A Usage-centric Take on Intent Understanding in E-Commerce

Anonymous ACL submission

Abstract

Identifying and understanding user intents is a pivotal task for E-Commerce. Despite its popularity, intent understanding has not been consistently defined or accurately benchmarked. In this paper, we focus on predicative user intents as "how a customer uses a product", and pose intent understanding as a natural language reasoning task, independent of product ontologies.

We identify two weaknesses of FolkScope, the SOTA E-Commerce Intent Knowledge Graph, that limit its capacity to reason about user intents and to recommend diverse useful products. Following these observations, we introduce a Product Recovery Benchmark including a novel evaluation framework and an example dataset. We further justify the above FolkScope weaknesses on this benchmark. [1]

1 Introduction

User intents are a crucial source of information for E-Commerce (Zhang et al., 2016; Hao et al., 2022). Intents reveal users' motivation in E-Commerce interactions: suppose a user plans to go for outdoor barbecue, their intent may not refer only to barbeque smoker grills but also to other items that can be useful, such as disposable cutlery or plates. In these cases, traditional product recommendation approaches would fail to handle these queries or to remind customers of the products they may need but have forgotten. *Intent Understanding* offers great benefits in recommending distinct products based on common user intents they fulfil. It involves identifying user intents and connecting them with products: a profile of user intents is extracted using user interactions (e.g. co-buy records, reviews) for each product listing. Then, a mapping from intents to product listings can be built to predict useful products based on user intents.

One significant challenge towards effective intent understanding is the vague definition of user intents, which precludes effective intent identification and can easily result in contaminated intent-product associations. In prior work (Yu et al., 2023; Luo et al., 2021), user intents are often blended with "product properties" or "similar products", which we argue are related to the products and not the users. These shortcuts may benefit existing product recommendation benchmarks, but are not aligned with the intent understanding objective, namely, to retrieve superficially distinct kinds of products serving common intents.

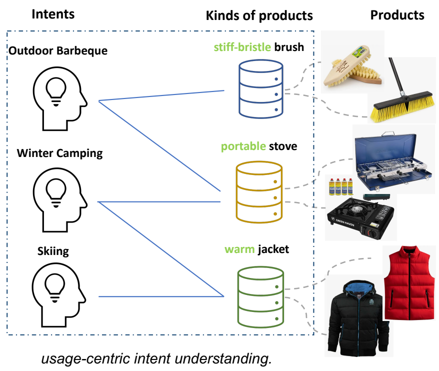

Figure 1: A graphic illustration of the usage-centric paradigm of intent understanding.

Therefore, we propose a usage-centric paradigm for intent understanding. In this paradigm, user intents are focused on natural language predicative phrases, i.e. how users **use** a product; also, instead of individual product listings, we aim to predict *kinds of products* **useful** for an intent. In particular, we define user intents as activities to accomplish (e.g. outdoor barbecue) or situations to resolve (e.g. lower-back pain); and, kinds of products as clusters of product listings of the same category (e.g. scrub brush) and property (e.g. stiff bristle). Conditioning on the kinds of products offers guarantees that the list of relevant predictions is not endless. Our task is a natural language reasoning task, closely related to commonsense reasoning (Sap et al., 2019; Bosselut et al., 2019): "The user has intent I" entails "The kind of product P is useful for the user."

[1] We will release our code and datasets.

We present an analysis of a SOTA E-Commerce intent KG, FolkScope (Yu et al., 2023), which reported promising results on an intrinsic co-buy prediction task. Refactoring their KG to build associations between kinds of products and their usage intents, we find two unsatisfactory characteristics in their KG topology: 1) *property-ambiguity*: generated user intents are poorly aligned with relevant product properties, such that the KG often maps user intents to kinds of products with relevant category but fairly random properties; 2) *category-rigidity*: each intent is strongly associated with a single category of product, such that the KG is unable to recommend diverse products that serve common intents.

