Knowledge-Augmented Relation Learning for Complementary Recommendation with Large Language Models

Chihiro Yamasaki
The University of
Electro-Communications
Chofu, Tokyo, Japan
c.yamasaki@librict.jp

Kai Sugahara
The University of
Electro-Communications
Chofu, Tokyo, Japan
research@kais.jp

Kazushi Okamoto
The University of
Electro-Communications
Chofu, Tokyo, Japan
kazushi@uec.ac.jp

Abstract

Complementary recommendations play a crucial role in e-commerce by enhancing user experience through suggestions of compatible items. Accurate classification of complementary item relationships requires reliable labels, but their creation presents a dilemma. Behavior-based labels are widely used because they can be easily generated from interaction logs; however, they often contain significant noise and lack reliability. While function-based labels (FBLs) provide high-quality definitions of complementary relationships by carefully articulating them based on item functions, their reliance on costly manual annotation severely limits a model's ability to generalize to diverse items. To resolve this trade-off, we propose Knowledge-Augmented Relation Learning (KARL), a framework that strategically fuses active learning with large language models (LLMs). KARL efficiently expands a high-quality FBL dataset at a low cost by selectively sampling data points that the classifier finds the most difficult and uses the label extension of the LLM. Our experiments showed that in out-of-distribution (OOD) settings, an unexplored item feature space, KARL improved the baseline accuracy by up to 37%. In contrast, in in-distribution (ID) settings, the learned item feature space, the improvement was less than 0.5%, with prolonged learning could degrade accuracy. These contrasting results are due to the data diversity driven by KARL's knowledge expansion, suggesting the need for a dynamic sampling strategy that adjusts diversity based on the prediction context (ID or OOD).

CCS Concepts

• **Information systems** → **Recommender systems**; *Business intelligence*; • **Applied computing** → *Online shopping*.

Keywords

complementary recommendation, active learning, large language models, function-based label, label augmentation

1 Introduction

The evolution of e-commerce has driven a paradigm shift in recommender systems, moving beyond individual item suggestions toward optimizing user experience through strategic item combinations [12, 31]. Complementary recommendation, a key technique in this domain, aims to identify functionally compatible item pairs to enhance user satisfaction and boost sales [12, 13], an impact proven by results such as a 9.56% increase in the Visit Buy Rate on the online supermarket platform Meituan Maicai [5] and a 0.23% boost in product sales on Amazon [9].

One of the most fundamental challenges in providing effective complementary recommendations is defining and creating labels that accurately describe the relationships between items [24]. Early studies in this field relied primarily on behavior-based labels (BBLs) [9, 16], which are derived from co-occurrence patterns in user interaction data, such as purchase or viewing histories. Although BBLs can be easily constructed with sufficient user behavioral data and have been widely used to train and evaluate complementary recommendation models, they often suffer from a lack of interpretability and potential noise owing to the inconsistencies or biases inherent in user behaviors [13, 20, 24]. Consequently, Sugahara et al. proposed function-based labels (FBLs) [24, 32]. Rather than merely indicating the presence or absence of a complementary relationship, FBLs classifies item pairs into nine functional categories. These labels are annotated by domain experts, and are completely independent of user interaction logs, making them reliable and noise-resistant alternatives to BBLs.

A practical drawback of FBLs is that they are not data-driven, resulting in substantial annotation costs. On e-commerce platforms with a large number of items, annotating all possible item combinations using FBLs is practically impossible. A more feasible approach would be to allocate human resources to annotate a subset of items concentrated in specific item categories on an e-commerce site. Recommendation models trained on such limited datasets perform well within the in-distribution (ID) feature space [32]; however, their generalization performance for out-of-distribution (OOD) items is often disappointing. From a different perspective, recent advancements in annotation techniques have demonstrated that large language models (LLMs) can function as zero-shot classifiers [13, 32], providing an alternative annotator for FBLs. Nevertheless, the inference cost of LLMs remains impractical for large-scale applications, and LLMs have inherent limitations in covering the entire range of items.

