GAVEL: Generative Attribute-Value Extraction Using LLMs on LLM-Augmented Datasets

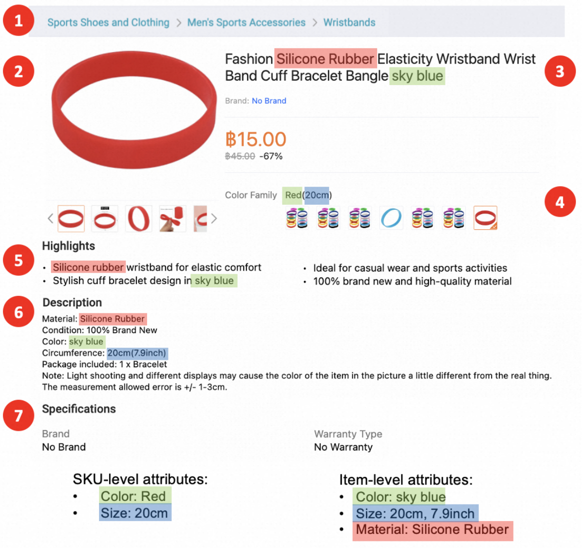
Pollawat Hongwimol, Dong Sheng, Zhang Li, Kai Liu, Xiufei Wang
Lazada, Alibaba Group
pollwat.h@alibaba-inc.com
frank.dong@lazada.com, zhangli835@gmail.com

Abstract

In the evolving e-commerce landscape, accurate product attribute-value extraction is crucial for enhancing user experience and increasing sales. This paper introduces GAVEL, a generative approach leveraging large language models (LLMs) to augment training data for attribute extraction from diverse textual sources. Our method extracts over 1,000 unique attributes across 2,000 product categories in multiple Southeast Asian languages, including Thai, Vietnamese, and Indonesian. Rigorous evaluations show significant improvements in accuracy and coverage compared to seller-provided attributes, with enhanced recall and F1 scores. Additionally, GAVEL reduces operational costs by minimizing instruction token usage and improves inference speed. The results of the A/B test indicate that our model has a positive impact on Gross Merchandise Value (GMV) per page view (PV) across all three operating countries. This research highlights the potential of generative techniques for optimizing attribute extraction in multi-language e-commerce applications.

1 Introduction

Product attributes significantly influence product search (Ai et al., 2019; Luo et al., 2022), recommendation systems (Luo et al., 2022; Gao et al., 2023), and buyers' decision-making processes (Zheng et al., 2020; Hafiz and Ali, 2019; Helfi et al., 2019). Inadequate or erroneous information on product pages can lead to adverse outcomes, such as a poor shopping experience, decreased visibility, and lower sales. For instance, Figure 1 illustrates a case where the color 'sky blue' is mentioned in the title, highlights, and description; however, it contradicts the Stock Keeping Unit (SKU) variant, which is listed as red. Such discrepancies can confuse potential buyers and negatively impact their purchasing decisions.



Sports Shoes and Clothing > Men's Sports Accessories > Wristbands

Fashion Silicone Rubber Elasticity Wristband Wrist Band Cuff Bracelet Bangle sky blue

Brand: No Brand

฿15.00
~~฿45.00~~ -67%

Color Family Red(20cm)

Highlights

- Silicone rubber wristband for elastic comfort
- Stylish cuff bracelet design in sky blue
- Ideal for casual wear and sports activities
- 100% brand new and high-quality material

Description
Material: Silicone Rubber
Condition: 100% Brand New
Color: sky blue
Circumference: 20cm(7.9inch)
Package included: 1 x Bracelet
Note: Light shooting and different displays may cause the color of the item in the picture a little different from the real thing.
The measurement allowed error is +/- 1-3cm.

Specifications

Brand	Warranty Type
No Brand	No Warranty

SKU-level attributes:

- Color: Red
- Size: 20cm

Item-level attributes:

- Color: sky blue
- Size: 20cm, 7.9inch
- Material: Silicone Rubber

Figure 1: An example of seller-provided information, including extracted attributes. The information consists of (1) category, (2) images, (3) title, (4) SKU variants, (5) highlights, (6) description, and (7) specifications. In this case, the color 'sky blue' appears in the title, highlights, and description; however, it contradicts the SKU variant's color, which is red.

In recent years, there has been significant research on Product Attribute-Value Extraction (PAVE) (Shinzato et al., 2023; Zou et al., 2024b; Zhu et al., 2020). Initial studies primarily employed sequence tagging methods using encoder-only models like BERT (Wang et al., 2020; Zhu et al., 2020; Chen et al., 2022; Embar et al., 2021; Xu et al., 2019; Deng et al., 2022). However, this approach has limitations in handling unseen and canonicalized values (Shinzato et al., 2023). To overcome these challenges, subsequent research has shifted from sequence tagging to sequence-to-sequence generation models, such as T5 and BART, which support more flexible output formats (Shinzato et al., 2023; Nikolakopoulos et al., 2023; Gong and Eldardiry, 2024; Wang et al., 2022; Roy et al., 2022; Sabeh et al., 2024; Roy et al., 2021). Addi-

Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing (KnowledgeNLP'25), pages 81–90
May 3, 2025 ©2025 Association for Computational Linguistics

