

Multi-Value-Product Retrieval-Augmented Generation for Industrial Product Attribute Value Identification

Huike Zou¹, Haiyang Yang¹, Yindu Su², Liyu Chen¹,
Chengbao Lian¹, Qingheng Zhang¹, Shuguang Han^{1,†}, Jufeng Chen¹

¹Xianyu of Alibaba ²Xiaohongshu Inc.

{zouhuike.zhk,qingheng.zqh,shuguang.sh}@alibaba-inc.com

Abstract

Identifying attribute values from product profiles is a key task for improving product search, recommendation, and business analytics on e-commerce platforms, which we called Product Attribute Value Identification (PAVI). However, existing PAVI methods face critical challenges, such as cascading errors, inability to handle out-of-distribution (OOD) attribute values, and lack of generalization capability. To address these limitations, we introduce Multi-Value-Product Retrieval-Augmented Generation (MVP-RAG), combining the strengths of retrieval, generation, and classification paradigms. MVP-RAG defines PAVI as a retrieval-generation task, where the product title description serves as the query, and products and attribute values act as the corpus. It first retrieves similar products of the same category and candidate attribute values, and then generates the standardized attribute values. The key advantages of this work are: (1) the proposal of a multi-level retrieval scheme, with products and attribute values as distinct hierarchical levels in PAVI domain (2) attribute value generation of large language model to significantly alleviate the OOD problem and (3) its successful deployment in a real-world industrial environment. Extensive experimental results on the dataset demonstrate that the proposed method performs better than the state-of-the-art baselines.

1 Introduction

Product attribute values are fundamental components in the e-commerce sector, serving as critical elements within the online retail ecosystem. From the perspective of a seller’s business strategy, meticulously curated product attribute values can significantly enhance traffic acquisition, increase product visibility, and ultimately drive transaction conversion. From the operational standpoint of

[†]Corresponding author.

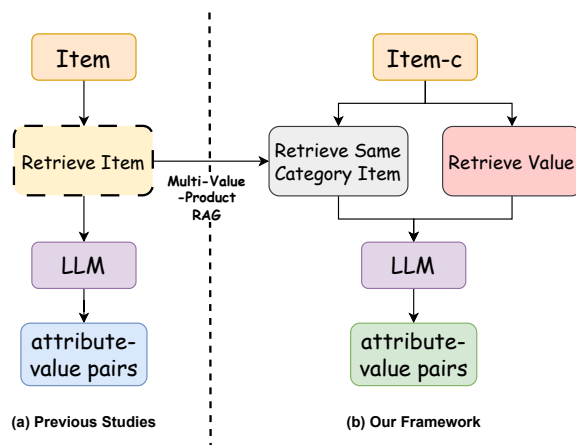


Figure 1: (a) shows the previous LLM and RAG methods, which either retrieve products from the product pool or directly generate attribute values. In contrast, (b) depicts our approach, which retrieves from both product and attribute value perspectives. This not only reduces the hallucination error caused by retrieving irrelevant products, but also ensures the stability of the results. Compared to previous studies, the core of our approach lies in the change in Retrieval.

e-commerce platforms, these attribute values provide fundamental data support for core functionalities, such as optimizing product display algorithms(Chen et al., 2024), constructing personalized recommendation systems(Truong et al., 2022), and enabling intelligent question-answering services(Gao et al., 2019), which are essential for improving user experience(Vashishth et al., 2024) and platform operational efficiency.

Professional sellers are usually able to standardize the selection of product attribute values. However, on second-hand trading platforms like Xianyu*, individual sellers often struggle to provide complete and accurate product attributes, which severely impacts the circulation efficiency of products. Therefore, achieving automated and precise product attribute value identification

*Xianyu is a C2C e-commerce platform in Alibaba

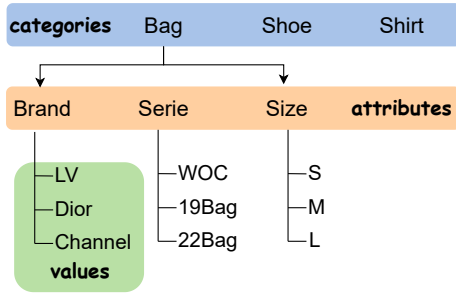


Figure 2: An illustration of a portion of the attribute taxonomy. Each category, such as *Bag*, is linked to multiple attributes, including *Brand*, *Serie*, and *Size*, with standardized values enumerated for each attribute (e.g., *LV*, *Dior*, and *Channel* for *Brand*).