In light of these findings, we develop a Product Recovery Benchmark, including an evaluation framework that aligns with the usage-centric paradigm, isolating product-specific confounders, such as product price or ratings. Also, we provide a dataset based on the Amazon Reviews Dataset (ARD) (Ni et al., 2019) where we further validate the impact of the weaknesses in FolkScope. All intent understanding methods developed on the ARD can be evaluated using this benchmark.

To summarize, in this paper: 1) we propose a usage-centric paradigm for intent understanding; 2) we introduce a product recovery benchmark featuring a novel evaluation framework, and report results with SOTA baselines; 3) we identify crucial weaknesses in existing SOTA as *category-rigidity* and *property-ambiguity*, where we propose intent mining from user reviews as a promising future direction.

2 Usage-Centric Intent Understanding

We propose a usage-centric paradigm of intent understanding, focusing on usage user intents and the kinds of useful products, where the goal is to ground usage user intents in kinds of useful products. Differently from the "informal queries" in Luo et al. (2021), and similarly to Ding et al. (2015), our usage user intents are generic eventualities/situations, independent of product ontologies.

We introduce *kinds of products* as the target granularity level, as it abstracts away the nuanced differences among individual listings, and yields a purely natural language setup, independent of product ontologies. It contains just enough information (category + property) to represent the product listings inside for intent understanding.

User intents rarely require combinations of properties in a product category. Therefore, to avoid generating factorial numbers of kinds of product, we impose a mild constraint that only one property is specified for each kind of product.

We demonstrate the specificity trade-off with an example below: for outdoor barbecues, a stiff-bristle scrub brush is useful for cleaning the grease on the grill. To that end, there are many listings of hard-bristle scrubs but the exact choice among them is irrelevant to the user intent and could be identified by downstream recommendation systems using other factors (customer habit, geo-location, etc.). However, the stiff bristle property is essential for a listing to be suitable for outdoor barbecues. In short, grouping based on *kinds of products* strikes a balance between sparsity that comes with specificity, and ambiguity that comes with generality.

3 FolkScope Analysis

3.1 KG Refactoring

We refactor FolkScope to our usage-centric intent understanding paradigm. FolkScope KG connects products with their user intents, which are generated with OPT-30B (Zhang et al., 2022) when given pairs of co-bought products sourced from ARD (Ni et al., 2019), along with commonsense relations.

Among their 18 commonsense relations, we filter out all "item" relations as well as 3 "function" relations (*SymbolOf*, *MannerOf*, and *DefinedAs*), since they are nominal in nature, and are irrelevant to product usage. We keep the remaining 5 predicative relations, *UsedFor*, *CapableOf*, *Result*, *Cause*, *CauseDesire*, as legitimate user intents.

To group the product listings into kinds of products, we take the fine-grained product categories from ARD (e.g. *Kids' Backpacks*), and borrow the attributes under the relation *PropertyOf* in the original FolkScope KG as properties.[2]

We compute the association strengths from selected user intents to common kinds of products by aggregation. Let $e(I_i, P_j)$ be the connection of intent I_i with product listing P_j, P_j belongs to a kind of products K_k. The association strength for edges in the refactored KG are then computed as: $e'(I_i, K_k) = \sum_{P_{j'} \in K_k} pmi(P_j, K_k) * e(I_i, P_j)$. [3]

[2]These attributes do not fit the criteria for usage user intents, but they are acquired through generic LLM prompted summarization, and thus are borrowed as product properties.

[3]The pmi term penalizes product listings with multiple kinds of products (e.g. multiple properties in one listing).