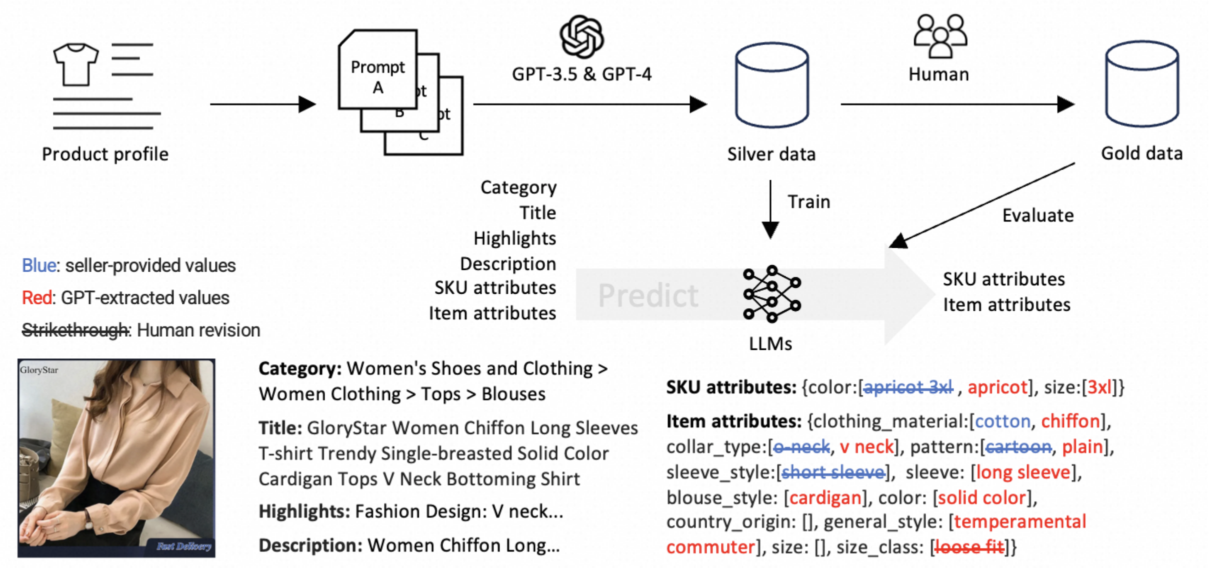

Product profile
Prompt A
B
C
pt
pt
GPT-3.5 & GPT-4
Silver data
Human
Gold data
Category
Title
Highlights
Description
SKU attributes
Item attributes
Train
Evaluate
Predict
LLMs
SKU attributes
Item attributes
Blue: seller-provided values
Red: GPT-extracted values
~~Strikethrough~~: Human revision
GloryStar
Fast Delivery
Category: Women's Shoes and Clothing > Women Clothing > Tops > Blouses
Title: GloryStar Women Chiffon Long Sleeves T-shirt Trendy Single-breasted Solid Color Cardigan Tops V Neck Bottoming Shirt
Highlights: Fashion Design: V neck...
Description: Women Chiffon Long...
SKU attributes: {color:[~~apricot 3xl~~ , apricot], size:[3xl]}
Item attributes: {clothing_material:[cotton, chiffon], collar_type:[~~o-neck~~, v neck], pattern:[~~cartoon~~, plain], sleeve_style:[~~short sleeve~~], sleeve: [long sleeve], blouse_style: [cardigan], color: [solid color], country_origin: [], general_style: [temperamental commuter], size: [], size_class: [~~loose fit~~]}

Figure 2: An overview of the GAVEL pipeline for generative attribute-value extraction using LLMs on LLM-augmented datasets. The GAVEL process begins with a product profile containing seller-provided information (indicated in blue) and employs prompts to extract and predict attributes (highlighted in red) utilizing GPT-3.5 and GPT-4. Silver data is used for training, while human evaluation is conducted to refine the final gold data for testing purposes.

tionally, current research leverages the zero-shot and few-shot capabilities of large-scale autoregressive models, such as GPT-3.5 and GPT-4, to enhance attribute extraction processes (Brinkmann et al., 2024b; Brinkmann et al., 2024a; Fang et al., 2024). This evolution has demonstrated that effective attribute value extraction significantly improves key e-commerce metrics, including Click-Through Rate (CTR) and Add-To-Cart Rate (ATC) (Fang et al., 2024).

Despite these advancements, existing publicly available datasets for PAVE face several limitations. For instance, the AE-110k dataset (Xu et al., 2019) is confined to the Sports & Entertainment category, offering data in the form of title-attribute-value triplets. While MEPAVE (Zhu et al., 2020) broadens its data sources to include images and descriptions, it still possesses a limited number of unique attributes. OA-Mine (Zhang et al., 2022) encompasses titles for 100 product types but lacks comprehensive information beyond the title itself. MAVE (Yang et al., 2021) includes 1,257 categories and various input types; however, it requires that explicit attribute values be present in the text, which complicates the extraction of unseen and canonicalized values. Although ImplicitAVE (Zou et al., 2024a) aims to address this shortcoming through a two-round human inspection process that annotates 25 attributes across five domains, it remains limited to titles and images. Furthermore, existing datasets predominantly rely on item-level information, which may lead to inaccuracies in attribute value extraction for a specific SKU variant, as demonstrated in Figure 1.