(PAVI)(Shinzato et al., 2023) is of great significance. Currently, PAVI models can be broadly categorized into two-stage(Putthividhya and Hu, 2011; Zhang et al., 2021) and one-stage paradigms. The two-stage methods involve two steps: product attribute value extraction (PAVE) and alignment. They first extract non-standard attribute values from text using entity recognition(Zheng et al., 2018; Xu et al., 2019; Yan et al., 2021) or question-answering(Wang et al., 2020; Yang et al., 2023) techniques, then map them to standard attribute value. The one-stage methods can be further divided into three strategies: classification-based, generation-based, and retrieval-based. The classification-based approach(Chen et al., 2022) treats each attribute value as an independent class for multi-label classification. The generation-based approach(Sabeh et al., 2024; Nikolakopoulos et al., 2023; Shinzato et al., 2023) directly generates standard attribute values based on product descriptions. The retrieval-based approach (Su et al., 2025) encodes product descriptions and attribute values into vector representations, and then selects the relevant attribute values through similarity calculations.

The existing PAVI methods have their own strengths and weaknesses, and face various technical challenges. The two-stage PAVI methods first extract non-standard attribute values and then align them to the standard attribute value. This approach suffers from cascading errors, where attributes that cannot be identified in the first stage will be completely lost in the second stage. The one-stage methods can address this problem but introduce other issues. The classification-based methods treat attribute values as independent classes, which is a simple approach but is limited in its ability to identify OOD attribute values, making it unsuitable for

evolving e-commerce platforms. The generation-based methods view PAVI as an end-to-end task, which can solve the OOD problem, but suffer from the inability to control the output. The retrieval-based methods consider PAVI as a matching task between product and attribute values, but face issues with undefined thresholds and insensitivity to new attribute values. In summary, the existing methods face various problems, including implicit value identification, OOD recognition, and generalization capability.

In response to the limitations of existing methods, this paper proposes a novel **Multi-Value-Product Retrieval Augmented Generation (MVP-RAG)** approach, which innovatively combines the strengths of retrieval, generation, and classification paradigms, as shown in Fig 1. Our method defines PAVI as a retrieval-generation task: the product title description serves as the query, while the product pool and attribute taxonomy act as the corpus. We first retrieve similar products with identical category and attribute values, and then generate the standardized attribute values. For attribute value retrieval, we leverage TACLRL(Su et al., 2025) to select the top-K attribute values with the highest prediction scores as the candidate set. For product retrieval, we use the BGE model(Xiao et al., 2024) to generate vector representations of the product corpus, and then retrieve similar products based on similarity. Based on the retrieval results, we use a large language model (LLM)(Zhao et al., 2025) to generate the attribute values. Additionally, we construct a separate batch of out-of-distribution attribute values and incorporate them into the model training, to enhance the model’s ability to discover and predict unknown attribute values.

Our contributions are threefold: (1) We introduce a Multi-Value-Product Retrieval-Augmented Generation (MVP-RAG) method for the PAVI task. This approach addresses the limitations of existing PAVI methods, which often suffer from cascading errors, inability to handle out-of-distribution (OOD) attribute values, and lack of generalization capability. (2) We incorporate multi-retrieval techniques into PAVI framework, using TACLRL for attribute value retrieval and universal representation models for product retrieval. The retrieved attribute values and product information provide valuable contextual cues to guide the subsequent attribute value generation. (3) We validate the effectiveness of MVP-RAG through extensive experiments on proprietary datasets. In addition, MVP-RAG has

been successfully deployed in a real-world industrial environment.

2 Approach

As shown in Fig3, the overall workflow of MVP-RAG can be divided into two main components: multi-level attribute value-product retrieval (§2.1) and attribute value generation (§2.2).

2.1 Multi-level Attribute Value-Product Retrieval

The main purpose of the multi-level attribute value-product retrieval in MVP-RAG is to provide candidate attribute values for the subsequent attribute value generation, as well as to offer product-level few-shot information to guide the selection and generation process, as shown in Fig3

2.1.1 Attribute Value Retrieval

In the standard retrieval task(Schütze et al., 2008), given a query, the goal is to retrieve a set of relevant documents from the corpus. Correspondingly, we treat product information as the query and the standard attribute values as the corpus, and retrieve the most relevant attribute values based on the product information. TACLRL(Su et al., 2025) is the current state-of-the-art approach for the PAVI task based on retrieval, and we follow its setup. For each product, we concatenate the product title t and description d into a sentence in the format *title description* as the input. For each attribute value, we fully utilize the attribute tree information and construct a prompt in the format *a category with attribute being value* as the corpus input.

