LLM-Ensemble: Optimal Large Language Model Ensemble Method for E-commerce Product Attribute Value Extraction

Chenhao Fang*†
University of Wisconsin-Madison
Wisconsin, USA
chenhao.fang@outlook.com

Xiaohan Li*
Walmart Global Tech
Sunnyvale, California, USA
xiaohan.li@walmart.com

Zezhong Fan
Walmart Global Tech
Sunnyvale, California, USA
zezhong.fan@walmart.com

Jianpeng Xu
Walmart Global Tech
Sunnyvale, California, USA
jianpeng.xu@walmart.com

Kaushiki Nag
Walmart Global Tech
Sunnyvale, California, USA
kaushiki.nag@walmart.com

Evren Korpeoglu
Walmart Global Tech
Sunnyvale, California, USA
ekorpeoglu@walmart.com

Sushant Kumar
Walmart Global Tech
Sunnyvale, California, USA
sushant.kumar@walmart.com

Kannan Achan
Walmart Global Tech
Sunnyvale, California, USA
kannan.achan@walmart.com

ABSTRACT

Product attribute value extraction is a pivotal component in Natural Language Processing (NLP) and the contemporary e-commerce industry. The provision of precise product attribute values is fundamental in ensuring high-quality recommendations and enhancing customer satisfaction. The recently emerging Large Language Models (LLMs) have demonstrated state-of-the-art performance in numerous attribute extraction tasks, without the need for domain-specific training data. Nevertheless, varying strengths and weaknesses are exhibited by different LLMs due to the diversity in data, architectures, and hyperparameters. This variation makes them complementary to each other, with no single LLM dominating all others. Considering the diverse strengths and weaknesses of LLMs, it becomes necessary to develop an ensemble method that leverages their complementary potentials.

In this paper, we propose a novel algorithm called LLM-ensemble to ensemble different LLMs' outputs for attribute value extraction. We iteratively learn the weights for different LLMs to aggregate the labels with weights to predict the final attribute value. Not only can our proposed method be proven theoretically optimal, but it also ensures efficient computation, fast convergence, and safe deployment. We have also conducted extensive experiments with various state-of-the-art LLMs on Walmart's internal data. Our offline metrics demonstrate that the LLM-ensemble method outperforms all the state-of-the-art single LLMs on Walmart's internal dataset. This method has been launched in several production models, leading to improved Gross Merchandise Volume (GMV), Click-Through Rate (CTR), Conversion Rate (CVR), and Add-to-Cart Rate (ATC).

*Both authors contributed equally to this research.
†Work done while at Walmart.

SIGIR '24, July 14–18, 2024, Washington, DC, USA.

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3661357

CCS CONCEPTS

• **Applied computing** → **E-commerce infrastructure**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Attribute Value Extraction, Large Language Models, E-commerce

ACM Reference Format:
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. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3661357

1 INTRODUCTION

With the development of Natural Language Processing (NLP) techniques and their applications in the e-commerce industry, the extraction of accurate product attribute values with NLP plays a critical role [10, 18, 21]. The quality and relevance of product recommendations, crucial to enhancing customer satisfaction, are heavily reliant on the precision of these attributes. However, a significant challenge faced by e-commerce platforms is the lack of access to precise attribute data, leading to less accurate recommendations. Existing methods for attribute extraction are evaluated on the high-quality datasets from other platforms [30, 31, 36]; however, these methods often falter when applied to Walmart's internal datasets, resulting in less accurate extractions.

Recent advancements in the field of NLP have seen the emergence of Large Language Models (LLMs), which have shown exceptional performance in a variety of NLP tasks, including attribute value exaction, notably without the necessity for domain-specific training data. These models, including Llama [23], GPT [4], and PaLM [7], have revolutionized the way attribute value extraction

is approached, offering a new level of efficiency and accuracy. Brinkmann et al. [3] have also demonstrated using LLMs can also significantly improve the accuracy of the product attribute value extraction and achieve state-of-the-art performance. Despite their effectiveness, these LLMs exhibit distinct strengths and weaknesses due to differences in their underlying data sources, architectural designs, and hyperparameters. This diversity results in a scenario where no single LLM is universally superior across all tasks.