In practice, only a small fraction of item pairs exhibit complementary or substitutable relationships, making it potentially sufficient to efficiently collect informative samples for training classification models. In other words, blindly annotating all possible item
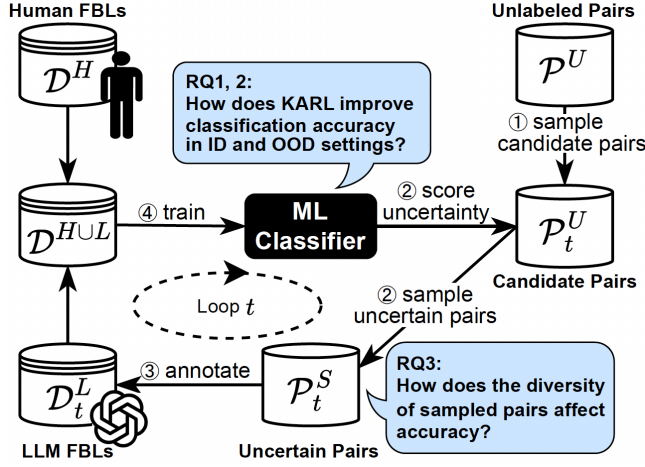
Figure 1: General overview of our study.

pairs with FBLs is inefficient for learning accurate decision boundaries. Actively selecting informative item pairs that help the model learn such boundaries, particularly for recognizing complementary relationships, can significantly reduce the need for exhaustive annotations. Given that the annotation cost, whether by human annotators or large language models (LLMs), depends on the selected item pairs, this active learning approach [22, 25] enables the construction of effective complementary recommendation models at a minimal scale while still covering diverse item categories.

Drawing from these perspectives, we propose KARL (Knowledge-Augmented Relation Learning), a framework designed to accurately and effectively classify complementary relationships under constrained annotation resources and limited training datasets regarding FBLs. KARL leverages active learning to iteratively (1) sample uncertain pairs from unlabeled item pairs for where the machine learning (ML) classifier struggles to assign FBLs, (2) assign FBLs to these pairs using LLMs, and (3) retrain the classifier on augmented data, as illustrated in Figure 1. This iterative process progressively improves the generalization capability of the model by gradually incorporating LLM-generated knowledge, addressing the risk that models trained on small, domain-limited datasets may only perform well in ID settings. To evaluate its effectiveness, we validated the accuracy of the classifier retrained iteratively through KARL on two FBL datasets corresponding to the ID and OOD scenarios with real-world data from a large e-commerce platform. Then, we analyzed the effect of the diversity of the training data induced by active learning on the classification accuracy in each scenario. This study is guided by the following research questions:

RQ1: To what extent does KARL enhance the generalization accuracy in ID feature spaces compared with the baseline?

RQ2: To what extent does KARL enhance the generalization accuracy in OOD feature spaces compared with the baseline?

RQ3: How does training data diversity driven by uncertainty sampling affect classification accuracy in ID and OOD settings?

2 Related Work

Complementary Recommendation. With the growth of e-commerce, complementary recommendations have become the key to enhancing user experience and increasing sales [7]. Previous studies in this field primarily utilized BBLs [9, 16], derived from user interaction logs such as purchasing or viewing histories. Although BBLs can be constructed at a low cost on a large scale and have been widely used for training and evaluating complementary recommender systems [2, 14, 15, 20, 28, 34], their reliability is often compromised by noise labels stemming from irregular user behaviors [13, 15, 20, 24, 31]. To address the challenges of BBLs, FBLs [24, 32] have been proposed to define item relationships based on the following nine functional categories:

(A) Items x and y have the same function and usage.

(B-1) Item x can be replenished with item y.

(B-2) Item y can be replenished with item x.