After constructing the product queries, we leverage the TACLRL encoder to obtain the vector representations. Then, for each product query, we calculate the similarity with the attribute values under its corresponding category, and select the top-k candidates as the results.

2.1.2 Product Retrieval

Attribute value retrieval can provide a candidate set for attribute value generation, but accurately selecting the most appropriate attribute value from these candidates remains a challenge. Product-level few-shot examples can provide reference cases to guide the model(OpenAI et al., 2024) in selecting the answers.

In a straightforward approach, we employ vector retrieval for similar product search. Specifically, we utilize the general-purpose representa-

tion model BGE(Xiao et al., 2024), to encode products into vector embeddings. We then compute the cosine similarity between the query product and a candidate product pool to select similar items. Furthermore, to ensure the consistency of attribute items, we restrict product retrieval within the same category.

2.2 Attribute Value Generation

After completing attribute value retrieval and product retrieval, we integrate them into a unified template as training data. The combined prompt consists of five parts: task definition, note, similar products in the same category, product information, and candidate attribute values. The specific details are shown in the appendixA.

We choose Qwen2.5(Qwen et al., 2025) as the base model, leveraging their strong performance and robust open-source ecosystem. The training objective of the model is the classic next-word prediction task. Specifically, we exclude the loss computation of the prompt prefix, and instead focus on the special tokens and model output tokens.

$$L = - \sum_{t=1}^T \log P(y_t | y_{<t}, X)$$

Here, X represents the input to the model, y represents the generated token, t represents the position of the generation, and T represents the final generated sequence.

3 Experiments

3.1 Data

We evaluate our model on the Xianyu-PAVI dataset and WDC-PAVI dataset. The Xianyu-PAVI dataset is sourced from the second-hand e-commerce platform Xianyu, where each product goes through a three-step process of model prediction, seller feedback, and manual review for attribute value assignment. The dataset contains 8,803 product categories, 26,645 category-attribute pairs, and 6.3 million category-attribute-attribute value triples. On average, each category has 3 attributes and 716 attribute values. The statistics of the dataset are shown in Tab 1. And the WDC-PAVI dataset comprises products spanning five distinct categories. The training set contains 1,066 products and 8,832 product-attribute pairs, among which 3,973 pairs have null values. The test set consists of 354 products and 2,937 product-attribute pairs, including 1,330 pairs with null values.

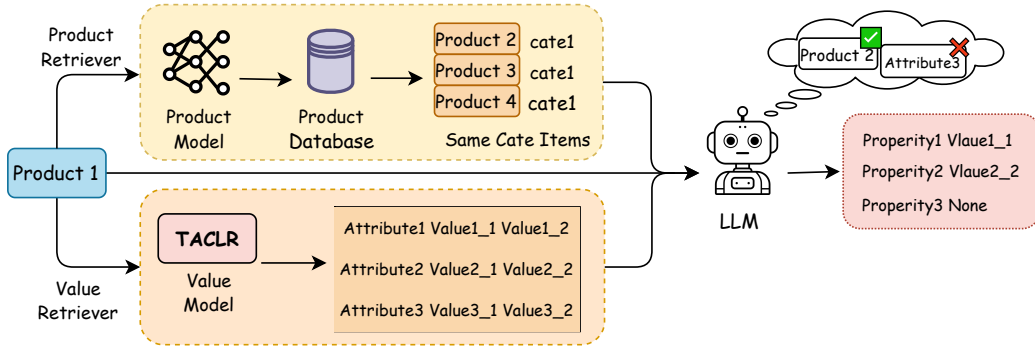


Figure 3: Pipeline of MVP-RAG

Table 1: Statistics of the Xianyu-PAVI.

