prompts for different commonsense relations.

buy pairs, we set the max generation length as 100 and generate 3 assertions using nucleus sampling ($p = 0.9$) (Holtzman et al., 2020). We post-process the candidates as follows. (1) We discard the generations without one complete sentence. (2) We use the sentence segmenter from Spacy library⁴ to extract the first sentence for longer generations. After removing duplicates, we obtain 16.64M candidate assertions for 293K item pairs and 4.06M unique tails among them. The statistics of the two categories are listed in Table 1.

2.3 Two-step Annotation and Population

As the generated candidates can be noisy or not rational, we apply the human annotation to obtain high-quality assertions and then populate the generated assertions. We use Amazon Mechanical Turk (MTurk) to annotate our data. Annotators are provided with a pair of co-buy items with each item’s title, category, shopping URL, and three images from our sampled metadata. Assertions with different relations are presented in the natural language form by using the prompts presented in Table 2. More details are listed in Appendix A.

Annotation. To filter out incorrect candidates, we begin by annotating plausibility in the first step. This step serves as a preliminary filter and reduces the annotation cost for the subsequent steps. We randomly sample 66K generations and collect three plausibility votes per generated candidate. The final plausibility score is derived by majority voting. The overall IAA score is 75.48% in terms of pairwise agreement proportion, while Fleiss’s Kappa (Fleiss, 1971) is 0.4872. Both metrics are

tational cost. Thus, we choose the best model we can use.

⁴<https://spacy.io/>

Stage	Category	# Annotation	Avg. Score
Plausibility	Clothing	44,337	0.6435
	Electronics	21,760	0.5467
	Total	66,097	0.6116
Typicality	Clothing	38,279	0.4407
	Electronics	22,995	0.4631
	Total	61,274	0.4491

Table 3: Statistics of annotated data.

	Plausibility	Typicality
RoBERTa-large	83.22%	81.96%
DeBERTa-large	85.12%	82.67%

Table 4: Classification results on validation sets (F1).

satisfiable for such large-scale annotations.

Different from the simple binary plausibility judgments, in the second step, we have more fine-grained and precise typicality indicators concerning *informativeness* and *causality*. Here we choose the candidates automatically labeled as plausible based on our classifier trained on the first step’s data. We ask the annotators to judge whether they are *strongly acceptable* (+1), *weakly acceptable* (0.5), *rejected* (0), or *implausible* (-1) that the assertion is informative and casual for a purchasing behavior. Considering the judgments might be subjective and biased with respect to different annotators, we collect five annotations for each assertion and take the average as the final typicality score.⁵ Similar to the first step, we collect around 60K assertions. Empirically, we find annotating more data does not bring significantly better filtering accuracy. The statistics are presented in Table 3.

Population. For plausibility population, we train binary classifiers based on the majority voting results in the first step, which can produce binary labels of the plausibility of unverified generations. For the typicality score, as we take the average of five annotators as the score, we empirically use scores greater than 0.8 to denote positive examples and less than 0.2 as negative examples. We split the train/dev sets at the ratio of 80%/20% and train binary classifiers using both DeBERTa-large (He et al., 2021, 2023) and RoBERTa-large (Liu et al., 2019) as base models. The best models are selected to maximize the F1 scores on the validation sets, and results are shown in Table 4 (more results can be found in Appendix B). DeBERTa-large achieves better performance than RoBERTa-large on both

⁵The annotators in this step are chosen from the high-quality annotators in the first step. We tried other options, such as using seven or nine annotators per generation in our pilot study. The results do not show much improvement.

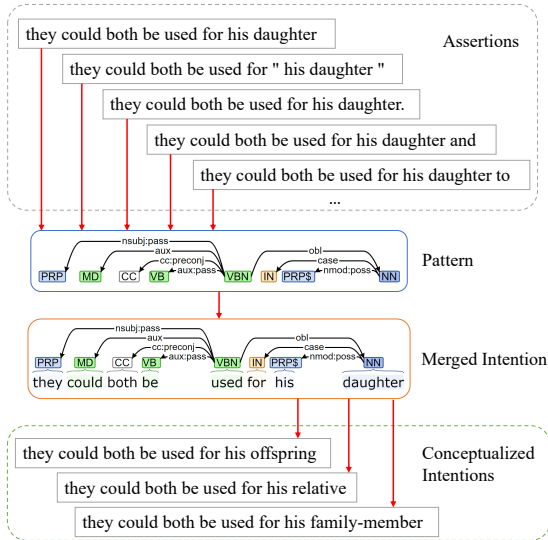


Figure 3: Illustration of knowledge aggregation.

plausibility and typicality evaluation. We populate the inference over the whole generated corpus in Table 1 and only keep the assertions whose predicted plausibility scores are above 0.5 (discarding 32.5% generations and reducing from 16.64M to 11.24M). Note that only plausible assertions are kept in the final KG. Using different confidence cutting-off thresholds leads to trade-offs between the accuracy of generation and the size of the corpus. After the two-step populations, we obtain the plausibility score and typicality score for each assertion. Due to the measurement of different aspects of knowledge, we observe low correlations between the two types of scores (Spearman correlation ρ : 0.319 for *clothing* and 0.309 for *electronics*).

2.4 Knowledge Aggregation

To acquire a KG with topology structures instead of sparse triplets, we aggregate semantically similar assertions. This is done by (1) pattern mining to align similar generated patterns and (2) conceptualization to produce more abstract knowledge.