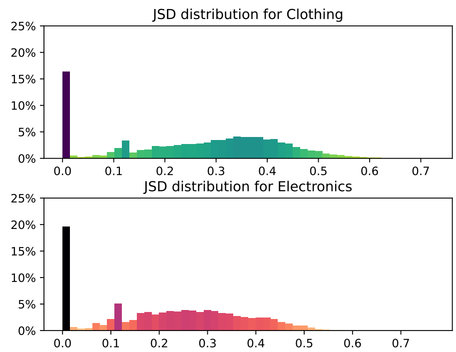

Figure 2: Histograms of Jensen-Shannon Divergence for each intent-category pair. Values are packed around 0: property-distributions of edge weights conditioned on intents are close to unconditioned frequency priors.

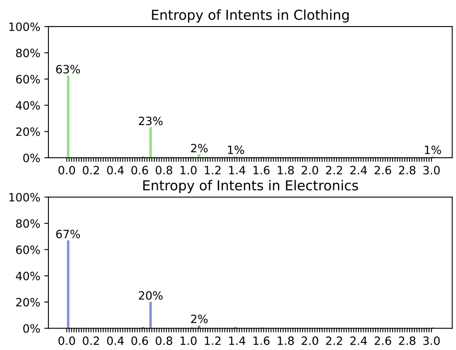

Figure 3: Histograms of category-entropy for each user intent. Values are concentrated at 0.0 and 0.7, meaning the intent is associated with only 1 / 2 categories.

3.2 Statistical Analysis

We identify two major weaknesses of FolkScope KG under the usage-centric paradigm: it is over-specific about categories of useful products, but under-specific about the required properties in these categories. Intents in FolkScope tend to be associated with products with vague properties from few categories, rather than specific kinds of products from diverse categories. (examples in App. B).

Property-Ambiguity For each user intent, we look into the distribution of its edge weights among kinds of products from one category with different properties. We compare these posterior edge-weight distributions, conditioned on the intent, against the prior distributions among the differently-propertied kinds of products in that category. We calculate Jensen-Shannon Divergence (JSD) between these conditional and prior distributions (see Figure 2): for up to 20% of cases, JSD is < 0.1, where only 2% of cases have JSD > 0.5.

This shows, the KG's edge weights among same-category different-property kinds of products are strongly predicted by their prior distribution, and are insensitive to the specific usages depicted by user intents. We credit this to the mismatch between property and intent mining: each product listing may have multiple properties and may serve multiple intents, but the mappings between these properties and intents are underspecified.

Category-Rigidity For each user intent, we measure the category-diversity of its edge weights in the refactored KG: we compute the entropy of its edge weights grouped by product categories.

Figure 3 shows the entropy meta-distributions: entropy values are concentrated in 2 narrow ranges, $[0, 0.02)$ and $[0.68, 0.70)$. We notice that an entropy in $[0, 0.02)$ indicates that the associations about this intent are focused on only one product category; $[0.68, 0.70)$ indicates that the associations are focused on two product categories. Therefore, from Figure 3 we can conclude that over 80% of the intents are associated with only one or two categories. This category-rigidity in FolkScope hampers its ability to recommend diverse kinds of products, as we will discuss in §4.2.

4 The Product Recovery Benchmark

4.1 Benchmark Design

Following our intent understanding paradigm in §2, we introduce a usage-centric evaluation framework, which aims to recover kinds of products based on retrieved user intents. Under this framework, an intent understanding method first predicts a profile of user intents for a product listing (using product description, user reviews, etc.). Then, using **solely** the predicted intent as input, the method recovers useful kinds of products based on its knowledge of E-Commerce demands (e.g. in symbolic KGs or LLMs). The predictions are compared against: 1) *bought-product-recovery*: kinds of product to which the current product belongs; 2) *co-bought-product-recovery*: kinds of co-bought products that belong to other categories.

We take *bought-product-recovery* as our main evaluation setup, since it focuses on intent-to-kinds-of-product associations. Compared to the product recommendation evaluation in Yu et al. (2023), this framework marginalizes factors inciting co-buy behaviour (e.g. brand loyalty, geolocation, etc.). We also include the co-bought-product-recovery setup to validate statistical findings on cross-category recommendation performance.