Statistic	Train	Valid	Test
# Products	809,528	81,699	85,024
# PA Pairs	3,584,462	358,582	458,954
# Null Pairs	2,345,577	228,534	272,285

3.2 Metrics

Following TACLR(Su et al., 2025), we use micro-averaged precision@1, recall@1 and F1 score@1 to evaluate PAVI methods. For each attribute, the ground truth is a set of values V from the taxonomy. If the ground truth set is empty, a correct prediction (True Negative, TN) occurs when the model also predicts an empty set; otherwise, it is a False Positive (FP). When the ground truth set is not empty, the model’s top-1 output is a True Positive (TP) if it matches any ground truth value. Predicting an empty set in this case results in a False Negative (FN), while mismatched predictions are both False Positives (FP) and False Negatives (FN)[†], as it simultaneously introduces an error and misses the correct value. Table 3 summarizes these outcomes. Final precision, recall, and F1 scores are computed by aggregating TP, FP, and FN counts across the dataset, providing a comprehensive performance evaluation.

In addition, for nonempty set of ground-truth attribute values, we introduce the Coverage metric to quantify the overlap between the retrieved attribute values and the ground-truth set; coverage is deemed achieved when the intersection of the two sets is nonempty.

[†]In previous work (Shinzato et al., 2023), evaluation metrics did not explicitly consider the false positive (FP) case, and instances involving both false positives (FP) and false negatives (FN) were categorized solely as false positives. In contrast, we employ more rigorous evaluation metrics that independently account for both FP and FN cases.

3.3 Baselines

We compare our model with the one-stage PAVI methods, including classification, generation, and the current state-of-the-art retrieval-based methods:

- **BERT-CLS** This model(Chen et al., 2022) treats PAVI as a multi-label classification task, where each attribute value is considered an independent label. It also performs masking on irrelevant attribute values based on the product category.
- **LLM** The basic LLM model treats PAVI as a generative task, learning the product attribute vocabulary during training. It then uses the product title description as input and the attribute-attribute value pairs as output.
- **Product-RAG** Building upon the LLM model, RAG recalls relevant products to provide more contextual information for attribute value identification.
- **TACLR** The current state-of-the-art PAVI model, which treats PAVI as a retrieval task. It combines contrastive learning and adaptive thresholding techniques to select attribute values by computing the similarity between the product description and the attribute values.

3.4 Implementation Details

Retrieve Module. For retrieving similar products, we employ BGE-base(Xiao et al., 2024) as representation model and utilize cosine similarity as the similarity measurement. Furthermore, to ensure attribute consistency, retrieval is restricted to products within the same category. For retrieving attribute values, we employ the current state-of-the-art retrieval model, TACLR, to recall k attribute values.

Generate Module. We trained our model based on Qwen2.5-7B-Instruct, using retrieved products of same category and retrieved values. We trained our model by full parameter fine-tuning, The core hyperparameters are as follows: 3 training epochs, batch size of 16, AdamW optimizer, maximum-learning rate of 2×10^{-5} , 1% warmup steps, cosine learning rate scheduler.

4 Results

4.1 Main Result

Tab 2. shows the comparative results of our model on the Xianyu-PAVI dataset and the WDC-PAVI dataset. And MVP-RAG outperforms all baselines, achieving the state-of-the-art F1-score on both datasets. We attribute the excellent performance of MVP-RAG to two advantages: One is the retrieval-based candidate value generation which ensures the comprehensiveness of the attribute values, the other is incorporating similar products with corresponding attribute values from the same category as supplementary inputs to further improve the accuracy of the identification. The combination of these two strategies ensure both the completeness and accuracy of the attribute value recognition. Specifically, MVP-RAG outperforms the product-retrieval-based methods such as Qwen2.5(Product-RAG) by 26.3% on F1-score and outperforms the attribute-retrieval-based methods like previous SOTA TACL R by 3.3% on Xianyu-PAVI dataset. Meanwhile, MVP-RAG also achieves an overall performance improvement on the WDC-PAVI dataset.