Assertions are typically expressed as free-form text phrases, some of which may have similar syntax and semantics. By extracting the skeleton and necessary modifiers, such as demonstrative pronouns, adjectives, and adverbs, we can reduce the noise generated by these phrases. For example, as shown in Figure 3, several generations can be simplified to “they could both be used for his daughter,” despite the presence of punctuation and incomplete content. To achieve this, we employ frequent graph substructure mining over dependency parse trees

Threshold	Clothing		Electronics		Total	
	Accept	Size	Accept	Size	Accept	Size
0.5	83.73%	7,986,031	82.74%	3,250,605	83.40%	11,236,636
0.7	90.27%	7,346,160	88.27%	2,868,256	89.40%	10,214,416
0.8	91.02%	6,947,606	89.50%	2,650,625	90.00%	9,598,231
0.9	95.60%	6,167,315	94.87%	2,230,423	95.36%	8,397,738

Table 5: Acceptance ratios of plausible assertions and the corresponding sizes of populated assertions with different cutting-off thresholds.

to discover linguistic patterns (More details in Appendix C).

After pattern mining, we can formally construct our knowledge graph, where the head is a pair of items (p_1, p_2) , the relation r is one of the relations shown in Table 2, and the tail is an aggregated assertion e that is originally generated and then mapped to a particular one among 256 patterns. Each of the knowledge triples is associated with two populated scores, i.e., plausibility and typicality.

To produce abstract knowledge generalizable to new shopping contexts, we also consider the conceptualization with the large-scale concept KG, Probase (Wu et al., 2012; He et al., 2022; Wang et al., 2023b). The conceptualization process maps one extracted assertion e to multiple conceptualized assertions with concepts c . For example, in Figure 3, “they could be used for his daughter” can be conceptualized as “they could be used for his offspring,” “they could be used for his relative,” and “they could be used for his family-member,” etc. The conceptualization weight $P(c|e)$ can be determined by the likelihood for $\text{IsA}(e, c)$ in Probase. This process has been employed and evaluated by ASER 2.0 (Zhang et al., 2022a). Finally, we obtain a KG with 184,146 items, 217,108 intentions, 857,972 abstract intentions, and 12,755,525 edges to explain 236,739 co-buy behaviors, where 2,298,011 edges from the view of original assertions and 9,297,500 edges from the angle of conceptualized ones, and 1,160,014 edges model the probabilities of the conceptualization.

3 Intrinsic Evaluations

In this section, we present some examples of our constructed KG and conduct comprehensive intrinsic evaluations of KG.

3.1 Examples in KG

We show two examples of co-purchasing products and their corresponding knowledge (§ 2.2) as well as populated scores (§ 2.3) in Table 7. We measure the quality of assertions using both plausibility and

Relation	Acc. Rate	# Edges	# Tails	Avg. Length
<i>Open</i>	87.54%	703,059	151,748	7.86
<i>HasA</i>	94.08%	710,331	68,516	5.53
<i>HasProperty</i>	79.13%	317,938	133,877	5.00
<i>RelatedTo</i>	91.89%	571,918	130,551	3.08
<i>SimilarTo</i>	86.35%	685,737	18,603	3.53
<i>PartOf</i>	79.60%	674,928	114,983	4.36
<i>IsA</i>	89.05%	591,037	98,262	3.82
<i>MadeOf</i>	90.05%	528,289	70,246	5.06
<i>CreatedBy</i>	95.15%	267,459	74,920	3.93
<i>DistinctFrom</i>	91.74%	861,929	80,295	4.66
<i>DerivedFrom</i>	85.54%	444,131	61,696	4.90
<i>UsedFor</i>	91.79%	630,462	45,206	2.58
<i>CapableOf</i>	87.73%	681,480	101,170	5.23
<i>SymbolOf</i>	78.04%	809,196	52,075	3.46
<i>MannerOf</i>	89.44%	371,892	122,829	4.38
<i>DefinedAs</i>	85.59%	288,411	151,986	6.31
<i>Result</i>	44.79%	568,523	166,018	8.80
<i>Cause</i>	80.50%	696,392	185,042	7.06
<i>CauseDesire</i>	67.23%	833,524	155,422	5.61
<i>Total</i>	83.40%	11,236,636	1,874,782	5.02

Table 6: Evaluation on plausible rate and size of the populated KG. The prompts in the generation are not included in the calculations of assertion lengths.

typicality scores, which are again shown they are not correlated. For example, “they are *SimilarTo* the product they bought” for the first pair and “they are *DistinctFrom* other similar products” for the second pair are plausible assertions but not typical explanations of why a user would buy them together. Moreover, some of the open relations are very good as well. Take the second pair as an example: the open relation shows “he was worried about his baby’s skin” as both products are related to baby skin protection. We also append more typical knowledge examples in Table 14 of the Appendix.

3.2 Human Evaluation

As we populate the whole generated assertions using classifiers based on DeBERTa-large model, we conducted human evaluations by sampling a small number of populated assertions from different scales of predicted scores to evaluate the effectiveness of the knowledge population.