We instantiate the proposed evaluation framework with a product recovery benchmark, based on

Models	Clothing	Electronics
FolkScope	0.192	0.263
FolkScope − properties	0.116	0.166
FolkScope + GPT	0.187	0.257

Table 1: MRR_{max} for *bought-product-recovery* task.

the ARD (Ni et al., 2019), using available resources. We utilise the pool of product listings in ARD, enriched with product descriptions, category information, anonymized user purchase records and reviews. We additionally borrow kinds of products from refactored FolkScope, as in §3.1.[4]

Evaluation metric Following prior work (Chen and Wang, 2013), we measure success by Mean Reciprocal Rank (MRR) of gold kinds of products in the predicted distributions. In case multiple gold kinds of products are assigned for a product listing, we calculate the MRR_{max} (see App. Eq. 2) using the highest-ranking hit.

4.2 Experiments and Results

We evaluate the FolkScope KG (refactored in §3.1) with the Product Recovery benchmark. We offer the baseline results in Table 1, and highlight below the impact of weaknesses discussed in §3.2.

Property-Ambiguity To understand how property ambiguity affects FolkScope performance, we compare it with another prior property baseline derived from it: for each evaluation entry, we corrupt the FolkScope predictions by replacing the property in the predicted kinds of products based on the property popularity. (see Appendix A.2 for details)

From Table 1, we observe that *FolkScope − properties* reached respectable performance with only moderate regression from FolkScope predictions. This limited MRR gap shows the impact of property-ambiguity, where performance gains could be expected with better property alignment.

Category-Rigidity To validate the category-rigidity observation in §3.2, we also evaluate the FolkScope KG in the co-bought-product-recovery setup, where we specifically use it to predict kinds of co-bought products in **other categories**.

In this setup, we observe low MRR_{max} of 0.077 and 0.033 for *Clothing* and *Electronics* domains, respectively: the FolkScope KG cannot effectively recommend superficially distinct kinds of products connected with the same user intents.

Notably, between the two domains, FolkScope reaches a slightly higher MRR_{max} in *Clothing*. This is consistent with our findings in Figure 3, where category-entropy values are slightly more spread than in *Electronics*.

LLM Rerank We also evaluate LLM performance in our benchmark, using GPT-3.5-turbo (Brown et al., 2020). Ideally, we would like the LLM to predict useful kinds of products end-to-end. However, due to the difficulty of reliably matching LLM predictions with gold kinds of products[5], we instead adopt a re-ranking paradigm, where we prompt the LLM to re-rank the top-10 kinds of products predicted by FolkScope (see App. A.4).

As Table 1 shows, we observe no clear benefit with LLM-reranking. We investigate this failure by looking into where hits are met in the predictions and find that most hits are either at first or not in top 10 (see App. B.3). These polarized distributions leave little room for re-ranking to take effect.

We raise the warning that dataset artefacts from the common source corpus (AWD) could be behind this abnormally high hit-at-1 rate (compared with the MRR_{max} value), where the reported MRR_{max} values may have been inflated. Due to the lack of another large E-Commerce Reviews corpus, we leave further investigations for future work.

5 Discussions and Conclusion

In this paper, we revisit intent understanding from a usage-centric perspective, as a natural language reasoning task, to detect superficially distinct kinds of products useful for common usage intents. We developed a Product Recovery benchmark, and investigated two weaknesses of the SOTA FolkScope KG in supporting usage-centric intent understanding: *Property Ambiguity* and *Category-Rigidity*.

We advocate for adopting the usage-centric intent understanding paradigm, and for considering user reviews, in addition to co-buy records. Desired product properties and their respective intents are likely to co-occur in product reviews, relieving property-ambiguity; the same usage intents tend to be described consistently in user reviews across different categories, relieving category-rigidity.