In light of these variations, it becomes essential to explore ensemble methods that leverage the complementary strengths of different LLMs. Ensemble methods in machine learning [9] can enhance the robustness, generalization, and accuracy of attribute extraction by aggregating the unique contributions of each model. They are particularly effective in mitigating biases, errors, and uncertainties inherent in individual models, thereby aligning the results more closely with human judgment and preferences. At Walmart, we offer hundreds of millions of items spanning numerous categories. Due to variations in the training datasets, different LLMs demonstrate varying levels of performance across these item categories. Therefore, experimenting with different LLMs for each category is not only costly but also time-intensive. This necessitates the development of an efficient strategy to seamlessly integrate multiple LLMs, enhancing overall effectiveness and efficiency.

Recently, there have been some papers on LLM fusion, as highlighted by works in [13, 25]. These studies primarily aim to enhance text generation capabilities in areas such as code generation and reasoning. However, when it comes to extracting product attribute values, the requirements are notably different. In this context, LLMs are tasked with generating concise outputs, such as a single word or a short phrase, representing the attribute values of products. To address this unique challenge, we draw inspiration from crowdsourcing techniques, treating each LLM as an individual worker. Through a process of voting and iterative refinement of predictions, our algorithm assigns weights to each LLM, calibrating their influence based on their demonstrated accuracy in specific tasks.

This paper introduces a novel algorithm called LLM-ensemble designed to ensemble the outputs of various LLMs for the purpose of attribute extraction. At its core, our approach is based on the Dawid-Skene Model [8], a structured latent variable model, to iteratively learn and assign weights to different LLM outputs. Our method is not only theoretically optimal but also boasts efficient computation, and rapid convergence, and ensures safe deployment. We validate our approach through extensive experimentation with leading LLMs on Walmart's internally labeled data. The results from these experiments clearly demonstrate that our LLM-ensemble method surpasses the performance of any single state-of-the-art LLM on Walmart's dataset.

Furthermore, we have successfully deployed this method, generating millions of highly accurate "age" and "gender" attribute labels for items in Walmart. The deployment of this method in various production models has yielded tangible benefits, including significant improvements in Gross Merchandise Volume (GMV), Click-Through Rate (CTR), Conversion Rate (CVR), and Add-to-Cart rate (ATC). The integration of this algorithm into Walmart's e-commerce platform marks a significant advancement in the field of NLP and e-commerce, setting a new standard for attribute extraction and recommendation quality.