3.2.1 Plausibility Evaluation

We randomly sample 200 plausible assertions from each relation in each of the clothing and electronics domains to test the human *acceptance rate*. The annotation is conducted in the same way as the construction step. As we only annotate assertions predicted to be greater than the 0.5 plausibility score, the IAA is above 85%, even greater than the one in the construction step. As shown in Table 5, different cutting-off thresholds (based on the plausibility score by our model) lead to the trade-offs between the accuracy and the KG size. Overall, FolkScope can achieve an 83.4% acceptance rate with a default threshold (0.5). To understand what

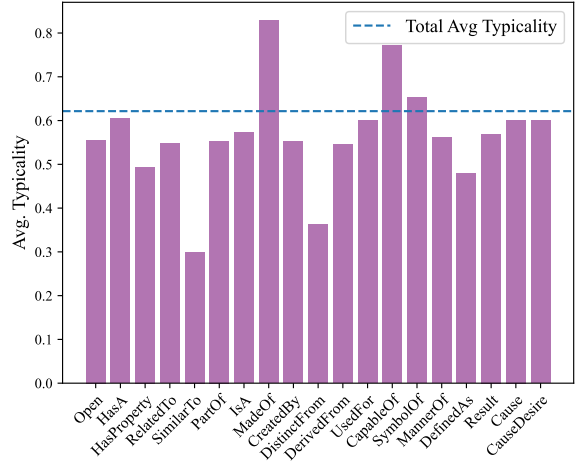


Figure 4: Average typicality score of each relation in the populated KG with the cutting-off threshold 0.8.

is filtered, we manually check the generations with low plausibility scores and find that OPT can generate awkward assertions, such as simply repeating the item titles or obviously logical errors regarding corresponding relations. Our classifier trained on annotated datasets helps resolve such cases. Using a larger threshold of 0.9, we attain a 95.35% acceptance rate, a nearly 11.96% improvement while still keeping above 8M plausible assertions. We also report the accuracy in terms of different relations in Table 6. We can observe that assertions concerning the relations of human beings’ situations like *Cause*, *Result*, and *CauseDesire* have relatively lower plausibility scores and longer lengths than the relations of items’ property, function, etc. This is because there exist some clues about items’ knowledge in the item titles, while it is much harder to generate (or guess) implicit human beings’ causal reasons using language generation.

3.2.2 Typicality Evaluation

The goal of the typicality population is to precisely recognize high-quality knowledge, and we evaluate whether assertions with high typicality scores are truly good ones. We randomly sample 200 assertions from each relation whose predicted typicality scores are above 0.8 for human evaluation. Each of the assertions is again annotated by five AMT workers, and the average rating is used. The results are shown in Table 8. It shows that average annotated scores are lower than the predicted ones due to harder judgments for typicality. Similarly, predicted typicality scores are less accurate than plausibility. Especially the typicality score will be





Item 1	Item 2	Relation	Tail	P.	T.
GGS III LCD Screen Protector glass for CANON 5D Mark III (link) 	ECC5D3B Secure Grip Camera Case for Canon 5D Mark III (link) 	<i>Open</i> <i>HasProperty</i> <i>SimilarTo</i> <i>PartOf</i> <i>UsedFor</i> <i>SymbolOf</i> <i>DefinedAs</i>	they can be used for the same purpose "easy to install" and "easy to remove" the product he bought his camera gear protect the camera from scratches and dust his love for his camera "Camera Accessories" on Amazon.com	0.67 0.80 0.95 0.93 0.97 0.99 0.99	0.35 0.85 0.09 0.99 0.99 0.88 0.67
Sun Smarties Baby UPF 50+ Non-Skid Sand Water Socks Pink (link) 	Schylling UV Play Shade, SPF 50+, Ultra portable, Blue (link) 	<i>Open</i> <i>SimilarTo</i> <i>DistinctFrom</i> <i>UsedFor</i> <i>CapableOf</i> <i>DefinedAs</i> <i>Result</i> <i>Cause</i>	he was worried about his baby's skin each other other similar products baby's outdoor activities blocking harmful UV rays sun protection products enjoy the sun sagely and comfortably want to use them for his/her baby	0.98 0.74 0.97 0.85 0.97 0.87 0.97 0.99	0.98 0.01 0.10 0.91 0.99 0.81 0.98 0.94

Table 7: Two examples from the constructed knowledge graph. "P." and "T." stand for the predicted plausibility and typicality scores. Generated tails with high typicality (in green) and low typicality (in red) scores are highlighted.

Threshold	Aggregated Knowledge	Conceptualization
0.8	0.6215	0.4571
0.9	0.6335	0.5567
0.99	0.7028	0.5775

Table 8: Average annotated typicality scores for assertions after pattern mining and conceptualization with different thresholds of predicted typicality scores.

further decreased after conceptualization. This is because, first, the conceptualization model may introduce some noise, and second, the more abstract knowledge tends to be less typical when asking humans to annotate. We also show the typicality scores of each relation in Figure 4. Different from plausibility, *SimilarTo*, *DistinctFrom*, *DefinedAs*, and *HasPropertyOf* are less typical compared to other relations. They describe items' general features but can not well capture typical purchasing intentions though they have high plausibility scores, whereas *CapableOf* and *MadeOf* are the most typical features that can explain purchasing intentions for the two domains we are concerned about.

More evaluation on the diversity of implicit generation and fine-grained subcategory knowledge aggregation can be found in Appendix D.

4 Extrinsic Evaluation

4.1 Experimental Setup

Data Preparation. We conduct extrinsic evaluation via knowledge-augmented recommendation tasks. Specifically, we use the same categories' user-item interaction data from the Amazon Review dataset (Ni et al., 2019) shown in Table 9. We split datasets into train/dev/test sets at a ratio of 8:1:1 and report averaged RMSE (root mean square error) scores over five runs.

To fairly evaluate the KG for recommendations, we sample the sub-graph from the original KG

	Clothing	Electronics
# Users	782,144	486,349
# Items	18,042	6,166
# Interactions	1,579,499	1,056,406
Density	0.011%	0.035%

Table 9: Statistics of the recommendation datasets.

where *co-buy* pairs are simultaneously purchased by at least one user in the recommendation training set. The detailed statistics of the matched KG are in Table 10. The item coverage computes the percentage of the items in the recommendation dataset that are covered by the matched KG. Moreover, we also filter the matched KG with the threshold of 0.5 or 0.9 on *plausibility* and *typicality* scores to evaluate the effectiveness of the knowledge population. From Table 10, we can observe the number of edges essentially reduces when the filters are applied, but the coverage of the items does not drastically drop.