[4]Our elicitation procedure is corpus-agnostic, we empirically select ARD as it is the largest available dataset; we acknowledge that re-using information from FolkScope may grant it an unfair advantage, however, we show below, that it nevertheless suffers from the aforementioned weaknesses and fails to perform intent understanding effectively.

[5]In Appendix C, we include an LLM-only baseline using GPT-4 as matching metric, where we find it underperforming FolkScope baseline, and find GPT-4 metric over permissive.

Limitations

In this paper, we have proposed to study E-Commerce intent understanding from a usage-centric perspective. Due to the lack of consistent task definition, we are only able to analyse one SOTA intent understanding KG (namely FolkScope) and one SOTA LLM. We encourage more research attention on the usage-centric E-commerce intent understanding task for a more diverse landscape.

We have established that weaknesses of Property Ambiguity and Category Rigidity exist in the SOTA KG, and we have offered a principled hypothesis that utilizing genuine user reviews could help with these weaknesses. However, due to limits to the scope of this paper, we do not provide empirical evidence for this hypothesis and leave it as a promising direction of future work.

We note that as this paper is related to recommendation, there exists risks that methods developed on the Product Recovery Benchmark may be used to bias customer decisions; on the other hand, we also note that our task definition is purely natural language and does not involve any individual product listings, therefore it would not bias customer choices among directly competing listings of the same kinds of products.

References

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 Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. 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. ArXiv:2005.14165 [cs].

Li Chen and Feng Wang. 2013. Preference-based clustering reviews for augmenting e-commerce recommendation. *Knowledge-Based Systems*, 50:44–59.

Xiao Ding, Ting Liu, Junwen Duan, and Jian-Yun Nie. 2015. Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). Number: 1.

Zhenyun Hao, Jianing Hao, Zhaohui Peng, Senzhang Wang, Philip S. Yu, Xue Wang, and Jian Wang. 2022. Dy-hien: Dynamic evolution based deep hierarchical intention network for membership prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 363–371, New York, NY, USA. Association for Computing Machinery.

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 *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3385–3393, Virtual Event Singapore. ACM.

Jianmo Ni, Jiacheng Li, and Julian 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)*, pages 188–197, Hong Kong, China. 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3027–3035.

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. ArXiv:2211.08316 [cs].

Chenwei Zhang, Wei Fan, Nan Du, and Philip S. Yu. 2016. Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 1373–1384, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona 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. 2022. Opt: Open pre-trained transformer language models.

A Implementation Details

A.1 Benchmark data split

We follow Yu et al. (2023), and we split product instance in FolkScope KG into training, validation and test splits with respective portions of 80%, 10% and 10%. Please refer to Table 2 for detailed statistics. Note that *Clothing* stands for the "Clothing, Shoes and Jewelry" domain in the Amazon Reviews Dataset, and *Electronics* simply stands for the "Electronics" domain in the Amazon Reviews Dataset.

Categories	**Train**	**Validation**	**Test**
Clothing	30296	2027	2088
Electronics	85086	7853	7900

Table 2: Number of product listings in the training, validation and test set. Please note that we drop product listings that lack related kinds of products, so the ratio of the number of instances across the splits are not exactly equal to 8:1:1.

A.2 Prior Property Baseline

For each kind of product in the prediction list, we corrupt its property part with respect to its prior popularity within its fine-grain category in the Amazon Reviews Dataset. Popularity is defined as the frequency of a property appearing with the product listing having this corresponding fine-grained category. To avoid repeated kinds of products in the predictions, when multiple predicted kinds of products from the same category are predicted, we draw properties top-down w.r.t. popularity for each prediction.

A.3 Evaluation Metric

Our evaluation metric $\mathrm{MRR}_{\max}$ can be formally defined as follows:

$$\mathrm{RR}_{\max}(l) = \max_{c \in C_{\text{gold}}(l)} \left(\mathrm{rank}(c)^{-1}\right) \tag{1}$$

$$\mathrm{MRR}_{\max} = \frac{\sum_{l \in L} \mathrm{RR}_{\max}(l)}{|L|} \tag{2}$$

where RR represents the Reciprocal Rank, $C_{\text{gold}}(l)$ are the gold clusters for the listing l and L is the set of all listings in the benchmark.