Expanding beyond these challenges, it is crucial to recognize the growing e-commerce landscape in Southeast Asia, where the need for PAVE solutions is becoming increasingly pressing. This region is home to a rich diversity of languages, including Thai, Vietnamese, Indonesian, and English. However, most PAVE research to date has primarily focused on English (Brinkmann et al., 2024b; Fang et al., 2024; Yang et al., 2023), Chinese (Zhu et al., 2020; Deng et al., 2023), and Japanese (Shinzato et al., 2023; Chen et al., 2022). To the best of our knowledge, there has been no prior study exploring the potential for enhancing PAVE in Southeast Asian languages.

To tackle these challenges, we propose an efficient pipeline for augmenting training data for generative attribute-value extraction, as depicted in Figure 2. Our key contributions are summarized as follows:

- We experiment with a comprehensive set of attributes, consisting of over 1,000 unique attributes across 2,000 categories.
- We incorporate successful methodologies

from prior research, including the utilization of generation-based models with zero-shot capabilities, the effective incorporation of textual data from product profiles, and the prediction of multiple attribute values.

- We reformulate the task to include seller-provided attribute values within input data. This enhancement improves output quality, as valuable information is sometimes misallocated under incorrect attribute values.
- We introduce GAVEL, a novel pipeline for generating high-quality training data for PAVE, enabling the training of smaller models with shorter, more concise prompts.

2 Related Work

2.1 Attribute Value Extraction

Attribute value extraction aims to derive specific values from product information. Traditionally, this has involved sequence tagging techniques employing models like Long Short-Term Memory (LSTM) or Bidirectional Encoder Representations from Transformers (BERT) (Embar et al., 2021; Xu et al., 2019; Wang et al., 2020). However, these methods often struggle with unseen values. To address these limitations, Roy et al. (2022) proposed generative frameworks that jointly extract attributes and values using the Text-to-Text Transfer Transformer (T5), demonstrating that such approaches outperform traditional tagging for single-value sentences.

Recent work has explored LLMs like GPT-3.5 and GPT-4 for attribute extraction, showing improved data efficiency and robustness to unseen values compared to traditional pretrained language models (Brinkmann et al., 2024b). Despite these advancements, many studies focus on English products, with few addressing the complexities of Southeast Asian languages (Brinkmann et al., 2024b; Yang et al., 2023), underscoring a significant gap in multilingual PAVE research.

2.2 Attribute Value Extraction Datasets

A variety of datasets have been created to support PAVE research. Table 1 offers a detailed summary of existing datasets alongside our datasets. Notably, these datasets vary significantly across several dimensions, including product categories, SKU counts, attribute counts, unique attributes, languages, and data sources.

OpenTag (Zheng et al., 2018) comprises 10k SKUs, with a total of 13k attribute instances, across three categories collected from Amazon. This dataset includes attributes extracted from product titles, descriptions, and highlights, and is exclusively presented in English. AE-110k (Xu et al., 2019) is concentrated within a single Sports & Entertainment category, featuring a considerably larger SKU count of 50k, yielding 110k attribute instances. Attributes derive solely from product titles, with data collected from AliExpress and created without human annotation. MEPAVE (Zhu et al., 2020) offers a broader variety of categories, containing seven categories with 34k SKUs and 87k attributes, with human annotations. Attributes in this dataset are predominantly extracted from images and descriptions and are published in Chinese. MAVE (Yang et al., 2021) includes an extensive catalog of 1.3k categories and a substantial total of 3 million attribute instances. The products are sourced from the Amazon Review Dataset and do not include any human annotations. OA-Mine (Zhang et al., 2022) encompasses ten distinct categories, comprising 11k attributes. This dataset focuses on attributes derived from titles, descriptions, and highlights, all collected from Amazon in English. Only the development and test sets are annotated by human. ImplicitAVE (Zou et al., 2024a) presents a dataset featuring five categories and 70k attributes, focusing on attributes sourced from title and image data. This dataset represents an enhanced version of MAVE, with the evaluation set re-annotated by a team of five Ph.D. students to improve reliability.

3 Methods

3.1 Attribute Mining

Our approach to enhancing seller-provided attributes utilizes three distinct prompts submitted to GPT-3.5 (Ouyang et al., 2024) and GPT-4 (OpenAI et al., 2024), as illustrated in Figure 3. The first prompt verifies brand names in the title and highlights. The second prompt focuses on SKU-level attributes, addressing frequent misclassifications by providing detailed instructions for the extraction of five common attributes. The third prompt targets item-level attributes from titles, highlights, and descriptions, requiring comprehensive instructions to avoid the model simply replicating examples instead of accurately extracting values. This prompt includes value examples and bullet-point instructions to regulate the expected output format.

Dataset	#Cate	#SKU	#Attr	#Unique	Lang	Source
OpenTag	3	10k	13k	4	en	title, desc, highlights
AE-110k	1	50k	110k	4	en	title
MEPAVE	7	34k	87k	26	zh	image, desc
MAVE	1.3k	2.2M	3M	2.5k	en	title, desc, highlights
OA-Mine	10	2k	11k	10	en	title, desc, highlights
ImplicitAVE	5	70k	70k	25	en	title, image
Lzd-ID-train (ours)	2k	163k	739k	1.2k	en, id	title, desc, highlights, sku attributes, specifications
Lzd-TH-train (ours)	2k	152k	648k	1.2k	en, th	
Lzd-VN-train (ours)	2k	152k	705k	1.2k	en, vi	
Lzd-ID-test (ours)	314	1k	6.3k	313	en, id	
Lzd-TH-test (ours)	352	1k	5.8k	372	en, th	
Lzd-VN-test (ours)	353	1k	6.3k	417	en, vi	
Total (ours)	**2k**	**470k**	**2.1M**	**1.2k**	**en, th, id, vi**	

Table 1: A summary of existing datasets compared to our datasets.