4.2 Ablation Study

To evaluate the effectiveness of each core module in the MVP-RAG model, we conduct a series of ablation studies on the Xianyu-PAVI and WDC-PAVI datasets. Specifically, we adopt a module-by-module removal strategy to quantify the contribution of each component to the overall performance. As shown in Table 3, removing the attribute value retrieval module leads to a clear drop in the F1 scores on both datasets. This indicates that the module effectively transforms the original attribute value generation task into a classification-like problem, thereby reducing prediction difficulty and improving output accuracy. Similarly, removing the product retrieval module also results in performance degradation, demonstrating that this module enhances the model’s generalization capability

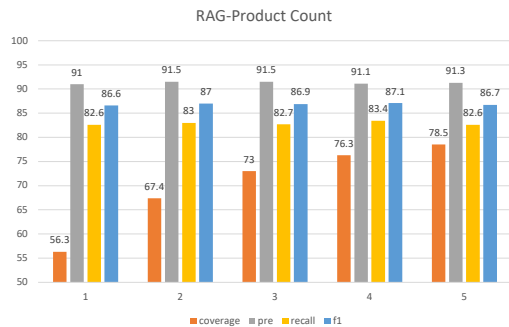


Figure 4: Performance analysis across Product Counts

for diverse expressions and long-tail attribute values by introducing relevant product instances as contextual references.

Furthermore, when the generation module is completely removed, causing the model to degenerate into TACL R, the F1 scores decrease by 3.3 and 6.8 on Xianyu-PAVI and WDC-PAVI, respectively. This highlights the critical role of the generation module in integrating retrieved information to produce accurate and fluent attribute values. These results collectively validate the necessity of each component in MVP-RAG and their synergistic effects, confirming the framework’s effectiveness and robustness in handling complex attribute value identification tasks.

4.3 Analysis

4.3.1 Impact of Product Counts

Fig 4. compares the impact on results as the number of retrieved products from the same category is varied. As the number of retrieved same-category products increases, the coverage of the test set’s attribute values rises from 56.3% to 78.5% and gradually stabilizes. In contrast, the model’s F1 score exhibits relatively minor fluctuations. This is because as the coverage increases with more similar product retrieval, attribute value generation becomes easier, leading to improved recall.

4.3.2 Impact of Attribute Value Counts

Fig 5. compares the impact on results as the number of retrieved attribute values is varied. As the number of retrieved attribute values increases, the coverage of the test set’s true attribute values gradually improves (94.9% -> 99.6%), approaching 100%. The model’s recall first increases and then decreases with the number of candidate attribute values, while precision declines overall. F1 and recall exhibit a similar pattern of initial increase

Table 2: Performance comparison of classification, generation, and retrieval methods on Xianyu-PAVI and WDC-PAVI

Paradigm	Method	Xianyu-PAVI			WDC-PAVI		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Classification	BERT-CLS	50.9	50.1	50.5	68.9	12.0	20.5
Generation	Qwen2.5 (zero-shot)	42.7	55.7	48.4	51.9	60.3	55.8
	Qwen2.5 (few-shot)	45.8	58.6	51.4	72.2	72.3	72.2
	Qwen2.5 (Product-RAG)	58.3	69.1	63.2	75.1	73.4	74.2
	Qwen2.5(fine-tune)	84.5	79.1	81.7	54.1	60.0	56.9
Retrieval	TACLRL	85.4	87.1	86.2	74.3	70.9	72.6
Multi-level Retrieval	MVP-RAG	93.8	85.3	89.5	81.1	77.7	79.4

Table 3: Ablation study of MVP-RAG

Method	Xianyu-PAVI			WDC-PAVI		
	Pre	Recall	F1	Pre	Recall	F1
Qwen2.5(fine-tune)	84.5	79.1	81.7	54.1	60.0	56.9
TACLRL	85.4	87.1	86.2	74.3	70.9	72.6
w/o product_retrieve	93.2	85.5	89.2	78.9	77.7	78.3
w/o value_retrieve	91.5	83.0	87.0	79.5	77.0	78.7
MVP-RAG	93.8	85.3	89.5	81.1	77.7	79.4

followed by decrease, with the F1 score peaking at 89.2% when the number of candidate values is 6. Clearly, as the number of candidate values increases, more information is introduced, with the initial information supplementation eventually leading to misinformation. Additionally, it is worth noting that compared to the TACLRL retrieval model that provides candidate values, the model’s precision improved by nearly 8% (85.4% -> 93.2%), indicating that MVP-RAG is effective and leads to a significant enhancement in the online user experience.