2 RELATED WORKS

2.1 Product Attribute Value Extraction

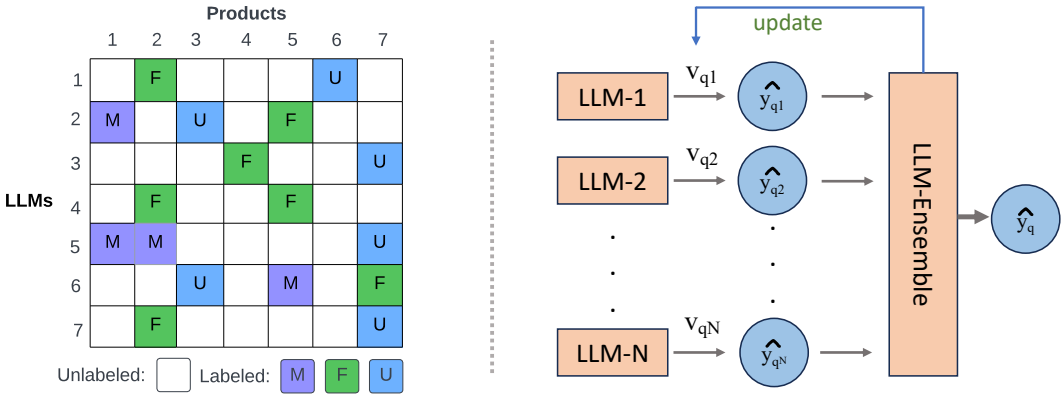
Early research on attribute value extraction relied on domain-specific rules to identify attribute-value pairs from product descriptions, as indicated in [24, 34]. However, the initial learning-based approaches required substantial feature engineering and struggled to adapt to previously unseen attributes and values [10, 20, 29]. Recently, some studies [15, 36] have shifted towards employing BiLSTM-CRF architectures for tagging attribute values in product titles. OpenTag [36] leverages a BiLSTM-CRF model enhanced by active learning. Further innovations include SU-OpenTag [30], which extends OpenTag by incorporating both a target attribute and the product title into a pre-trained language model. AdaTag [31] introduces a combination of a language model and a mixture-of-experts module for attribute value extraction, while TXtract [14] integrates a product taxonomy into its model. Approaches like AVEQA [26] and MAVEQA [33] frame attribute value extraction as a question-answering task, utilizing diverse pre-trained language models to process the target attribute, product category, and title. OA-Mine [35] explores the mining of unknown attribute values and attributes using a language model. More recent research has explored soft prompt tuning to fine-tune a minimal number of trainable parameters within a language model [2, 32]. Additionally, Brinkmann et al. [3] have demonstrated the extraction of product attribute values using Large Language Models (LLMs) like GPT-3.5 and GPT-4, highlighting the ongoing evolution in this field. Zou et al. [37, 38] propose ImplicitAVE and EIVEN to extract implicit attribute values with multimodal large language models.

2.2 Feature Extraction with LLMs

LLMs often outperform other models in zero-shot learning tasks and exhibit greater robustness when faced with unseen examples [4]. This superior performance is attributed to their extensive pre-training on vast text corpora and the emergent abilities [28] arising from their substantial model sizes. LLMs have demonstrated effectiveness across various application domains, particularly in information extraction tasks. For instance, Wang et al. [27] and Parekh et al. [19] have utilized OpenAI's LLMs for extracting structured event data from unstructured text sources. Agrawal et al. [1] have applied InstructGPT, leveraging zero-shot and few-shot learning prompts, to extract information from clinical notes efficiently. Similarly, Chen et al. [5] have used LLMs to identify relationships between products in the e-commerce sector. Maragheh et al. [17] extract keywords for products that are inferred from their textual data. Furthermore, LLMs have been employed to rank items by extracting features directly from prompts [12]. Despite these advances, current methodologies primarily focus on employing a single LLM, overlooking the potential benefits of ensemble approaches that could combine multiple LLMs to achieve enhanced results.

3 METHODOLOGY

In this section, we introduce how the proposed LLM-ensemble algorithm utilizes multiple LLMs to obtain better predictive performance on the attributes in e-commerce. Motivated from [6, 16, 22],



(a) Input data matrix W

(b) LLM-ensemble algorithm

Figure 1: (a) The input data matrix W. We take the attribute "gender" as an example, and its labels are "Male" (M), "Female" (F), and "Unisex" (U). (b) The illustration of LLM-ensemble procedures. To learn the label of a product for attribute q, we have N LLMs as inputs to the LLM-Ensemble algorithm. After several rounds of iteration, the algorithm generates the weights for each LLM and aggregates the labels with weights to predict the final label $\hat{y}_q$.

we leverage the crowdsourcing techniques to ensemble multiple LLMs to achieve enhanced results.