Knowledge Representation. As our constructed KG can be represented as the triplet $((p_1, p_2), r, e)$, where the head (p_1, p_2) is the co-buy pair, the relation r is from relations in Table 2 and e refer to generated tails. To combine both structural and textual information from KG, we modify the original TransE model (Bordes et al., 2013) to the following objective:

$$\mathcal{L} = \gamma + d\left(\frac{p_1 + p_2}{2} + r, e\right) - d\left(\frac{p'_1 + p'_2}{2} + r, e\right)$$

where γ is a margin parameter, and p_1, p_2, p'_1, p'_2 are item embeddings for positive head (p_1, p_2) , and negative corrupted head (p'_1, p'_2) . Meanwhile, r is the relation embedding for the relation r , e is the embedding for the tail e , and the function d is Euclidean distance. Moreover, the node embeddings for e are initialized by Sentence-BERT (Reimers

Knowledge Graph	Clothing		Electronics	
	# Edges	Coverage	# Edges	Coverage
Matched Knowledge Graph	432,119	79.83%	117,836	82.40%
+ Plau. >0.5	323,263	79.83%	78,908	82.40%
+ Plau. >0.5 and Typi. >0.5	141,422	79.67%	40,978	80.20%
+ Plau. >0.9	269,210	79.83%	58,013	82.39%
+ Plau. >0.9 and Typi. >0.9	103,262	79.36%	27,288	76.94%

Table 10: Details of matched KG subsets. “Plau.” means plausibility and “Typi” means typicality.

and Gurevych, 2019) representations. After training the modified TransE model, all the item embeddings \mathbf{p} can be used as extra features to enhance recommendations.

4.2 Experimental Results

Baselines. We adopt commonly-used NCF (He et al., 2017) and Wide&Deep model (Cheng et al., 2016) as our baselines. As our goal is to evaluate the effectiveness of features derived from KG, we leave advanced KG fusion methods, such as hyperedges or meta path-enhanced, to future work.

Ablation Study. We conduct two ablation studies to evaluate the effect of structural information provided by the co-buy pairs and the semantic information provided by the tails’ text only. For the former, we train a standard TransE model solely on co-buy pairs to learn the graph embeddings of items. For the latter, for each item in the matched KG, we conduct average pooling of its neighbor tails’ Sentence-BERT embeddings as its semantic representations. The experimental results are shown in Table 11, and we have the following observations. First, the textual information contained in intentional assertions is useful for product recommendations. This can be testified as the W&D model can perform better even when only features of the assertions are provided. Second, our KG, even before annotations and filtering, can produce better item embeddings than solely using the co-buy item graphs. As we can see, the performance of our matched KG is better than that of the co-buy pair graphs. Third, the two-step annotation and population indeed help improve the item embeddings for recommendations. The higher the scores are, the larger improvement the recommendation system obtains.

5 Related Work

Knowledge Graph Construction. An early approach of commonsense KG construction is proposed in ConceptNet (Liu and Singh, 2004; Speer

Method	Clothing	Electronics
NCF (He et al., 2017)	1.117	1.086
W&D (Cheng et al., 2016)	1.104	1.071
+ Co-Buy Structure Only	1.096	1.067
+ Textual Features Only	1.093	1.068
+ Matched Knowledge Graph	1.093	1.058
+ Plau. > 0.5	1.087	1.060
+ Plau. > 0.5 and Typi. > 0.5	1.081	1.053
+ Plau. > 0.9	1.086	1.053
+ Plau. > 0.9 and Typi. > 0.9	1.081	1.052

Table 11: Recommendation results in RMSE.

et al., 2017) where both text mining and crowdsourcing are leveraged. In 2012, a web-scale KG, Probase, which focuses on *IsA* relations (Yu et al., 2020), is constructed based on pattern mining (Wu et al., 2012), which can model both plausibility and typicality of conceptualizations (Song et al., 2011). Recently, situational commonsense knowledge, such as Event2Mind (Rashkin et al., 2018) and ATOMIC (Sap et al., 2019), has attracted more attention in the field of AI and NLP. Then their extensions and neural generative models are developed (Bosselut et al., 2019; Hwang et al., 2021). Meanwhile, information extraction can be used to extract event-related knowledge from large-scale corpora, such as KnowlyWood (Tandon et al., 2015), WebChild (Tandon et al., 2017), and ASER (Zhang et al., 2020b, 2022a). The extracted knowledge can then be transferred to other human-annotated knowledge resources (Zhang et al., 2020a; Fang et al., 2021b,a).

In e-commerce, Amazon Product Graph (Zalmout et al., 2021) is developed to align Amazon catalog data with external KGs such as Freebase and to automatically extract thousands of attributes in millions of product types (Karamanolakis et al., 2020; Dong et al., 2020; Zhang et al., 2022c). Alibaba also develops a series of KGs including AliCG (Zhang et al., 2021), AliCoCo (Luo et al., 2020, 2021), AliMeKG (Li et al., 2020a), and OpenBG (Deng et al., 2022; Qu et al., 2022). As we have stated in the introduction, there is still a gap between collecting factual knowledge about products and modeling users’ purchasing intentions.