(C-1) Items x and y must be combined to be usable.

(C-2) When combined with item y, item x becomes more useful.

(C-3) When combined with item x, item y becomes more useful.

(C-4) Combining x and y makes them more useful.

(D) Items x and y have no relationship.

(E) Items x and y seem to have a relationship, but it is difficult to verbalize.

FBLs are highly reliable because they are defined by domain experts based on functional relationships and are independent of noisy user behaviors, leading to high classification accuracy when used for training [32]. However, this reliance on manual expert annotation, although integral to their quality, makes expanding FBL datasets prohibitively expensive.

To address the high cost of FBLs, a recent work demonstrated that LLMs render effective annotators for FBL creation [32]. Specifically, GPT-4o-mini [18] achieved a macro-F1 score of 0.849 with human ground truth in 3-class classification (complementary, substitute, and unrelated). Meanwhile, similar studies applying LLMs to relationship classification still face unresolved challenges. Although a study using an LLM directly as a classifier reported high accuracy and explainability for the identified relationships [13], this approach faces scalability issues when applied to numerous item combinations in e-commerce.

LLM-based Active Learning. Active learning is a technique for efficiently training models while minimizing annotation costs [22]. This works by having the model actively select samples that are difficult to predict, which are then labeled by a domain expert and added to the training set. [17, 25]. Common sampling strategies include uncertainty sampling, which selects samples near the decision of the model boundary [10, 17, 22, 25], and diversity sampling, which prevents data bias [4, 25, 29]. Traditionally, active learning has relied on human annotators; this is a highly reliable approach, but remains costly [11, 25].

In response, the remarkable inference capabilities of LLMs have shown promise as cheaper alternatives to human annotators [11, 27, 29, 30]. For instance, using GPT-4-Turbo for low-resource language annotation has been reported to significantly reduce costs compared with human annotation [11]. In addition, the benefits

may extend beyond cost reduction, as some studies have reported a synergistic effect in which a model trained on LLM-generated labels surpasses the performance of the LLM itself [6, 33]. However, the promising performance of LLMs is not universal. A comprehensive experiment confirmed that their effectiveness is highly dependent on the specific task and data characteristics [19]. This highlights the critical need to validate the applicability and limitations of the LLM-based annotations in specific domains.

3 Methodology

We propose KARL, an active learning framework designed to achieve an accurate and effective classification of item relationships by efficiently leveraging limited FBL datasets and annotation resources. The framework utilizes the following two data sources: $\mathcal{P}^U$, which consists of all possible unlabeled item pairs, and $\mathcal{D}^H$, a dataset of human-annotated FBLs used to train the classifier. The framework initially trains a classifier on $\mathcal{D}^H$ for 3-class classification: *complementary*, *substitute*, and *unrelated*. Following the methodology and the prompt from previous studies [32], we used Bayesian-optimized [1] logistic regression with a 424-dimensional content-based feature vector, and performed LLM-based annotation using GPT-4o-mini [18]. Once initialized, KARL enhances the classifier through the following iterative four-step process, which is illustrated in Figure 1:

Step 1: Candidate Pair Sampling Because processing the entire set of all possible pairs $\mathcal{P}^U$ is computationally infeasible owing to memory constraints, we sample a subset $\mathcal{P}^U_t$ in each round t. We employed a two-stage hierarchical sampling approach to ensure that this subset was not biased towards specific categories. We selected up to 10 query items from each of the 368 fine-grained categories (e.g., ballpoint pens), and thereafter paired each with up to 100 candidates from the same broad category (e.g., Office Supplies).

Step 2: Uncertainty Sampling The uncertainty scores were calculated based on the model prediction probabilities for each pair in $\mathcal{P}^U_t$ and the most uncertain pair per fine-grained category was selected as $\mathcal{P}^S_t$. This ensures diverse relationship representation while preventing similar-pair dominance.