3.1 Problem Definition

Our goal is to extract specific attribute values from unstructured text data, such as product profiles that include titles and descriptions, based on a set of pre-defined target attributes (e.g., gender, age, style). We aim to identify attribute values that have not been previously labeled. To achieve this, we have predetermined a set of applicable attribute values for products within the domain. For example, given the inputs,

- target attributes: gender, age, and size
- product title: "Garanimals Toddler Girl Short Sleeve Graphic T-Shirt, Sizes 18M-5T"
- product description: "Bring an instant smile to her face with this colorful Graphic T-shirt from Garanimals. Cute and comfortable in a soft knit fabric $\cdots$"

Based on this unstructured text data, we want to extract 'female' (gender), 'child' (age), and 't-shirt' (style) as the corresponding values as output from LLMs. Formally, our problem is defined as

Definition 1 (Attribute Value Extraction with LLM ensemble.). *Given a set of products $\mathcal{P}$, we utilize their unstructured text data $\mathcal{T} = \{t_1, \cdots, t_p : p \in \mathcal{P}\}$ and a set of attributes $\mathcal{Q} = \{q_1, \cdots, q_m\}$ to extract the corresponding attribute-values $\mathcal{V}_p = \{v_{p,1}, \cdots, v_{p,q}\}$ for $p \in \mathcal{P}$ and $q \in \mathcal{Q}$. m is the number of the pre-defined attributes. The attribute values in $\mathcal{V}_p$ are selected from $\mathcal{L}_q$, which is the set of pre-defined labels.*

3.2 LLM Ensemble

As an example of ensemble learning, we assume that a set of LLMs is assigned to perform a labeling task extracting the specific values from the product attributes. In the following parts, we use "label" to represent the attribute values as they are selected from a limit set. Based on the Dawid-Skene Model [8], a structured latent variable model, our algorithm iteratively learns and assigns weights to different LLM outputs. We assume that there is a set of LLMs $\mathcal{N}$ and $|\mathcal{P}|$ products for the labeling task of attribute q with $|\mathcal{L}_q|$ label classes. The extended label set includes missing values represented by 0, which is defined as $\bar{\mathcal{L}}_q = \mathcal{L}_q \cup \{0\}$.

Subsequently, we take y_{qp} as the ground-truth label of the p-th product for attribute q, and $\hat{y}_{qp}$ as the predicted label for the p-th product and attribute q by an LLM. The input data matrix is denoted by $W \in \bar{\mathcal{L}}_q^{N \times P}$ where $N = |\mathcal{N}|$ and $P = |\mathcal{P}|$. W_{qij} is the label provided by i-th LLM to the j-th product for attribute q. The missing corresponding label is represented by 0, which means the i-th LLM hasn't labeled the j-th product yet. We introduce the indicator matrix $T = (T_{qij})_{N \times P}$ for attribute q, where $T_{qij} = 1$ indicates that entry (i, j) is observed, and $T_{qij} = 0$ indicates entry (i, j) is unobserved. Please note that W and T are observed together. To learn the weight v_i for the i-th LLM, we illustrate our LLM-ensemble algorithm in Algorithm 1.

Based on the findings from [16], we prove that our algorithm is theoretically optimal to ensemble multiple LLMs. Under our problem settings, the oracle Maximum A Posteriori (MAP) rule [16] approximately optimizes the upper bound on the mean error rate of weighted majority voting of LLMs. Our iterative weighted LLM-ensemble algorithms further optimize the error rate bound and approximate the oracle MAP rule. Therefore, with our LLM-ensemble method, we can assign higher weights to the "superior" LLMs, while mitigating the influence of the "spammers" (referring to LLMs whose accuracy is comparable to random guessing).

4 EXPERIMENTS

We conduct two experiments to verify the effectiveness of our method. The first is comparison experiments, in which we compare the LLM-ensemble with all single LLMs as well as traditional methods to extract attribute values. In the second experiment, we show

Algorithm 1: LLM-ensemble algorithm for product attribute-value extraction of attribute q.