4.3.3 Accuracy Of Attribute Values And Products

As the number of recalled attribute values increases, the coverage of true attribute values gradually increases. When the true result appears in the candidate attribute value, f1 can reach 92.6%, which is 6.3% higher than when it does not appear. At the same time, we control the number of recalled products to be 4, and analyze whether the accuracy of the attribute value of the recalled product has an impact on the results. For simple attributes with clear distinguishing meanings, such as brand and model, when the attribute value of the recalled product is 75% wrong, MVP-RAG can still return the correct attribute value; when the error is higher than 75%, MVP-RAG has a probability of returning an incorrect attribute value. At the same time, think-

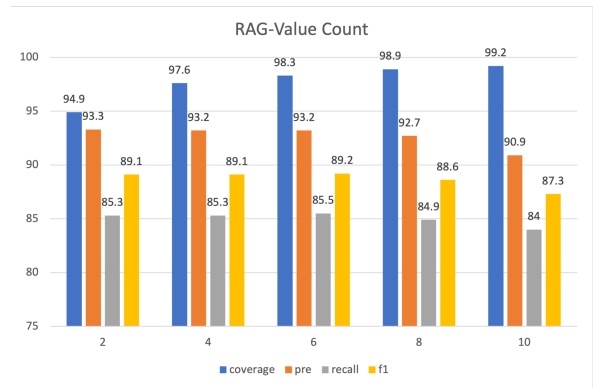


Figure 5: Performance analysis across Rag Value Counts

ing models such as DeepSeek will first analyze whether the attribute value of the given retrieved product is wrong, and then make corrections to predict the attribute value. For attributes that do not have clear meanings, such as color function, MVP-RAG focuses on the product’s own information and can still return the correct result when the attribute value of the recalled product is wrong. DeepSeek focuses on analyzing the information of the recalled product and is more inclined to be consistent with the attributes of the recalled product. The specific details are shown in the appendixB.

5 Related Work

Product Attribute Value Identification Product Attribute Value Identification (PAVI) is a critical task(Klemm and Kaufman, 2024) in e-commerce, with current methods broadly categorized into one-stage and two-stage paradigms. The two-stage approach first performs product attribute value extraction(PAVE)(Yang et al., 2022), followed by standard value mapping, which can lead to cascading errors. The one-stage paradigm includes classification-based, generation-based, and

retrieval-based methods. The classification-based approach (Fuchs and Acriche, 2022; Chen et al., 2022) defines PAVI as a multi-label classification problem over a finite set of values, but suffers from a significant limitation in its inability to identify out-of-distribution (OOD) values, a critical shortcoming in the rapidly evolving e-commerce domain. Retrieval-based method (Su et al., 2025) define PAVI as an attribute value retrieval task, but exhibit insensitivity to new attribute values and require redeployment to accommodate changes in the attribute library. Generation-based approaches cast PAVI as an end-to-end generation task, directly generating single or multi-attribute values from product descriptions, but face challenges with hallucination and output instability. The proposed model builds upon both generation-based and retrieval-based approaches, incorporating information from similar products to perform robust product attribute value identification.

Retrieval-Augmented Generation Retrieval-based approaches (Leto et al., 2024) have been widely demonstrated to be highly effective for knowledge-intensive tasks, particularly in output-constrained settings, where the introduction of relevant knowledge can significantly mitigate the issue of hallucination. (Sabeh et al., 2024) incorporated Retrieval-Augmented Generation (RAG) technologies in the PAVI task, utilizing a pre-trained T5 (Raffel et al., 2023) to retrieve similar products and leverage the information to enhance the generation, outperforming direct attribute value generation methods by a substantial margin. In contrast, the unique contribution of current research lies in its utilization of both product-level and attribute-value-level information for the retrieval-augmentation process. By successfully applying this approach in the e-commerce domain, the proposed model is able to address the hallucination problem while also improving performance on the specific task at hand.

6 Conclusion

In this work, we present multi-value-product retrieval-augmented generation (MVP-RAG) method for the PAVI task. This approach addresses the limitations of existing PAVI methods, which often suffer from cascading errors, inability to handle out-of-distribution (OOD) attribute values, and lack of generalization capability.

Comprehensive experiments on proprietary and

public datasets demonstrated MVP-RAG’s superiority over classification- and generation-based baselines. Notably, method achieved an F1 score of 89.5% on the large-scale Xianyu-PAVI dataset. Beyond these experimental results, MVP-RAG has been successfully deployed on the real-world e-commerce platform Xianyu, processing millions of product listings daily and seamlessly adapting to dynamic attribute taxonomies, making it a practical solution for large-scale industrial applications.