Language Models as Knowledge Bases. Researchers have shown LLMs trained on large corpus encode a significant amount of knowledge in their parameters (AlKhamissi et al., 2022; Ye et al., 2022). LLMs can memorize factual and commonsense knowledge, and one can use prompts (Liu et al., 2023) to probe knowledge from them (Petroni et al., 2019). It has been shown that we can derive factual KGs at scale based on LLMs for factual

knowledge (Wang et al., 2020; Hao et al., 2022a) and distill human-level commonsense knowledge from GPT3 (West et al., 2022). None of the above KGs are related to products or purchasing intention. We are the first to propose a complete KG construction pipeline from LLMs and several KG refinement methods for e-commerce commonsense discovery.

6 Conclusion

In this paper, we propose a new framework, FolkScope, to acquire intention commonsense knowledge for e-commerce behaviors. We develop a human-in-the-loop semi-automatic way to construct an intention KG, where the candidate assertions are automatically generated from large language models, with carefully designed prompts to align with ConceptNet commonsense relations. Then we annotate both plausibility and typicality scores of sampled assertions and develop models to populate them to all generated candidates. Then the high-quality assertions will be further structured using pattern mining and conceptualization to form more condensed and abstractive knowledge. We conduct extensive evaluations to demonstrate the quality and usefulness of our constructed KG. In the future, we plan extend our framework to multi-domain, multi-behavior type, multilingual (Huang et al., 2022; Wang et al., 2023a) and temporal (Wang et al., 2022b,a) scenarios for empowering more e-commerce applications.

Limitations

We outline two limitations of our work from *user behavior sampling* and *knowledge population* aspects. Due to huge-volume user behavior data produced every day in the e-commerce platform, it is crucial to efficiently sample significant behaviors that can indicate strong intentions and avoid random co-purchasing or clicking etc. Though in this work we adopt the criteria of selecting nodes whose degree are more than five in the *co-buy* graph, it is still coarse-grained and more advanced methods remain to be explored in order to sample representative co-buy pairs for intention generation. Some potential solutions are to aggregate frequent *co-buy* category pairs and then sample product pairs within selected category pairs. Moreover, our proposed framework can be generalized to other types of abundant user behaviors such as *search-click* and *search-buy*, which requires to design corresponding

prompts. We leave these designs to future work.

For open text generation from LLMs, it becomes common practices to label high-quality data for finetuning to improve the quality and controllability of generation such as LaMDA (Thoppilan et al., 2022), InstructGPT (Ouyang et al., 2022), and ChatGPT⁶. However, computation cost is the major bottleneck to use annotated data as human feedback for language model finetuning with billions of parameters, like OPT-30b in our work. Hence we adopt a trade-off strategy to populate human judgments by training effective classifiers and conducting inferences over all the generation candidates. With impressive generation performance of ChatGPT, we expect efficient methods to directly optimize LLMs with human feedback in more scalable way like reinforcement learning (RLHF), and enable LLMs to generate more typical intention knowledge with less annotation efforts.

Ethics Statement

As our proposed framework relied on large language models, text generation based on LLMs often contains biased or harmful contexts. We argue that our work largely mitigated the potential risks in the following ways. First, our careful-designed prompting leads to rather narrow generations constrained on small domains, i.e., products in e-commerce. Second, we also had a strict data audit process for annotated data from annotators and populated data from trained classifiers. On a small scale of inspections, we found none belongs to significant harmful contexts. The only related concern raised here is that some generated knowledge is irrelevant to the products themselves. The major reason is due to imprecise product titles written by sellers for search engine optimization, such as adding popular keywords to attract clicks or purchases. Our human-in-the-loop annotation identified such cases and the trained classifier further assisted machines in detecting bias, as we hope our intention generations can be safe and unbiased as much as possible.

Acknowledgements

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong, the MHKJFS (MHP/001/19) from

⁶<https://openai.com/blog/chatgpt/>

ITC of Hong Kong and the National Key R&D Program of China (2019YFE0198200) with special thanks to HKMAAC and CUSBLT. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

- Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. [A review on language models as knowledge bases](#). *CoRR*, abs/2204.06031.
- Jacob Andreas. 2022. [Language models as agent models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5769–5779. Association for Computational Linguistics.
- 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, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.
- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. [Translating embeddings for modeling multi-relational data](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. [COMET: commonsense transformers for automatic knowledge graph construction](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Yohan Chalier, Simon Razniewski, and Gerhard Weikum. 2020. [Joint reasoning for multi-faceted commonsense knowledge](#). In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.
- Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishikesh Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. [Wide & deep learning for recommender systems](#). In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2016, Boston, MA, USA, September 15, 2016*, pages 7–10. ACM.
- Honghua (Kathy) Dai, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, and Ying Li. 2006. [Detecting online commercial intention \(OCI\)](#). In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 829–837. ACM.
- Shumin Deng, Chengming Wang, Zhoubo Li, Ningyu Zhang, Zelin Dai, Hehong Chen, Feiyu Xiong, Ming Yan, Qiang Chen, Mosha Chen, Jiaoyan Chen, Jeff Z. Pan, Bryan Hooi, and Huajun Chen. 2022. [Construction and applications of billion-scale pre-trained multimodal business knowledge graph](#). *CoRR*, abs/2209.15214.
- Xin Luna Dong, Xiang He, Andrey Kan, Xian Li, Yan Liang, Jun Ma, Yifan Ethan Xu, Chenwei Zhang, Tong Zhao, Gabriel Blanco Saldana, Saurabh Deshpande, Alexandre Michetti Manduca, Jay Ren, Surenender Pal Singh, Fan Xiao, Haw-Shiuan Chang, Giannis Karamanolakis, Yuning Mao, Yaqing Wang, Christos Faloutsos, Andrew McCallum, and Jiawei Han. 2020. [Autoknow: Self-driving knowledge collection for products of thousands of types](#). In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2724–2734. ACM.
- Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. [Benchmarking commonsense knowledge base population with an effective evaluation dataset](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.
- Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. [DISCOS: bridging the gap between discourse knowledge and commonsense knowledge](#). In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