A.4 GPT-3.5-turbo Re-ranking

For each product listing l, when there is no predicted kind of products given a set of related user intents, we mark the $\mathrm{RR}_{\max}(l)$ as 0 both before and after re-ranking.

A.4.1 Re-ranking Prompt

> A product is suitable for the following purposes:
> {*Intents*}
>
> Please rank the following categories in order of likelihood that the product belongs to them (most likely to least likely):
> {*kinds of products list*} ...
> Answer:
> 1.

We fill *Intents* with a set of mined user intents and *kinds of products list* with the top 10 predictions for kinds of products.

Note that in this setting and in § C.1.1, we still use the term "category" in LLM prompts to refer to kinds of products, because during preliminary experiments we found that LLMs do not respond well to the term "kind of product".

B Details on the Statistical Analyses

B.1 Property Ambiguity

We hypothesize that if the edges from a user intent to kinds of products are specific about the properties, then, the distribution of property edge weights should typically be distinct from the prior frequencies of these properties.

For example, for the user intent outdoor barbecues, its edge weights distribution among different kinds of scrub brush products should be skewed, where stiff bristle scrub brush receives much higher weights than other kinds of scrub brushes. On the other hand, out of the context of outdoor barbecues, there are more diverse kinds of scrub brushes on the market, e.g. soft bristle, wooden handle, etc.

In §3.2, we compare the per-category edge weight distributions against the corresponding frequency priors, and find that for a large portion of edges, the edge weight distributions, conditioned on the user intent, are almost identical to the prior distributions. This indicates low sensitivity of the FolkScope edges to the relevant properties.

B.2 Category Rigidity

In a E-Commerce user intent KG, a non-negligible amount of usage user intents should entail the demand for diverse products from different categories.

For the example of outdoor barbecues, for outdoor barbecues one may need not only scrub brush, but also other categories of products, such as picnic blankets, grill gloves, etc.

Therefore, we take the category-entropy of the edges for each user intent, to measure how diverse the KG edges are w.r.t. categories. We add up the edge weights grouped by product categories (e.g. edge weights to stiff bristle scrub brush and scrub brush with wooden handle are added together), and compute the entropy of the converted category distribution. As discussed in §3.2, we found severe category rigidity in the FolkScope KG, where very few user intents have diverse category distributions, the majority of user intents are associated with only one category, followed by those associated with two.

B.3 Hit distribution in predictions

	Clothing	Electronics
hit at 1st	16%	22%
hit out top 10	73%	63%

Table 3: The ratio of hit being the first in the prediction list and not in the top-10 of the prediction list.

C GPT End-to-End Evaluation

We perform an additional experiment to directly predict kinds of products in an end-to-end setup, with an LLM, for our proposed product recovery task. Again, we use GPT-3.5-turbo as the LLM and design the zero-shot prompt as in §C.1.1. However, due to the absence of the complete ontology of the Amazon Reviews Dataset, it is challenging for GPT-3.5-turbo to predict the exact ground truth kinds of products. To sidestep the difficulty of evaluating whether the predicted strings are semantically identical to the ground truth labels, we use GPT-4 to judge whether there is a match between predicted and ground truth labels. The relevant prompt is specified in §C.1.2. The detailed evaluation results is presented in Table 4.

From Table 4, we can observe that GPT-3.5-turbo does not outperform the FolkScope KG baseline on the product recovery benchmark. Compared to the strict string matching results in Table 1,

	Clothing	Electronics
GPT-3.5-turbo	0.511	0.543
FolkScope	0.527	0.671

Table 4: $\mathrm{MRR}_{\mathrm{max}}$ score when evaluating using GPT-4 as the judge for matching. Values for GPT-3.5-turbo and our baseline refactored FolkScope KG are both higher in absolute values due to the more benign matching criterion; the LLM baseline with GPT-3.5-turbo does not outperform the KG baseline.