3.2 Dataset Construction

The attributes mined in Section 3.1, along with seller-provided attributes, are categorized as silver labels. A rule-based processing algorithm resolves conflicts by prioritizing mined attributes; if there are no contradictions, both types are retained. Experienced e-commerce annotators from Indonesia, Thailand, and Vietnam evaluate these silver-labeled attributes to generate gold labels.

To optimize token efficiency during training and inference, we compile all attribute information into a concise prompt (see Figure 3). This prompt includes seller-provided data such as titles, highlights, descriptions, SKU attributes, and item specifications, allowing models to predict multiple attributes and values simultaneously.

Table 1 summarizes our three datasets, encompassing 2,000 categories and over two million attribute instances. Each training category includes an equal number of SKUs, while the test set consists of 1,000 randomly selected and mutually exclusive SKUs, ensuring diverse representation and comprehensive evaluation of model performance.

3.3 Model Fine-tuning

In this study, we fine-tune four LLMs with comparable parameter counts, which are accessible on Hugging Face[1]. The models include *Gemma-2-9b* (Team et al., 2024), *Llama-3.1-8B* (Dubey et al., 2024), *Qwen2.5-7B* (Yang et al., 2024), and *SeaLLMs-v3-7B* (Zhang et al., 2024). Our training employs Low-Rank Adaptation (LoRA) (Hu et al., 2022) with bf16 precision, specifically setting `lora_target=all` to facilitate comprehensive adaptation across all layers of the models. We split the dataset into training and validation sets, allocating 15% for validation to ensure robust evaluation of model performance. The training process utilizes a cosine learning rate scheduler and incorporates early stopping criteria to further optimize performance. We set the training and evaluation batch sizes to 2 and employ gradient accumulation over 8 steps, leading to a total of 10 training epochs, which allows for thorough learning from the dataset. Additionally, evaluations and logging are conducted at intervals of 500 steps to monitor convergence effectively. Notably, this experiment utilizes 4 PPU 810 cards provided by Alibaba Cloud to fine-tune the models.

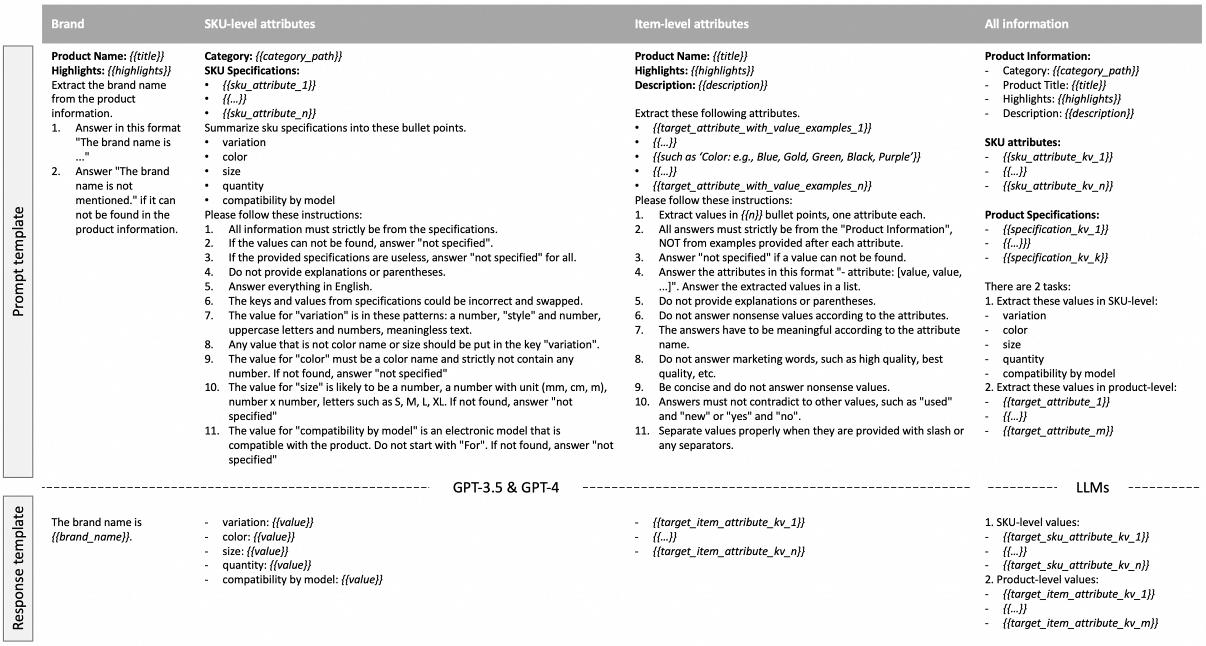
4 Experimental Setup

In this section, we outline the experimental framework employed in this study, utilizing data sourced from Lazada, a prominent e-commerce platform in Southeast Asia. The information collected encompasses product profiles and various attributes relevant to our analysis, establishing a comprehensive basis for the subsequent investigations.