- Jonathan Gordon, Benjamin Van Durme, and Lenhart K. Schubert. 2010. [Learning from the web: Extracting general world knowledge from noisy text](#). In *Collaboratively-Built Knowledge Sources and Artificial Intelligence, Papers from the 2010 AAAI Workshop, Atlanta, Georgia, USA, July 11, 2010*, volume WS-10-02 of AAAI Technical Report. AAAI.
- Shibo Hao, Bowen Tan, Kaiwen Tang, Hengzhe Zhang, Eric P. Xing, and Zhiting Hu. 2022a. [Bertnet: Harvesting knowledge graphs from pretrained language models](#). *CoRR*, abs/2206.14268.
- Zhenyun Hao, Jianing Hao, Zhaohui Peng, Senzhang Wang, Philip S. Yu, Xue Wang, and Jian Wang. 2022b. [Dy-hien: Dynamic evolution based deep hierarchical intention network for membership prediction](#). In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 363–371. ACM.
- Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. [Acquiring and modelling abstract commonsense knowledge via conceptualization](#). *CoRR*, abs/2206.01532.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: decoding-enhanced bert with disentangled attention](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. [Neural collaborative filtering](#). In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 173–182. ACM.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, and Wei Wang. 2022. [Multilingual knowledge graph completion with self-supervised adaptive graph alignment](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 474–485. Association for Computational Linguistics.
- Daniel Hutto and Ian Ravenscroft. 2021. Folk Psychology as a Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University.
- 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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. [Survey of hallucination in natural language generation](#). *CoRR*, abs/2202.03629.
- Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. [Textract: Taxonomy-aware knowledge extraction for thousands of product categories](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8489–8502. Association for Computational Linguistics.
- Yoshihisa Kashima, Allison McKintyre, and Paul Clifford. 1998. The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1(3):289–313.
- Feng-Lin Li, Hehong Chen, Guohai Xu, Tian Qiu, Feng Ji, Ji Zhang, and Haiqing Chen. 2020a. [Alimekg: Domain knowledge graph construction and application in e-commerce](#). In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2581–2588. ACM.
- Lei Li, Yongfeng Zhang, and Li Chen. 2020b. [Generate neural template explanations for recommendation](#). In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 755–764. ACM.
- Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Comput. Surv.*, 55(9):195:1–195:35.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.

- Xusheng Luo, Le Bo, Jinhang Wu, Lin Li, Zhiy Luo, Yonghua Yang, and Keping Yang. 2021. [Alicoco2: Commonsense knowledge extraction, representation and application in e-commerce](#). In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 3385–3393. ACM.
- Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinhang 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 Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 313–327. ACM.
- Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 188–197. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *NeurIPS*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.
- Yincen Qu, Ningyu Zhang, Hui Chen, Zelin Dai, Chengming Wang, Xiaoyu Wang, Qiang Chen, and Huajun Chen. 2022. [Commonsense knowledge salience evaluation with a benchmark dataset in e-commerce](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 14–27. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. [Event2mind: Commonsense inference on events, intents, and reactions](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 463–473. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- 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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.
- Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hong-song Li, and Weizhu Chen. 2011. [Short text conceptualization using a probabilistic knowledgebase](#). In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2330–2336. IJCAI/AAAI.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.
- Niket Tandon, Gerard de Melo, Abir De, and Gerhard Weikum. 2015. [Knowlywood: Mining activity knowledge from hollywood narratives](#). In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 223–232. ACM.
- 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, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pages 115–120. Association for Computational Linguistics.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao

- Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguer-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. [Lamda: Language models for dialog applications](#). *CoRR*, abs/2201.08239.
- Luke Vilnis, Zachary Fisher, Bhargav Kanagal, Patrick Murray, and Sumit Sanghai. 2022. [Impakt: A dataset for open-schema knowledge base construction](#). *CoRR*, abs/2212.10770.
- Chenguang Wang, Xiao Liu, and Dawn Song. 2020. [Language models are open knowledge graphs](#). *CoRR*, abs/2010.11967.
- Ruijie Wang, Zheng Li, Dachun Sun, Shengzhong Liu, Jinning Li, Bing Yin, and Tarek F. Abdelzaher. 2022a. [Learning to sample and aggregate: Few-shot reasoning over temporal knowledge graphs](#). In *NeurIPS*.
- Ruijie Wang, Zheng Li, Jingfeng Yang, Tianyu Cao, Chao Zhang, Bing Yin, and Tarek F. Abdelzaher. 2023a. [Mutually-paced knowledge distillation for cross-lingual temporal knowledge graph reasoning](#). In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2621–2632. ACM.
- Ruijie Wang, Zheng Li, Danqing Zhang, Qingyu Yin, Tong Zhao, Bing Yin, and Tarek F. Abdelzaher. 2022b. [RETE: retrieval-enhanced temporal event forecasting on unified query product evolutionary graph](#). In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 462–472. ACM.
- 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](#). *CoRR*, abs/2305.04808.
- Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. [Symbolic knowledge distillation: from general language models to commonsense models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.
- Isaac Wilhelm. 2022. Typical: A theory of typicality and typicality explanation. *The British Journal for the Philosophy of Science*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.
- Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. [Probase: a probabilistic taxonomy for text understanding](#). In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.
- Xifeng Yan and Jiawei Han. 2002. [gspan: Graph-based substructure pattern mining](#). In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pages 721–724. IEEE Computer Society.
- Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. [MAVE: A product dataset for multi-source attribute value extraction](#). In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1256–1265. ACM.
- Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. [Generative knowledge graph construction: A review](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1–17. Association for Computational Linguistics.
- Changlong Yu, Jialong Han, Peifeng Wang, Yangqiu Song, Hongming Zhang, Wilfred Ng, and Shuming Shi. 2020. [When hearst is not enough: Improving hypernymy detection from corpus with distributional models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6208–6217. Association for Computational Linguistics.
- 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 *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 4090–4091. ACM.
- Chenwei Zhang, Wei Fan, Nan Du, and Philip S. Yu. 2016. [Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach](#). In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016*,

Montreal, Canada, April 11 - 15, 2016, pages 1373–1384. ACM.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. [Transomcs: From linguistic graphs to commonsense knowledge](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4004–4010. ijcai.org.