GPT-4 evaluation has a significantly more permissive criterion on matching, yielding much higher $\mathrm{MRR}_{\mathrm{max}}$ values. We find many of these "matched" verdicts by GPT-4 to be spurious (see Table 5), and conclude that GPT-4 cannot easily achieve reliable matching for the product recovery benchmark, and more robust criteria are needed before replacing the exact match criterion.

C.1 Prompt Examples

C.1.1 Kinds of Products Prediction

Intents:
{*intents*}
Given the intents, please predict the top 10 kinds of products that will be useful for these intents.
A kind of product is the concatenation of a fine-grained category from the Amazon Review Dataset and a useful property. For example: Clothing, Shoes & Jewelry|Men|Watches|Wrist Watches ### leather.
Kinds of products:
1.

C.1.2 Prediction Evaluation

Here is a list of predicted categories:
{*prediction*}
Validate each prediction based on the ground truth categories[T/F].
Each prediction can be considered true when it is similar to one of the ground truth categories.
Ground truth categories:
{*ground truth*}

Ground truth kinds of products
1. Clothing, Shoes & Jewelry|Costumes & Accessories|Men|Accessories ### Wandering Gunman 2. Clothing, Shoes & Jewelry|Costumes & Accessories|Men|Accessories ### Holster 3. Clothing, Shoes & Jewelry|Costumes & Accessories|Men|Accessories ### Western
GPT-3.5-turbo prediction
1. Clothing, Shoes & Jewelry|Men|Costumes|Western ### authentic ...
Ground truth kinds of products
1. Clothing, Shoes & Jewelry|Women|Jewelry|Earrings|Stud ### Jewelry 2. Clothing, Shoes & Jewelry|Women|Jewelry|Earrings|Stud ### Gemstone 3. Clothing, Shoes & Jewelry|Women|Jewelry|Earrings|Stud ### Sterling Silver
GPT-3.5-turbo prediction
1. Clothing, Shoes & Jewelry|Women|Earrings|Stud Earrings ### elegant and beautiful ...

Table 5: Here we list two examples that GPT-4 validate with $\mathrm{RR}_{\max} = 1$. In the first example, it validates the first prediction as true by matching the "property" part of the ground truth 3 with the main category of prediction 1. In the second example, the "property" part of prediction 1 is too general compared to all the ground truth kinds of products, but it still validates it as true.

D Computational Budget

D.1 Main Experiments

All the benchmark construction and evaluation has been performed using 2 x Intel(R) Xeon(R) Gold 6254 CPUs @ 3.10GHz.

FolkScope KG Refactoring We converted all the intents generated by FolkScope without applying any of its proposed filters based on the graph evaluation results on the validation set. The whole graph generation for both domains takes around 24 hours in total.

FolkScope Intents Evaluation We need around 71 and 6 hours for evaluating the intents for the test set of the Clothing and Electronics domain respectively.

D.2 LLM Experiments

We mainly use GPT-3.5-turbo and GPT-4 for our LLM-related experiments. Please refer to Table 6 for details about the relevant costs. For both models, we keep the default parameters from OpenAI, and set the temperature to 0 to facilitate reproducability.

Experiment	Clothing	Electronics
LLM Rerank	3.86 $	1.38 $
LLM End-to-End	15.57 $	14.56$

Table 6: API costs of our LLM-related experiments. For the LLM Rerank experiment, we re-rank all the data samples in the test set while for the End-to-End evaluation, we only sample 1000 data samples in the test set.

E Artifact Licenses

Amazon Reviews Dataset: Limited license for academic research purposes and for non-commercial use (subject to Amazon.com Conditions of Use)
FolkScope: MIT license