4.1 Data Sources

The product pages on Lazada contain extensive information furnished by sellers, which includes both textual and visual components, as illustrated in Figure 1. SKU-level attributes refer to specific variants of a product, while item-level attributes cover common characteristics shared across all variants. We extract SKU-level attributes from individual

[1] https://huggingface.co/



	Brand	SKU-level attributes	Item-level attributes	All information
Prompt template	**Product Name:** *{{title}}* **Highlights:** *{{highlights}}* Extract the brand name from the product information. 1. Answer in this format "The brand name is ..." 2. Answer "The brand name is not mentioned." if it can not be found in the product information.	**Category:** *{{category_path}}* **SKU Specifications:** • *{{sku_attribute_1}}* • *{{...}}* • *{{sku_attribute_n}}* Summarize sku specifications into these bullet points. • variation • color • size • quantity • compatibility by model Please follow these instructions: 1. All information must strictly be from the specifications. 2. If the values can not be found, answer "not specified". 3. If the provided specifications are useless, answer "not specified" for all. 4. Do not provide explanations or parentheses. 5. Answer everything in English. 6. The keys and values from specifications could be incorrect and swapped. 7. The value for "variation" is in these patterns: a number, "style" and number, uppercase letters and numbers, meaningless text. 8. Any value that is not color name or size should be put in the key "variation". 9. The value for "color" must be a color name and strictly not contain any number. If not found, answer "not specified" 10. The value for "size" is likely to be a number, a number with unit (mm, cm, m), number x number, letters such as S, M, L, XL. If not found, answer "not specified" 11. The value for "compatibility by model" is an electronic model that is compatible with the product. Do not start with "For". If not found, answer "not specified"	**Product Name:** *{{title}}* **Highlights:** *{{highlights}}* **Description:** *{{description}}* Extract these following attributes. • *{{target_attribute_with_value_examples_1}}* • *{{...}}* • *{{such as 'Color: e.g., Blue, Gold, Green, Black, Purple'}}* • *{{...}}* • *{{target_attribute_with_value_examples_n}}* Please follow these instructions: 1. Extract values in *{{n}}* bullet points, one attribute each. 2. All answers must strictly be from the "Product Information", NOT from examples provided after each attribute. 3. Answer "not specified" if a value can not be found. 4. Answer the attributes in this format "- attribute: [value, value, ...]". Answer the extracted values in a list. 5. Do not provide explanations or parentheses. 6. Do not answer nonsense values according to the attributes. 7. The answers have to be meaningful according to the attribute name. 8. Do not answer marketing words, such as high quality, best quality, etc. 9. Be concise and do not answer nonsense values. 10. Answers must not contradict to other values, such as "used" and "new" or "yes" and "no". 11. Separate values properly when they are provided with slash or any separators.	**Product Information:** - Category: *{{category_path}}* - Product Title: *{{title}}* - Highlights: *{{highlights}}* - Description: *{{description}}* **SKU attributes:** - *{{sku_attribute_kv_1}}* - *{{...}}* - *{{sku_attribute_kv_n}}* **Product Specifications:** - *{{specification_kv_1}}* - *{{...}}}* - *{{specification_kv_k}}* There are 2 tasks: 1. Extract these values in SKU-level: - variation - color - size - quantity - compatibility by model 2. Extract these values in product-level: - *{{target_attribute_1}}* - *{{...}}* - *{{target_attribute_m}}*
	GPT-3.5 & GPT-4			LLMs
Response template	The brand name is *{{brand_name}}*.	- variation: *{{value}}* - color: *{{value}}* - size: *{{value}}* - quantity: *{{value}}* - compatibility by model: *{{value}}*	- *{{target_item_attribute_kv_1}}* - *{{...}}* - *{{target_item_attribute_kv_n}}*	1. SKU-level values: - *{{target_sku_attribute_kv_1}}* - *{{...}}* - *{{target_sku_attribute_kv_n}}* 2. Product-level values: - *{{target_item_attribute_kv_1}}* - *{{...}}* - *{{target_item_attribute_kv_m}}*

Figure 3: Structured templates for extracting product information, including brand, SKU-level attributes, item-level attributes, and all relevant details. The first three prompts are designed for dataset development using GPT models, while the last prompt is intended for training and inference with internal models.

product variants, whereas item-level attributes are derived from the product specifications. All SKUs associated with a particular item share identical title, highlights, description, and specifications.

It is imperative to recognize that the attributes supplied by sellers may be the least reliable source of information. This unreliability stems from potential inaccuracies, such as incorrect categorization of values or arbitrary selections from platform-provided dropdown menus. In instances where an attribute value contradicts information presented in the title, highlights, or description, there exists a considerable probability that the attribute value is erroneous. Consequently, this study does not treat seller-provided attributes as definitive ground truth; rather, these attributes are meticulously revised for accuracy and subsequently used as golden labels.

4.2 Large Language Models

GPT-3.5 and GPT-4, developed by OpenAI, are advanced large language models that employ deep learning to generate human-like text. They demonstrate exceptional performance in zero-shot and few-shot contexts on datasets like OA-Mine and AE-110k (Brinkmann et al., 2024b). Following previous research, we utilize these models to extract and verify product attributes, which are then combined with seller-provided data to create silver datasets for training and evaluation.

To assess the performance of various LLMs, we selected multilingual models proficient in languages including English, Thai, Indonesian, and Vietnamese. SeaLLMs 3 (Zhang et al., 2024) from Alibaba's DAMO Academy excels in Southeast Asian languages. Qwen2.5 (Yang et al., 2024), developed by Alibaba Cloud, offers decoder-only models ranging from 0.5 to 72 billion parameters with capabilities in natural language understanding, coding, and mathematics. Gemma 2 (Team et al., 2024) from Google DeepMind includes lightweight models with 2 to 27 billion parameters, utilizing architectural innovations and knowledge distillation. Llama 3.1 (Dubey et al., 2024) from Meta AI features multilingual models competitive with leading closed-source variants, excelling in coding, reasoning, and mathematics.