Step 3: LLM-Based Annotation The LLM annotates each pair from $\mathcal{P}^S_t$ into nine FBLs classes using a prompt that incorporates the description of the pair. The 9-class output is then systematically mapped to the traditional 3-class: FBLs(A) map to *substitute*, FBLs(B-*,C-*) map to *complementary*, and FBLs(D,E) map to *unrelated*. We applied a consistency protocol to ensure the reliability of the LLM-annotated labels [19]. Under this protocol, only pairs in which the three independent labels are identical are adopted into $\mathcal{D}^L_t$.

Step 4: Model Retraining The accumulated $\mathcal{D}^L$ is integrated with $\mathcal{D}^H$ for retraining. We employed bagging ensemble to address the class imbalance in $\mathcal{D}^L$ [3]: ten balanced subsets from $\mathcal{D}^L$ via random undersampling were each combined with $\mathcal{D}^H$ to train separate classifiers. The final predictions were averaged across the classifiers, with the aggregated probabilities used in Step 2 of the next round.

4 Experiments

4.1 Experimental Setup

We observed changes in the ML classifier accuracy and training data diversity over 20 active learning loop rounds to evaluate the efficiency and convergence properties of KARL.

Datasets. Our study utilized an item dataset provided by ASKUL Corporation[1] including rich item attributes such as title, description, hierarchical categories, and more. To test how well KARL works in ID and OOD settings, we used two human-annotated FBLs datasets, $\mathfrak{D}^H_{id}$ and $\mathfrak{D}^H_{ood}$[2] [24, 32]:

- $\mathfrak{D}^H_{id}$ (ID): This set contains 2,625 labeled pairs, sampled based on high co-occurrence patterns and supplemented with web-sourced pairs. This dataset comprises 591 complementary, 410 substitute, and 1,624 unrelated pairs.
- $\mathfrak{D}^H_{ood}$ (OOD): This set contains 2,790 labeled pairs created by sampling one query item from each of 366 fine-grained categories and pairing it with another item based on BBLs. This dataset comprises 375 complementary, 2,024 substitute, and 391 unrelated pairs.

Both comprised item pairs from the "Office Supplies/Stationery" and "Household Goods/Kitchenware" categories, with three independent human annotations per pair, from which we retained only pairs with majority-agreed labels. To evaluate ID accuracy, we employed 5-fold nested cross-validation [8] on $\mathfrak{D}^H_{id}$, training models on each training fold and reporting averaged results across all test folds. For OOD evaluation, we tested these models on the entire $\mathfrak{D}^H_{ood}$. The severe distributional shift between these datasets is evident: a classifier trained on $\mathfrak{D}^H_{id}$ achieved only 0.44 macro-F1 on $\mathfrak{D}^H_{ood}$, as detailed in Section 4.3. In addition, our unlabeled pair pool $\mathcal{P}^U$ consisted of item pairs within the same item categories found on the same e-commerce source as the FBL datasets.

Uncertainty Sampling Methods. We compared three uncertainty sampling methods to score uncertainty in Step 2:

1. **Random**: Baseline method for randomly selecting pairs by assigning uniform random scores to each pair.
2. **Query-by-Committee (QBC)** [23]: Select pairs with the highest prediction variance across ten bagging models, where variance is computed over the predicted class probabilities.
3. **Margin** [21, 26]: Select pairs with the smallest probability margin between the top two predicted classes.