Input: Number of LLMs = N; Number of products= P; Attribute p; input data matrix: $W \in \bar{\mathcal{L}}_q^{N\times P}$
Output: The predicted attribute values for attribute q: $\{\hat{y}_{q1}, \cdots, \hat{y}_{qP}\}$
Initialization: $v_{qi} = 1, \forall i \in \mathcal{N}; T_{qij} = I(W_{qij} \neq 0), \forall i \in \mathcal{N}, \forall j \in \mathcal{P}$ // I is the indicator matrix

/* loop the following steps to learn weights for LLMs */

while *not converges* ***or*** *reaches maximum iterations* **do**
  $\hat{y}_{qj} \leftarrow \text{argmax} \sum_{i=1}^{N} v_{qi} I(W_{qij} = k), \forall j \in \mathcal{P}$
  $\hat{\alpha}_{qi} \leftarrow \frac{\sum_{j=1}^{P} I(W_{qij}=\hat{y}_{qj})}{\sum_{j=1}^{P} T_{qij}}, \forall i \in \mathcal{N}$
  $v_{qi} \leftarrow L\hat{\alpha}_{qi} - 1, \forall i \in \mathcal{N}$
return the predictions $\{\hat{y}_{qj}\}_{j\in\mathcal{P}}$ by $\{\hat{y}_{qj}\} = \text{argmax}_{k\in\mathcal{L}_q} \sum_{i=1}^{N} v_{qi} I(W_{ij} = k)$

the results of the A/B test on the similar item recommendation model in Walmart.

4.1 Comparison Experiments

Our proposed LLM-ensemble is compared with its base LLMs: Llama2-13B, Llama2-70B [23], PaLM-2 [7], GPT-3.5, GPT-4 [4]. We also include logistic regression [11] and our internal rule-based method in the baseline models. The datasets we employ are Walmart-Age and Walmart-Gender, which contain products sensitive to the ages or genders of customers[1]. Each dataset has 20K items for offline evaluation. The ground-truth label is created by the crowdsourcing results.

The experiment results are shown in Table 1. From this table, we can find that our proposed LLM-ensemble achieves the best performance compared to all other baseline models, which demonstrates the effectiveness of the LLM ensemble algorithm. Moreover, the logistic regression and rule-based method are much worse than LLM-based model, which means LLMs' emergent abilities can dramatically improve the accuracy of the product attribute value extraction.

4.2 A/B test on Walmart Recommendation Model

Having demonstrated state-of-the-art performance in our offline experiments, we've chosen to advance to an online A/B test. Within these trials, we apply the age and gender labels as top-layer filtration on the similar item recommendation model, which is one of the recommendation models in Walmart, to filter out misclassified products. By employing our method, the recommendation model can enhance user engagement and increase conversion rates by ensuring the relevance of recommendation results. We implement this top-layer filtration across over 200 product categories on Walmart's e-commerce platform.

Table 2 shows the outcomes of the online experiment, demonstrating a statistically significant enhancement across various key metrics in e-commerce, including Gross merchandise volume (GMV), Click-Through Rate (CTR), Conversion Rate (CVR) and Add-to-Cart Rate (ATC). These findings strongly support the effectiveness of the LLM-ensemble methodology, emphasizing its importance and potential impact on e-commerce applications. As a result of the success of the A/B test, we have now launched this feature on Walmart's online platform.

[1]Due to the privacy policy in Walmart, we only disclose these two attributes. The actual product in practice has more attributes involved.

Models	Walmart-Age	Walmart-Gender
Logistic regression	0.653	0.681
Rule-based method	0.710	0.759
Llama2-13B	0.753	0.798
Llama2-70B	0.887	0.910
PaLM-2	0.875	0.894
GPT-3.5	0.911	0.933
GPT-4	0.934	0.952
LLM-ensemble	**0.956**	**0.979**
Improvement	2.36%	2.76%

Table 1: Comparison experiments of different models on the prediction accuracy. The underlined model is the second-best one and the bold model is the best of all models.

Table 2: Online experiment results

Metrics	Percentage Lift	P-Value
GMV	0.38%	0.039
CTR	2.16%	0.028
CVR	0.26%	0.043
ATC	1.42%	0.036

5 CONCLUSION

this paper introduces an innovative ensemble method, LLM-ensemble, designed to optimize product attribute value extraction by ensembling the outputs of various Large Language Models (LLMs). By dynamically learning the weights for different LLMs and aggregating their labels, our algorithm not only achieves theoretical optimality but also excels in efficiency, convergence speed, and deployment safety. Through comparison experiments with state-of-the-art LLMs on Walmart's internal dataset, the LLM-ensemble method has demonstrated superior performance over all individual LLMs. The A/B test of this method in production models has also improved the performance of our recommendation product.