Licensing is pivotal for compliance and intellectual property respect. OpenAI's terms for GPT-3.5 and GPT-4 restrict modifications but allow usage for specific tasks. Our work involves generating a small-sized model that does not directly compete with OpenAI, aligning with their guidelines. Other models have varying licenses: SeaLLMs 3 permits modification under a worldwide, non-exclusive, non-transferable agreement; Gemma 2 allows reproduction and modification within certain limits; Qwen2.5 operates under the permissive Apache License 2.0; and Llama 3.1's Community License

Agreement permits modifications with specified conditions. These licenses provide us the flexibility to innovate while ensuring compliance with each organization's guidelines.

4.3 Evaluation Metrics

We evaluate our models based on Precision (*P*), Recall (*R*), and the F1 score (*F1*), consistent with prior research (Brinkmann et al., 2024b; Yang et al., 2021). In addition, we calculate accuracy (*Acc*) and coverage (*Cov*) based on the first predicted value for each attribute.

Our predictions are classified into five distinct categories: no prediction when there is no attribute (*NN*), incorrect prediction where no attribute exists (*NV*), no prediction despite the existence of an attribute (*VN*), correct prediction that matches the attribute (*VC*), and incorrect prediction that does not align with the attribute (*VW*). The subscripted numbers (e.g., VC_1, NV_1, etc.) denote the counts of correct and incorrect predictions associated with the first predicted value. The evaluation metrics are computed as follows:

$$\begin{aligned} P &= VC\,/\,(NV{+}VC{+}VW) \\ R &= VC\,/\,(VN{+}VC{+}VW) \\ F1 &= 2PR\,/\,(P{+}R) \\ Acc &= VC_1\,/\,(NV_1{+}VC_1{+}VW_1) \\ Cov &= (NV_1{+}VC_1{+}VW_1)\,/\,All_1 \end{aligned}$$

Importantly, we follow standard practice by assessing accuracy solely on the attributes provided by sellers, which allows us to focus our evaluation on the correctness of the available information without penalizing for any missing attributes. This practice is also applied to outputs generated by LLMs. In contrast, coverage accounts for both available and missing information, providing a comprehensive view of the model's performance.

This evaluation framework enables us to assess model performance through metrics such as precision, recall, and F1 score, while also providing insights into attribute quality via accuracy and coverage.

5 Results

This section presents performance metrics and evaluations of four selected LLMs across three datasets. We benchmark the models against the seller attribute values and assess their effectiveness in predicting multiple attributes, the quality of the first predicted value, and inference speed.

5.1 Human Annotation

To assess the quality of Seller Attribute Values (SAV) and Mined Attribute Values (MAV), we conducted a human annotation evaluation focusing on SKU-level and item-level attributes across three distinct test datasets.

The results, presented in Table 2, show acceptance rates for SAV and MAV, denoted as %ASAV and %AMAV. While SAV exhibits higher acceptance rates, its volume is considerably lower than that of MAV, with accepted SAV totaling 7,710 compared to 21,742 for MAV. This gap highlights the significant potential of MAV in identifying new attribute values that sellers may overlook, as approximately 75% of accepted attribute values arise from our mining pipeline, even though they are evaluated as less accurate. These findings underscore the complementary functions of SAV and MAV; SAV provides reliable attributes, while MAV enriches the dataset by introducing three times more newly identified values.

Dataset	**SAV**	**MAV**	**%ASAV**	**%AMAV**
		SKU-level attributes		
ID-test	1221	512	48.16	99.22
TH-test	885	583	58.53	25.73
VN-test	735	608	35.92	74.18
		Item-level attributes		
ID-test	3170	6062	96.97	86.69
TH-test	1871	9473	81.88	77.56
VN-test	1950	12443	88.92	64.54
Total	**9832**	**29681**	**78.42**	**73.25**

Table 2: Results of human annotation for Seller Attribute Values (SAV) and Mined Attribute Values (MAV), categorized into SKU-level and item-level attributes. Acceptance rates (%ASAV and %AMAV) indicate the proportion of attribute values recognized as accurate by human annotators for both SAV and MAV.

5.2 Multiple Attribute Values Prediction

To evaluate the models' efficacy in predicting multiple attribute values, we focus on precision, recall, and F1 score as key performance metrics. These metrics serve as indicators of the models' capabilities to accurately identify and extract valid attribute values from product listings.