Evaluation Metrics. We evaluated KARL using two metrics: classification accuracy and training data diversity. Classification accuracy was measured by macro-F1, which averages the F1-score across all classes. Training data diversity was quantified using a Pearson correlation metric $\rho : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow [-1, 1]$:

$$diversity(X) = 1 - \frac{1}{n(n-1)} \sum_{i \neq j} \left| \rho(X_{i\cdot}, X_{j\cdot}) \right|$$

where n is the number of training pairs and $X_{i\cdot}$ denotes the feature vector of the i-th pair. Higher $diversity(X)$ show greater diversity. While other diversity metrics might provide complementary
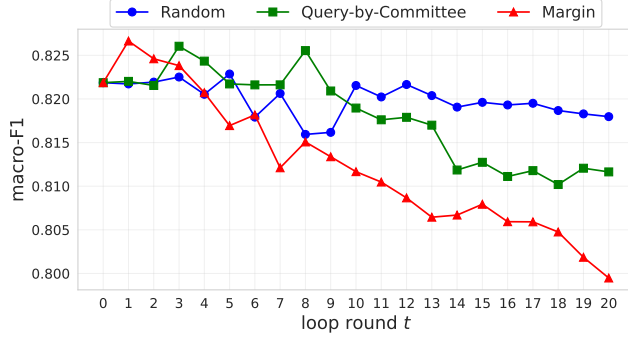
Figure 2: ID macro-F1 score on $\mathfrak{D}_{id}^{H}$, averaged across the 5 test folds of nested cross-validation.

Figure 3: OOD macro-F1 score on $\mathfrak{D}_{ood}^{H}$.

insights, we adopted this correlation-based approach as a simple baseline for quantifying diversity.

4.2 ID Accuracy Analysis (RQ1)

Figure 2 shows the macro-F1 progression on the ID test set over 20 rounds. Whereas the baseline model (loop 0) started with high accuracy, consistent with previous studies [32], KARL offered only marginal gains ($\leq 0.5\%$) before steadily degrading the accuracy. This degradation was most pronounced with uncertainty-based sampling methods such as QBC and Margin. This result suggests that the model has already captured sufficient ID relationships, making additional diverse data counterproductive. Rather than providing useful information, ambiguous samples from uncertainty sampling acted as "noise" that disrupted the model's stable distribution of the model and caused an accuracy-degrading shift.

4.3 OOD Accuracy Analysis (RQ2)

In contrast to the ID scenario, KARL proved highly effective for the OOD test set. As shown in Figure 3, KARL dramatically improved macro-F1 by up to 37% compared to the baseline. The superiority of the uncertainty sampling methods (QBC, Margin) over Random is particularly noteworthy, offering two distinct advantages. First, they are significantly more cost-efficient; they achieve any given level of accuracy in far fewer rounds than Random, thus minimizing additional annotation costs. Second, they increase the peak accuracy of the model. Although all the methods eventually plateaued at around loop 15, the final accuracy achieved by the uncertainty-based methods was up to 6.6% higher than that of Random, resulting in a more capable classifier. This evidence demonstrates that selective sampling in an OOD context not only accelerates learning but also raises the ceiling of the potential of the model.

4.4 Diversity–Accuracy Relation Analysis (RQ3)

To understand the mechanisms behind the contrasting results in RQ1 and RQ2, we analyzed the correlation between training data diversity and classifier accuracy gains. In Figures 4 and 5, we plot the accuracy gain against the diversity gain, with both metrics measured as changes from the baseline at loop 0. Each scatter plot shows the correspondence for a specific model at a specific loop
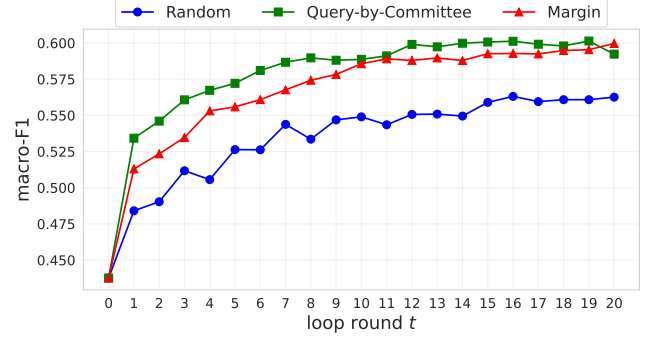