Limitations

MVP-RAG can accurately identify product attribute values, but there are still many limitations. First, only text information is used, and image or video information of the product has not been used. Multimodal input can provide information for attributes that are difficult to infer from text alone (such as color, material, or shape). Second, the reasoning output based on LLM takes a long time, and the model still needs to be accelerated and optimized in the future.

References

- Kang Chen, Qingheng Zhang, Chengbao Lian, Yixin Ji, Xuwei Liu, Shuguang Han, Guoqiang Wu, Fei Huang, and Jufeng Chen. 2024. Ipl: Leveraging multimodal large language models for intelligent product listing. *arXiv preprint arXiv:2410.16977*.
- Wei-Te Chen, Yandi Xia, and Keiji Shinzato. 2022. [Extreme multi-label classification with label masking for product attribute value extraction](#). In *Proceedings of the Fifth Workshop on e-Commerce and NLP (EC-NLP 5)*, pages 134–140, Dublin, Ireland. Association for Computational Linguistics.
- Gilad Fuchs and Yoni Acriche. 2022. [Product titles-to-attributes as a text-to-text task](#). In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 91–98, Dublin, Ireland. Association for Computational Linguistics.
- Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 429–437.
- Celine Klemm and Stefan Kaufman. 2024. The importance of circular attributes for consumer choice of fashion and textile products in australia. *Sustainable Production and Consumption*, 45:538–550.
- Alexandria Leto, Cecilia Aguerrebere, Ishwar Bhati, Ted Willke, Mariano Tepper, and Vy Ai Vo. 2024.

- Toward optimal search and retrieval for rag. *Preprint*, arXiv:2411.07396.
- Athanasios N. Nikolakopoulos, Swati Kaul, Siva Karthik Gade, Bella Dubrov, Umit Batur, and Suleiman Ali Khan. 2023. [Sage: Structured attribute value generation for billion-scale product catalogs](#). *Preprint*, arXiv:2309.05920.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Duangmanee Putthividhya and Junling Hu. 2011. [Bootstrapped named entity recognition for product attribute extraction](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Preprint*, arXiv:1910.10683.
- Kassem Sabeh, Robert Litschko, Mouna Kacimi, Barbara Plank, and Johann Gamper. 2024. [An empirical comparison of generative approaches for product attribute-value identification](#). *Preprint*, arXiv:2407.01137.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.
- Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. 2023. [A unified generative approach to product attribute-value identification](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6599–6612, Toronto, Canada. Association for Computational Linguistics.
- Yindu Su, Huike Zou, Lin Sun, Ting Zhang, Haiyang Yang, Liyu Chen, David Lo, Qingheng Zhang, Shuguang Han, and Jufeng Chen. 2025. [Taclr: A scalable and efficient retrieval-based method for industrial product attribute value identification](#). *arXiv preprint arXiv:2501.03835*.
- Quoc-Tuan Truong, Tong Zhao, Changhe Yuan, Jin Li, Jim Chan, Soo-Min Pantel, and Hady W Lauw. 2022. [Ampsum: Adaptive multiple-product summarization towards improving recommendation captions](#). In *Proceedings of the ACM Web Conference 2022*, pages 2978–2988.
- Tarun Kumar Vashishth, Kewal Krishan Sharma, Bhupendra Kumar, Sachin Chaudhary, Rajneesh Panwar, and 1 others. 2024. [Enhancing customer experience through ai-enabled content personalization in e-commerce marketing](#). *Advances in digital marketing in the era of artificial intelligence*, pages 7–32.
- Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. [Learning to extract attribute value from product via question answering: A multi-task approach](#). *KDD '20*, page 47–55, New York, NY, USA. Association for Computing Machinery.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. [Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, Florence, Italy. Association for Computational Linguistics.
- Jun Yan, Nasser Zalmout, Yan Liang, Christian Grant, Xiang Ren, and Xin Luna Dong. 2021. [AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4694–4705, Online. Association for Computational Linguistics.
- Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabza, Sinong Wang, Zenglin Xu, and Dongfang Liu. 2023. [MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada. Association for Computational Linguistics.
- 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 *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1256–1265.
- Danqing Zhang, Zheng Li, Tianyu Cao, Chen Luo, Tony Wu, Hanqing Lu, Yiwei Song, Bing Yin, Tuo Zhao, and Qiang Yang. 2021. [Queaco: Borrowing treasures from weakly-labeled behavior data for query attribute value extraction](#). In *Proceedings of*

the 30th ACM International Conference on Information and Knowledge Management, CIKM '21, page 4362–4372. ACM.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. *A survey of large language models*. Preprint, arXiv:2303.18223.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. *Opentag: Open attribute value extraction from product profiles*. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, page 1049–1058, New York, NY, USA. Association for Computing Machinery.