As illustrated in Table 3, the models exhibit varying performance across distinct datasets. For Lzd-ID, Llama 3.1 emerges as the top performer, achieving the highest F1 score of 77.36, which indicates

Dataset	Model	P	R	F1	Acc	Cov	Speed (s)
Lzd-ID	*Seller attribute values*	**84.92**	74.16	**79.18**	85.09	69.40	-
	Gemma 2	76.29	76.19	76.24	90.61	78.44	5587
	Llama 3.1	77.34	**77.38**	77.36	**91.30**	78.51	**642**
	Qwen2.5	74.33	74.49	74.41	90.34	78.88	859
	SeaLLMs 3	77.68	76.87	77.27	91.27	**78.96**	832
Lzd-TH	*Seller attribute values*	**67.26**	44.54	53.59	67.63	50.86	-
	Gemma 2	50.18	56.24	53.04	**77.42**	77.09	8287
	Llama 3.1	52.81	56.00	54.36	73.52	**81.21**	**1226**
	Qwen2.5	51.38	52.47	51.92	75.54	76.77	1914
	SeaLLMs 3	55.27	**57.20**	**56.22**	74.64	79.16	2098
Lzd-VN	*Seller attribute values*	**72.26**	45.07	55.52	72.17	44.00	-
	Gemma 2	67.77	**69.57**	**68.66**	**78.04**	72.31	6251
	Llama 3.1	64.45	67.61	66.00	75.85	**73.95**	**716**
	Qwen2.5	65.33	67.64	66.47	75.93	73.38	1017
	SeaLLMs 3	65.66	67.38	66.51	77.09	72.36	1042

Table 3: Performance of four LLMs fine-tuned and evaluated on three datasets, compared against seller-provided attribute values.

its effectiveness in this specific context. Conversely, SeaLLMs 3 demonstrates superior performance on Lzd-TH, suggesting a heightened suitability for processing Thai-language attributes. Meanwhile, Gemma 2 excels in the Lzd-VN dataset, highlighting the necessity of aligning model selection with the unique linguistic and contextual features inherent in each dataset.

The models consistently outperformed seller-provided attributes in terms of recall, highlighting the potential for generative approaches. Although these models may identify a broader range of potential attribute values, they often fall short of the precision achieved by sellers for attributes that encompass multiple values. This observation underscores the fundamental trade-offs between precision and recall in automated prediction systems.

5.3 First Predicted Attribute Value

In our evaluation, we also examine the accuracy and coverage of the first predicted attribute value for each attribute. This perspective is crucial for assessing how effectively the models retrieve the most relevant attribute value when multiple options are available.

The results shown in Table 3 reveal significant variability in the accuracy of the first predicted value across different models and datasets. Notably, Llama 3.1 achieves the highest accuracy of 91.30% alongside a commendable coverage of 78.51% on Lzd-ID. This model consistently demonstrates strong coverage across Lzd-TH and Lzd-VN. Conversely, Gemma 2 excels in terms of accuracy on Lzd-TH and Lzd-VN, underscoring its effectiveness in these contexts.

Overall, our findings indicate that all models enhance the attribute values provided by sellers, reflecting improvements in both accuracy and coverage. This highlights the potential of utilizing automated models to complement seller-supplied data, thereby enriching the attribute extraction process across various datasets.

5.4 Inference Speed Analysis

Inference speed is a crucial consideration for deploying model solutions in real-world scenarios. In this analysis, we measure the inference time for each model while processing data from 1,000 SKUs under specific testing conditions. The inference was conducted using a batch size of 2 and a single PPU 810 card on Alibaba Cloud. The evaluation was performed using the following parameters: `temperature=0.2`, `top_p=0.1`, and `top_k=100`.

The results reveal considerable variation across models, as indicated in Table 3. Gemma 2 exhibits the longest inference time, whereas Llama 3.1 demonstrates significantly faster processing capabilities. The prolonged inference time associated with Gemma 2 may be attributed to the incompatibility between Flash Attention 2 (Dao, 2024) and Gemma 2, resulting in a marked decrease in processing efficiency.

5.5 Comparative Analysis of Models

The performance of the four selected LLMs varies significantly across evaluated datasets, presenting an opportunity to analyze their strengths and weaknesses in real-world applications. Llama 3.1 stands out on Lzd-ID, achieving an F1 score of 77.36 and an impressive accuracy of 91.30% for the first predicted value. This consistent performance indicates its suitability for applications demanding precision in multi-attribute extraction. In contrast, SeaLLMs 3 excels in Lzd-TH, highlighting the importance of language and locale.

Gemma 2 displays high accuracy in specific contexts but struggles with inference speed, making it less viable for real-time applications. Conversely, Llama 3.1 maintains swift processing times without sacrificing accuracy, making it an optimal choice for environments requiring rapid decision-making. Qwen2.5 delivers moderate performance across datasets but lacks standout features, suggesting its suitability for general applications.

Considering the trade-offs in accuracy, coverage, and speed, Llama 3.1 is the most balanced model for deployment. Its combination of high accuracy, solid coverage, and efficient processing makes it ideal for commercial applications that require reliable attribute extraction and the ability to handle large data volumes swiftly.

5.6 Online Performance

An A/B experiment was conducted to evaluate the impact of Llama 3.1 on online performance. Orders per item page view, also known as conversion rates (CVR), showed a 0.70% increase for ID, a 0.68% decrease for VN, and a 1.19% increase for TH. Additionally, orders per page view improved by 0.47% in ID and 1.40% in TH, with VN experiencing a decrease of 0.79%. Gross Merchandise Value (GMV) per page view saw substantial increases, with VN leading at 6.73%, followed by ID at 1.61% and TH at 1.44%. These findings underscore the ability of LLMs to enhance user engagement and optimize business outcomes, thereby contributing to overall revenue growth.