(a) ID setting on $\mathcal{D}_{id}^{H}$ (b) OOD setting on $\mathcal{D}_{ood}^{H}$
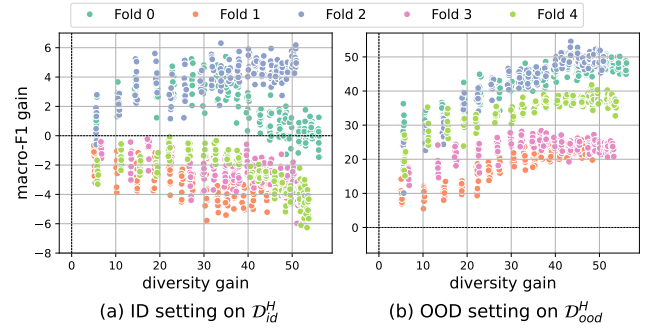
Figure 4: Correlation between accuracy gain and diversity gain in the training set for QBC.

round, comprising 200 data points–derived from the product of outer folds, loop rounds, and bagging models.

In the ID setting (panel (a)), although some folds showed a modest initial accuracy improvement, adding diversity beyond a certain threshold proved counterproductive, leading to a consistent decline in accuracy. Conversely, in the OOD setting (panel (b)), the diversity gain exhibits a strong and consistent positive correlation with the macro-F1 gain across all folds, resulting in a substantial improvement in accuracy of up to approximately 50%. This demonstrates that while diversity is a key driver of knowledge expansion in unfamiliar feature spaces, it can disrupt well-learned distributions in familiar ones.

5 Conclusion

This study presented KARL, a framework that addresses the cost and scalability challenges of FBLs by synergizing a cost-effective LLM annotator with an active learning strategy prioritizing informative samples. Our experiments demonstrated the high effectiveness of KARL in OOD settings, where increased data diversity directly promoted the acquisition of new knowledge, leading to improved generalization. Conversely, its effectiveness was limited in ID settings, as the excessive pursuit of diversity proved counterproductive to knowledge refinement by disrupting the learned data distribution. These contrasting results suggest that future frameworks should implement context-aware dual modes: preserving
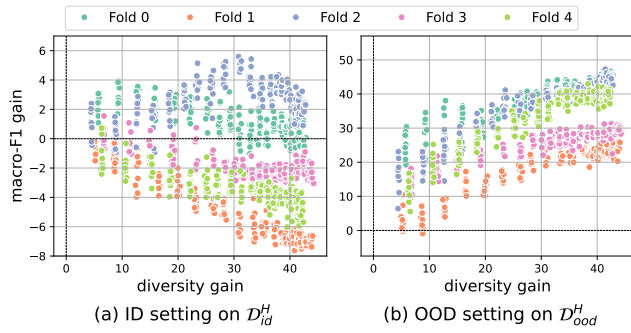
Figure 5: Correlation between accuracy gain and diversity gain in the training set for Margin.

learned distributions in ID settings while aggressively exploring in OOD settings. Such adaptive strategies could use confidence thresholds to automatically switch between conservative and exploratory sampling.

This study has several limitations that point to future research. First, the use of logistic regression may limit accuracy in OOD settings, as its linear decision boundaries may be insufficient for capturing complex non-linear complementary relationships. Future work should explore non-linear models such as gradient boosting or neural networks. Second, while our framework relies on GPT-4o-mini for annotation, the generalizability to other LLMs remains unexplored. Future studies should compare different LLMs and develop better methods for estimating label quality.

Acknowledgments

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 2623–2631. doi:10.1145/3292500.3330701

[2] K. Bibas, O. S. Shalom, and D. Jannach. 2023. Semi-supervised Adversarial Learning for Complementary Item Recommendation. In *Proc. ACM Web Conf. 2023.* 1804–1812. doi:10.1145/3543507.3583462

[3] L. Breiman. 1996. Bagging Predictors. *Mach. Learn.* 24 (1996), 123–140. doi:10.1007/BF00058655

[4] K. Brinker. 2003. Incorporating Diversity in Active Learning with Support Vector Machines. In *Proc. 20th Int. Conf. Int. Conf. Mach. Learn.* 59–66.