A Example Appendix

Task Description

Given the product description, category, similar same category product, attribute set, and candidate attribute values for each attribute, generate the attribute values of the product. Unrecognizable attributes can be returned as unknown.

Note

1. The attribute value of the product does not necessarily appear in the reference product attribute value and candidate attribute value
2. If the attribute value of the product exists but does not appear in the reference product attribute value and candidate attribute value, it can be generated, but it is best not to do so
3. If the attribute value of the product does not exist, return *None*
4. The probability of the given candidate attribute value decreases from the front to the back

Reference product information

Product description: 2 Sony E-mount Tamron back covers, 1 body front cover. New Year's sundries!

Category: SLR body cover

Attribute value:

Brand: Sony

Condition: None

Product description: Canon 60d back cover set button set button

Category: SLR body cover

Attribute value:

Brand: Canon

Condition: None

Product information

Product description: Nikon D series SLR camera front and rear covers, dust cover,

body cover, suitable for Nikon SLR camera D810/D7100/D610/D3200/D5100/D800 body cover, lens rear cover

Category: SLR body cover

Attribute collection: brand, condition

Attribute candidate value

Brand: Nikon, None

Condition: None, slight signs of use

Brand:Nikon

Condition:None

B Impact of incorrect product attribute values

Product.

Chanel new sale S engraved 19 black gold small size, how hard to buy not to say, no bargaining, only inspection treasure

Similar same category Product.

Chanel 20a Haute Couture, size 36 black short jacket with gold buttons, the same style as Zhou Xun, basically brand new, very good-looking, authentic guaranteed, supports identification!

Chanel four-pocket black jacket, a classic among Chanel classics, size 34, 99% new, purchased in a domestic counter, because it is an old model, it cannot be inspected, and those who can distinguish the authenticity can come, no bargaining, out of print collection, suitable for thin girls under 100 pounds, SF Express cash on delivery, no refund or exchange.

23 autumn and winter Chanel mer select Italy jin import black navy blue glitter silk 50 wool small fragrance care products anonymous eyiu bought size s if it is not suitable after arrival, rush to sell it. The goods have been received and sold.

Ready Stock

Chanel 23a black denim cc jacket size 36 Brand new full set, direct style can be sent for free inspection within 3-5 working days, direct style delivery, serious buyers only.

Explicit Attribute-Brand

MVP-RAG

When the brand of the given three products is changed to Dior, the result returned is always Chanel. However, when the brand of all four products is changed to Dior, the result returned is Dior.

DeepSeek

rethink

First, there are four examples in the reference product information provided by the user. Although "Chanel" is mentioned in the product description,

the brand in the attribute value is marked as "Dior", which is obviously inconsistent. This may be an error or a test case. However, according to the task description, I need to generate attribute values based on the given information and return unknown when it cannot be recognized. The description of the target product is "Chanel New Sale S Engraved 19 Black Gold Small", which clearly mentions "Chanel", so the brand should be "chanel". But the brands in the reference examples are all Dior. Here we need to judge whether to follow the error in the reference example or fill in the brand correctly according to the product description. According to the second point of the task description, if the attribute value exists but does not appear in the reference, it can be generated but it is better not to. But here the brand is clearly in the description, so it should be filled in correctly instead of following the wrong example.

Answer

No matter how many wrong results are given, it can always return to Chanel.

Explicit Attribute-Quality

MVP-RAG

Always returns a new attribute value that is consistent with the correct result.

DeepSeek

rethink

Let's first look at the information of the reference products. The description of the first product mentions "basically new", and the condition in the attributes is "almost new". The description of the second product says "99% new", and the condition is also "almost new". The descriptions of the third and fourth products mention "out" and "brand new complete set" respectively, and the condition is also "almost new". It seems that no matter whether the description is "basically new", "99% new" or "brand new", the condition is indicated by "almost new" in the attributes. Is this because in the second-hand or resale market, these descriptions are all classified as almost new, rather than brand new? Or it may be a unified processing method in the task, so even if the description says "brand new", the attributes still use "almost new".

Answer

When almost new appears, DeepSeek will tend to give almost new results.