6 Conclusion

In this study, we introduced an innovative approach for attribute-value extraction by leveraging generative LLMs on augmented datasets. Our method capitalizes on the zero-shot capabilities of advanced LLMs, facilitating the extraction of over 1,000 unique attributes across diverse categories with enhanced accuracy and speed. The empirical results demonstrate significant improvements in the quality of attributes provided by sellers, with notable increases in accuracy, coverage, and overall market performance metrics. By fine-tuning smaller models, we not only reduced operational costs but also enhanced efficiency, allowing for rapid inference while maintaining high prediction quality. The successful outcomes from our experiments underscore the viability of our GAVEL pipeline for wide-scale implementation in multilingual e-commerce platforms. This research paves the way for further exploration of generative approaches to attribute extraction, offering organizations valuable insights into optimizing their inventory and enhancing customer experiences.

7 Limitations

Despite the promising results of this study, several limitations should be noted. Firstly, while our augmented datasets cover a diverse range of attributes, performance may vary significantly across different product categories and languages, limiting the generalizability of our findings, particularly in regions underrepresented in the training data. Future research should aim to enhance model robustness across a broader spectrum of inputs. Additionally, our approach does not currently incorporate visual data, which is vital in e-commerce. The lack of image data may hinder comprehensive attribute extraction, especially in categories where visual representation is critical. Integrating multimodal data in future studies could enhance extraction accuracy. Another important limitation is the potential generation of erroneous data through LLM augmentation, which could result in misleading product attributes, damaging sellers' reputations and causing customer dissatisfaction. Implementing strategies for validation and verification of generated data is essential to mitigate these risks. Addressing these limitations will enable further refinement of attribute-value extraction models, enhancing their applicability in the e-commerce sector.

References

Qingyao Ai, Daniel N. Hill, S. V. N. Vishwanathan, and W. Bruce Croft. 2019. A zero attention model for personalized product search. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.*

Alexander Brinkmann, Nick Baumann, and Christian Bizer. 2024a. *Using LLMs for the Extraction and Normalization of Product Attribute Values*, page 217–230. Springer Nature Switzerland.

Alexander Brinkmann, Roee Shraga, and Christian Bizer. 2024b. Extractgpt: Exploring the potential of large language models for product attribute value extraction. *Preprint*, arXiv:2310.12537.

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.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Zhongfen Deng, Wei-Te Chen, L. Chen, and Philip S. Yu. 2022. Ae-smnsmlc: Multi-label classification with semantic matching and negative label sampling for product attribute value extraction. *2022 IEEE International Conference on Big Data (Big Data)*, pages 1816–1821.

Zhongfen Deng, Hao Peng, Tao Zhang, Shuaiqi Liu, Wenting Zhao, Yibo Wang, and Philip S. Yu. 2023. Jpave: A generation and classification-based model for joint product attribute prediction and value extraction. *2023 IEEE International Conference on Big Data (BigData)*, pages 1087–1094.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Varun Embar, Andrey Kan, Bunyamin Sisman, Christos Faloutsos, and Lise Getoor. 2021. Diffxtract: Joint discriminative product attribute-value extraction. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 271–280.

Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. *Preprint*, arXiv:2403.00863.

Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. 2023. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Trans. Recomm. Syst.*, 1(1).

Jiaying Gong and Hoda Eldardiry. 2024. Multi-label zero-shot product attribute-value extraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2259–2270, New York, NY, USA. Association for Computing Machinery.

Khalilah Abd Hafiz and Khairul Anuar Mohd Ali. 2019. The influence of product attributes on young consumers' purchase decision of makeups among malaysian: The mediating effects of perceived brand image, ceo's image, and quality.

Yudi Helfi, Fatihatul Akbar, Dinda Mutiara Pratiwi, and Fakhri Mujahid Maolani. 2019. How product attributes affect consumer decision to purchase a premium scooter matic? *JEMA: Jurnal Ilmiah Bidang Akuntansi dan Manajemen*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chen Luo, William Headden, Neela Avudaiappan, Haoming Jiang, Tianyu Cao, Qingyu Yin, Yifan Gao, Zheng Li, Rahul Goutam, Haiyang Zhang, and Bing Yin. 2022. Query attribute recommendation at amazon search. In *RecSys 2022*.

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, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kalyani Roy, Pawan Goyal, and Manish Pandey. 2021. Attribute value generation from product title using language models. In *Proceedings of the 4th Workshop on e-Commerce and NLP*, pages 13–17, Online. Association for Computational Linguistics.

Kalyani Roy, Tapas Nayak, and Pawan Goyal. 2022. Exploring generative models for joint attribute value extraction from product titles. *ArXiv*, abs/2208.07130.

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.

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.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit K. Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jonathan L. Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022. SMARTAVE: Structured multimodal transformer for product attribute value extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 263–276, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, 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 K. Sanghai, Bin Shu, Jonathan L. Elsas, and Bhargav Kanagal. 2021. Mave: A product dataset for multi-source attribute value extraction. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *Preprint*, arXiv:2407.19672.

Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. 2022. Oa-mine: Open-world attribute mining for e-commerce products with weak supervision. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3153–3161, New York, NY, USA. Association for Computing Machinery.

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.

Qiujie Zheng, Junhong Chen, Robin Zhang, and H. Holly Wang. 2020. What factors affect chinese consumers' online grocery shopping? product attributes, e-vendor characteristics and consumer perceptions. *China Agricultural Economic Review*, 12:193–213.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for E-commerce product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2129–2139, Online. Association for Computational Linguistics.

Henry Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip Yu, and Cornelia Caragea. 2024a. ImplicitAVE: An open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 338–354, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Henry Zou, Gavin Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024b. EIVEN: Efficient implicit attribute value extraction using multimodal LLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 453–463, Mexico City, Mexico. Association for Computational Linguistics.