[5] H. Chen, J. He, W. Xu, T. Feng, M. Liu, T. Song, R. Yao, and Y. Qiao. 2023. Enhanced Multi-Relationships Integration Graph Convolutional Network for Inferring Substitutable and Complementary Items. *Proc. AAAI Conf. Artif. Intell.* 37, 4 (2023), 4157–4165. doi:10.1609/aaai.v37i4.25532

[6] Y. Cui, F. Liu, P. Wang, B. Wang, H. Tang, Y. Wan, J. Wang, and J. Chen. 2024. Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Models. In *Proc. 18th ACM Conf. Recommender Systems.* 507–517. doi:10.1145/3640457.3688118

[7] N. Entezari, Evangelos E. Papalexakis, H. Wang, S. Rao, and S. K. Prasad. 2021. Tensor-based Complementary Product Recommendation. In *Proc. 2021 IEEE Int. Conf. Big Data.* 409–415. doi:10.1109/BigData52589.2021.9671938

[8] P. Filzmoser, B. Liebmann, and K. Varmuza. 2009. Repeated Double Cross Validation. *J. Chemom.* 23, 160–171. Issue 4.

[9] J. Hao, T. Zhao, J. Li, X. L. Dong, C. Faloutsos, Y. Sun, and W. Wang. 2020. P-Companion: A Principled Framework for Diversified Complementary Product Recommendation. In *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.* 2517–2524. doi:10.1145/3340531.3412732

[10] A. Hein, S. Röhrl, T. Grobel, M. Lengl, N. Hafez, M. Knopp, C. Klenk, D. Heim, O. Hayden, and K. Diepold. 2022. A Comparison of Uncertainty Quantification

[11] Methods for Active Learning in Image Classification. In *Proc. 2022 Int. Jt. Conf. Neural Netw.* 1–8. doi:10.1109/IJCNN55064.2022.9892240

[11] N. Kholodna, S. Julka, M. Khodadadi, M. N. Gumus, and M. Granitzer. 2024. LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages. In *Proc. Mach. Learn. Knowl Discov. Databases Appl. Data Sci. Track: Eur. Conf., ECML PKDD 2024, Part X.* 397–412. doi:10.1007/978-3-031-70381-2_25

[12] L. Li and Z. Du. 2024. Complementary Recommendation in E-commerce: Definition, Approaches, and Future Directions. arXiv:2403.16135 [cs.IR]

[13] Z. Li, Y. Liang, M. Wang, S. Yoon, J. Shi, X. Shen, X. He, C. Zhang, W. Wu, H. Wang, J. Li, J. Chan, and Y. Zhang. 2024. ECCR: Explainable and Coherent Complement Recommendation Based on Large Language Models. In *Proc. KDD 2024 Workshop Gener. AI Recomm. Syst. Pers.*

[14] H. Luo, X. Meng, S. Wang, H. Cao, W. Zhang, Y. Wang, and Y. Zhang. 2024. Spectral-Based Graph Neural Networks for Complementary Item Recommendation. arXiv:2401.02130 [cs.IR]

[15] L. Ma, J. Xu, J. H.D. Cho, E. Korpeoglu, S. Kumar, and K. Achan. 2021. NEAT: A Label Noise-resistant Complementary Item Recommender System with Trustworthy Evaluation. In *Proc. 2021 IEEE Int. Conf. Big Data.* 469–479. doi:10.1109/BigData52589.2021.9671870

[16] J. McAuley, R. Pandey, and J. Leskovec. 2015. Inferring Networks of Substitutable and Complementary Products. In *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 785–794. doi:10.1145/2783258.2783381

[17] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal. 2023. Human-in-the-loop Machine Learning: a State of the Art. *Artif. Intell. Rev.* 56 (2023), 3005–3054.