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

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

Ningyu Zhang, Qianghui 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 *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 3895–3905. ACM.

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

Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. 2022c. [Oa-mine: Open-world attribute mining for e-commerce products with weak supervision](#). In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3153–3161. ACM.

Assertion 1 (similarTo):

PersonX bought a product of Item A and a product of Item B because they both are similar to what he needs.

Question 1 Plausibility

Is this a plausible assertion? (Implausible if it does not make sense to you)

Yes, this sentence satisfies all the requirements of a plausible assertion.

No, this sentence is implausible as it falls into one of the reasons that make an implausible assertion.

Unfamiliar with Item A or B, so I can't decide.

Figure 5: The question card in our plausibility annotation round.

Appendix

A Annotation Guideline

Workers satisfying the following three requirements are invited to participate: (1) at least 90% lifelong HITs approval rate, (2) at least 1,000 HITs approved, and (3) achieving 80% accuracy on at least 10 qualification questions, which are carefully selected by authors of this paper. Qualified workers will be further invited to annotate 16 tricky assertions. Based on workers’ annotations, they will receive personalized feedback containing explanations of the errors they made along with advice to improve their annotation accuracy. Workers surpassing these two rounds are deemed qualified for main-round annotations. To avoid spamming, experts will provide feedback for all workers based on a sample of their main rounds’ annotations from time to time. Finally, we recruited more than 100 workers in the us-east district. It takes \$0.2 on average for each assertion, and the annotators are paid \$7.7 per hour on average, which satisfies the local minimum wage under local laws.

We conducted human annotations and evaluations on the Amazon Mechanical Turk as Figure 5 for the first-step plausibility annotation and as Figure 6 for the second-step typicality annotation.

B Knowledge Population

Using different confidence cutting-off thresholds leads to trade-offs between the accuracy of generation and the size of the corpus. Higher values result in conservative selections that favor precision over recall, whereas lower ones tend to recall more plausible assertions. We plotted four cutoff points in Figure 7.

C Pattern Mining Details

We apply the frequent graph substructure mining algorithm over dependency parse trees to discover

Assertion 1 (capableOf):

PersonX bought a product of Item A and a product of Item B because they both are capable of providing him comfort and joy.

Question 1

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

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

Figure 6: The question card in our typicality annotation round.

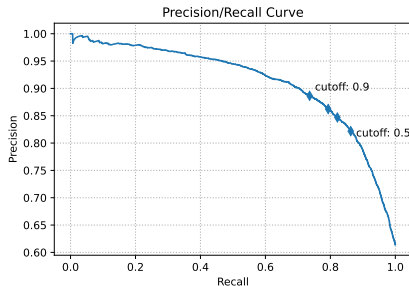


Figure 7: The precision-recall curve of our plausibility population classifier on the human-labeled validation set. The annotated points show the different thresholds (cutoffs) to filter the generated assertions, i.e. from left to right: 0.9, 0.8, 0.7, 0.5 respectively.

the linguistic patterns. We sample 90,000 candidates for each relation to analyze patterns and then parse each candidate into a dependency tree. In addition, the lemmatized tokens, pos-tags, and named entities are acquired for further use. To reduce the time complexity of pattern mining, we mine high-frequency patterns for each relation. To meet the two requirements of the knowledge with high precision but non-trivial, patterns are required to perfectly match more than 500 times. One perfect match means that this pattern is the longest pattern, and no other candidate patterns can match. Therefore, the pattern mining pipeline consists of three passes: (1) a graph pattern mining algorithm, Java implementation of gSpan (Yan and Han, 2002),⁷ to mine all candidate patterns with the frequency more than 500, (2) a subgraph isomorphism algorithm, C++ implementation of VF2 algorithm in igraph,⁸ with a longest-first greedy strategy to check the perfect match frequency, and (3) human evaluation and revision. Finally, we obtain 256 patterns that cover 80.77% generated candidates. Details can be found in Table 12.

⁷<https://github.com/timtdah/parsemis>

⁸<https://igraph.org/>

Type	Relation	# of Patterns	Coverage
Item	<i>RelatedTo</i>	14	96.94
	<i>IsA</i>	15	97.20
	<i>HasA</i>	12	99.30
	<i>PartOf</i>	8	99.83
	<i>MadeOf</i>	13	99.45
	<i>SimilarTo</i>	7	22.22
	<i>CreatedBy</i>	14	98.59
	<i>HasProperty</i>	16	63.20
	<i>DistinctFrom</i>	9	97.30
	<i>DerivedFrom</i>	20	100.00
Function	<i>UsedFor</i>	2	96.57
	<i>CapableOf</i>	13	74.68
	<i>DefinedAs</i>	27	95.99
	<i>SymbolOf</i>	9	99.76
	<i>MannerOf</i>	34	98.56
Human	<i>Cause</i>	21	93.68
	<i>Result</i>	0	0
	<i>CauseDesire</i>	0	0
Overall	<i>/</i>	256	80.77

Table 12: Frequent linguistic patterns and corresponding coverage on human-annotated knowledge.