6 PRESENTER BIO

Xiaohan Li is a Senior Data Scientist at the personalization team of Walmart Global Tech. He received his Ph.D. in Computer Science from the University of Illinois at Chicago (UIC) and B.Eng in Computer Science from Beijing University of Posts and Telecommunications (BUPT). His research interests are recommender systems, large language models, diffusion models, and graph neural networks.

REFERENCES

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 1998–2022.

[2] Ansel Blume, Nasser Zalmout, Heng Ji, and Xian Li. 2023. Generative Models for Product Attribute Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 575–585.

[3] Alexander Brinkmann, Roee Shraga, and Christian Bizer. 2023. Product Attribute Value Extraction using Large Language Models. *arXiv preprint arXiv:2310.12537* (2023).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason HD Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs. *arXiv preprint arXiv:2305.09858* (2023).

[6] Ziqi Chen, Liangxiao Jiang, and Chaoqun Li. 2022. Label augmented and weighted majority voting for crowdsourcing. *Information Sciences* 606 (2022), 397–409.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[8] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[9] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

[10] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. 2006. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter* 8, 1 (2006), 41–48.

[11] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.

[12] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845* (2023).

[13] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. *arXiv preprint arXiv:2306.02561* (2023).

[14] Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. *arXiv preprint arXiv:2004.13852* (2020).

[15] Zornitsa Kozareva, Qi Li, Ke Zhai, and Weiwei Guo. 2016. Recognizing salient entities in shopping queries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 107–111.

[16] Hongwei Li and Bin Yu. 2014. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086* (2014).

[17] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, et al. 2023. LLM-TAKE: Theme-Aware Keyword Extraction Using Large Language Models. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 4318–4324.

[18] Ajinkya More. 2016. Attribute extraction from product titles in ecommerce. *arXiv preprint arXiv:1608.04670* (2016).

[19] Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking Generalizability for Event Argument Extraction with Hundreds of Event Types and Argument Roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3664–3686.

[20] 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*. 1557–1567.

[21] Martin Rezk, Laura Alonso Alemany, Lasguido Nio, and Ted Zhang. 2019. Accurate product attribute extraction on the field. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1862–1873.

[22] Tian Tian and Jun Zhu. 2015. Max-margin majority voting for learning from crowds. *Advances in neural information processing systems* 28 (2015).

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[24] Damir Vandic, Jan-Willem Van Dam, and Flavius Frasincar. 2012. Faceted product search powered by the semantic web. *Decision Support Systems* 53, 3 (2012), 425–437.

[25] Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models. *arXiv preprint arXiv:2401.10491* (2024).

[26] 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. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 47–55.

[27] Xingyao Wang, Sha Li, and Heng Ji. 2023. Code4struct: Code generation for few-shot event structure prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3640–3663.

[28] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[29] Yuk Wah Wong, Dominic Widdows, Tom Lokovic, and Kamal Nigam. 2009. Scalable attribute-value extraction from semi-structured text. In *2009 IEEE international conference on data mining workshops*. IEEE, 302–307.

[30] Huimin Xu, Wenting Wang, Xinnian 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*. 5214–5223.

[31] Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding. *arXiv preprint arXiv:2106.02318* (2021).

[32] 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*. 9978–9991.

[33] 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*. 1256–1265.

[34] Liyi Zhang, Mingzhu Zhu, and Wei Huang. 2009. A Framework for an Ontology-based E-commerce Product Information Retrieval System. *J. Comput.* 4, 6 (2009), 436–443.

[35] 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*. 3153–3161.

[36] 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*. 1049–1058.

[37] Henry Peng Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip S. Yu, and Cornelia Caragea. 2024. ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction. *arXiv preprint arXiv:2404.15592* (2024).

[38] Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024. 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: Industry Track*.