[18] OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, et al. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL]

[19] N. Pangakis, S. Wolken, and N. Fasching. 2023. Automated Annotation with Generative AI Requires Validation. arXiv:2306.00176 [cs.IR]

[20] R. Papso. 2023. Complementary Product Recommendation for Long-tail Products. In *Proc. 17th ACM Conf. Recomm. Syst.* 1305–1311. doi:10.1145/3604915.3608864

[21] T. Scheffer, C. Decomain, and S Wrobel. 2001. Active Hidden Markov Models for Information Extraction. *Adv. Intell. Data Anal.* (2001), 309–318. doi:10.1007/3-540-44816-0_31

[22] Burr Settles. 2009. *Active Learning Literature Survey.* Technical Report. University of Wisconsin-Madison.

[23] H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proc. 5th Annu. Workshop Comput. Learn. Theory.* 287–294. doi:10.1145/130385.130417

[24] K. Sugahara, C. Yamasaki, and K. Okamoto. 2024. Is It Really Complementary? Revisiting Behavior-based Labels for Complementary Recommendation. In *Proc. 18th ACM Conf. Recomm. Syst.* 1091–1095. doi:10.1145/3640457.3691705

[25] A. Tharwat and W. Schenck. 2023. A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions. *Math.* 11, 4 (2023), 820. doi:10.3390/math11040820

[26] S. Tong and D. Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2 (2002), 45–66. doi:10.1162/153244302760185243

[27] C. Tseng, J. Song, Z. Bi, T. Wang, C. X. Liang, and M. Liu. 2025. Active Learning Methods for Efficient Data Utilization and Model Performance Enhancement. arXiv:2504.16136 [cs.LG]

[28] Z. Wang, Z. Jiang, Z. Ren, J. Tang, and D. Yin. 2018. A Path-Constrained Framework for Discriminating Substitutable and Complementary Products in E-Commerce. In *Proc. 11th ACM Int. Conf. Web Search Data Min.* 619–627. doi:10.1145/3159652.3159710

[29] Y. Xia, S. Mukherjee, Z. Xie, J. Wu, X. Li, R. Aponte, H. Lyu, J. Barrow, H. Chen, F. Dernoncourt, et al. 2025. From Selection to Generation: A Survey of LLM-based Active Learning. arXiv:2502.11767 [cs.LG]

[30] R. Xiao, Y. Dong, J. Zhao, R. Wu, M. Lin, G. Chen, and H. Wang. 2023. FreeAL: Towards Human-Free Active Learning in the Era of Large Language Models. In *Proc. 2023 Conf. Empir. Method. Nat. Lang. Process.* 14520–14535. doi:10.18653/v1/2023.emnlp-main.896

[31] D. Xu, C. Ruan, J. Cho, E. Korpeoglu, S. Kumar, and K. Achan. 2020. Knowledge-Aware Complementary Product Representation Learning. In *Proc. 13th Int. Conf. Web Search Data Min.* 681–689. doi:10.1145/3336191.3371854

[32] C. Yamasaki, K. Sugahara, Y. Nagi, and K. Okamoto. 2025. Function-based Labels for Complementary Recommendation: Definition, Annotation, and LLM-as-a-Judge. arXiv:2507.03945 [cs.IR]

[33] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou. 2023. LLMaAA: Making Large Language Models as Active Annotators. In *Proc. 2023 Conf. Empir. Method. Nat. Lang. Process.* 13088–13103. doi:10.18653/v1/2023.findings-emnlp.872

[34] Y. Zhang, H. Lu, W. Niu, and J. Caverlee. 2018. Quality-aware Neural Complementary Item Recommendation. In *Proc. 12th ACM Conf. Recomm. Syst.* 77–85. doi:10.1145/3240323.3240368