D More Evaluations

D.1 Implicit Generation Evaluation

As we know, language model based generation capture spurious correlation given the condition of the generation (Ji et al., 2022). Hence we simply quantify the diversity as the novelty ratio of generated tails not appearing in the item titles, i.e., novel generations. Different from explicit attribute extraction (Vilnis et al., 2022; Yang et al., 2022), our generative method is able to extract implicit knowledge behind item titles or descriptions. For example, the title “Diesel Analog Three-Hand - Black and Gold Women’s watch” contains specific attributes like “Black and Gold” or type information “women’s watch.” Such knowledge can be easily extracted by off-the-shelf tools. Traditional information extraction based approaches mostly cover our knowledge if the generation simply copies titles to reflect the attributes. Otherwise, it means that we provide much novel and diverse information compared with traditional approaches. The novelty ratio increases from 96.85% to 97.38% after we use the trained classifiers for filtering. Intuitively, filtering can improve the novelty ratio. For the assertions whose typicality scores are above 0.9, we also observe that the novelty ratio reaches 98.01%. These findings suggest that FolkScope is indeed an effective framework for mining high-quality implicit knowledge.

Subcategory	Generation
(Costumes, Toys)	he wants to disguise himself as a superhero they can be used to make a crown costume he wanted to be a star war character for Halloween they are both a manner of Christmas decoration he wants kids to have fun and enjoy the Easter holiday he is able to dress up as a pirate
(Dresses, Dresses)	they are symbol of the fashion trend they can both be worn to formal events they can both be worn for casual occasions they are both used for wedding dress they are both capable of giving a good fit they can both being worn by girls of any age

Table 13: Generated knowledge in same subcategory.

to use human annotation with middle-size LLMs.

D.2 Fine-grained Subcategory Knowledge

Since the items are organized in multilevel fine-grained subcategories in the catalog of shopping websites, we are interested in whether our constructed KG contains high-quality common intentions among items belonging to subcategories. The common knowledge can be useful to have intention-level organizations besides category-level and further help downstream tasks. The co-buy item-pairs in our sampled *clothing* category fall into 15,708 subcategory pairs, such as (*necklaces, earring*) or (*sweater, home & kitchen*), where most of them are different subcategories in one pair. We select frequent common assertions with high typicality scores to demonstrate the abstract knowledge. Two examples are shown in Table 13. Though costumes and toys belong to two different types, they are complementary because of the same usage, such as “Halloween,” “Easter holiday,” and “Christmas,” or sharing the same key feature like “star war character,” “pirate.” On the other hand, if two items fall in the same subcategory, like “dresses” in Table 13, the generated assertions share some common characteristics, such as being suitable for certain events and complementing each other when worn together.

D.3 Use Different LLMs as Knowledge Source

We are interested in whether different sizes of language models have large impact on the generation. Hence we empirically analyze the plausible rate of generation using four language models: GPT-J (6b), OPT-30b, OPT-66b and text-davinci-003. We can observe that: 1) OPT-30b outperformed GPT-J over 10% (51% vs. 41%) while OPT-66b did not improve OPT-30b. 2) text-davinci-003 achieved nearly perfect results and make little mistakes when recognizing products given title information. Though impressive results, we have to balance between knowledge size and cost hence the takeaway from our work is

Relation	Clothing	Electronics
Open	a fan of Harry Potter / Star Wars give gifts for his girlfriends / his son go to a costume party / wedding / be a father	make a robot, make a remote control, build a PC know how to play guitar / take better photos learn code / microcontroller programming
UsedFor	outdoor activities, hiking, camping, travel daily use, formal occasions, winter sports babies, maternity wear, sleepwear Halloween costumes, Christmas cosplay jewelry making, leather care, weight loss nursing, working out, polishing shoes	outdoor use, navigation, education, networking personal use, office work, home theater, 3D movies baby photography, underwater photography Arduino projects, Raspberry Pi, Samsung headphone water cooling, cable management, screen protection framing, storing data, mounting camera, prototyping
CapableOf	keeping cool, keeping dry, keeping warm being worn with jeans / dress / shorts holding up pants, holding a lot of stuff protecting from rain / sun / harmful UV rays making him look like wizard / price / Batman	taking pictures, printing labels, boosting signals being used in car / boat / computer / water / emergency holding radio / CDs / GoPro camera / phones / devices tracking location / heart beat rate / cycling activities controlling light / TV / home automation / device
SymbolOf	his love for daughter / wife / mother / family luxury, friendship, childhood, the 80s modern life, American culture, graduation	his passion for gaming / aviation / cycling / sports security, reliability, durability, high performance latest technology, hacker culture, music industry

Table 14: More examples of high-frequency typical assertions for different relations. Note we omit the prompts for space and simplicity .

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Limitation section.
- A2. Did you discuss any potential risks of your work?
Ethics section
- A3. Do the abstract and introduction summarize the paper’s main claims?
Introduction section
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Sec. 2 and Sec 4

- B1. Did you cite the creators of artifacts you used?
Sec. 2, Sec. 4 and Sec. 5
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
open access
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Sec.2 and Sec. 3
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics section
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sec. 2, Sec. 3 and Sec. 4

C Did you run computational experiments?

Sec.2, Sec.3 and Sec4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sec. 2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sec. 2, Sec. 3 and Sec. 4
 - C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sec 4 (multiple run averaged)
 - C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sec2 and Sec 4.
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Sec. 2 and Sec.3
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A
 - D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A
 - D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Sec 2 and Sec 3
 - D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.
 - D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A