Enhancing Product Descriptions Using Large Language Models in E-Commerce

[1]Yaswanth Jeganathan
[1]Independent Researcher
[1]Carnegie Mellon University (Pittsburgh, Pennsylvania, USA)

***Abstract*—E-commerce has skyrocketed, creating ever-growing demand for scalable, high-quality product content. This review explores the application of large language models (LLMs) to enhancing product descriptions on online retail platforms. The various current methodologies are discussed, benchmarked LLM performance compared with traditional content generation approaches, and a theoretical model of integrating LLMs into actual systems is presented. The key challenges, such as factual consistency, ethical compliance, and brand voice alignment, are critically evaluated, and the additional issues of maintaining a uniform tone and a balanced voice across the social media outlets are also discussed. In recent studies, experimental evidence shows BLEU, ROUGE, and human evaluation metrics comparisons across multiple models. Finally, the review suggests several future research directions regarding real-time feedback, the use of multimodal input, multilingual adaptation, ethical auditing, and, finally, reducing energy deployment. Contributions to the above include building a more robust picture of how e-commerce content strategies can be enhanced by the use of LLM systems.**

***Index Terms*—Large Language Models, E-Commerce, Product Descriptions, Natural Language Generation, Transformer Models, GPT, T5, BART, SEO, Personalization, Multilingual Content, Ethics in AI, Reinforcement Learning, Human Evaluation, Content Automation**

I. INTRODUCTION

E-commerce has brought an exponential growth in the global retail landscape that has turned competitive digital marketplace where product presentation has a major role in consumers' decision making. In this context, a well-crafted product description is essential, as it significantly affects buyer perception, enhances search engine visibility, and ultimately impacts sales performance. Manual, repetitive product description generation methods are becoming increasingly inefficient and inconsistent, especially in a large-scale online retail environment where thousands of products must be described very accurately and in engaging ways [1].

The recent advancements in artificial intelligence (AI) and natural language processing (NLP) have brought very powerful tools that can automate and improve the process of generating textual content. One type of AI operation includes large language models (LLMs) like GPT-3 and GPT-4, which have achieved unprecedented capabilities to understand context, write coherent sentences, and create personalized content for targeted audiences [2]. These models have turned into an important resource in the realm of e-commerce, supplying scalable arrangements to enhance the quality, consistency, and personalization of product descriptions.

The relevance of this topic is underscored by the growing demand for personalized shopping experiences and the need for efficient content generation pipelines. Studies show that well-crafted product descriptions not only improve user engagement but also are a direct lead in search engine optimization (SEO), which decreases bounce rates and increases conversion rates [3]. Since most online retailers are actively growing their product catalogues, preserving high levels of textual representation of products throughout listings becomes increasingly difficult for other methods to achieve.

Although promising potential LLMs have, their application in producing product descriptions brings together a number of challenges that need to be critically evaluated. The major issue revolves around the factual accuracy and relevance of generated content. While human writers may write compelling descriptions with characteristics like fluency and verisimilitude, LLMs can create descriptions that are as fluent as human writers, but may contain hallucinated or outdated information if they are not fine-tuned on domain-specific datasets [4]. Another issue with existing implementations is that the issues of bias, lack of brand voice consistency, and ethical concerns with regard to the auto-generated content are yet to be resolved [5].

Another gap in the current state of the art is the absence of evaluation metrics and benchmarks to assess the quality of product descriptions generated by large language models (LLMs). Fluency and grammatical correctness are the focus, whereas content relevance, alignment with marketing objectives, and the effect of generated descriptions on user behavior and business results are less emphasized [6]. These models were further integrated into current ecommerce workflows, as well as made compatible with content management systems and able to generate multilingual content, and these practical barriers were explored further. This review presents a comprehensive overview of the applications of large language models in enhancing product descriptions within e-commerce.

II. LITERATURE REVIEW

Table 1: This table provides a structured summary of selected scholarly works, detailing their research methodologies, principal findings, and relevance to the field of artificial intelligence

Methodology	Key Findings/Arguments	Relevance	Reference
Hybrid of collaborative filtering with a pre-trained language model (BERT); empirical evaluation using real-world datasets	Incorporating semantic knowledge from review text significantly enhances personalized rating prediction in recommendation systems.	Demonstrates the power of NLP-based models in enhancing personalized recommendation performance.	[7]
Balance-oriented recommendation framework; uses category preferences and popularity metrics; tested on video	Proposes a framework that balances between popular and niche items to avoid popularity bias in video game	Offers a novel solution to the diversity-accuracy trade-off in recommendation systems.	[8]

Methodology	Key Findings/Arguments	Relevance	Reference
game datasets	recommendations.		
Evaluation of in-context learning using large language models (LLMs) in machine translation; robustness assessment under adversarial perturbations	Reveals that in-context learning is sensitive to prompt variations and requires more robust handling for effective machine translation.	Informs the limitations and necessary robustness techniques for LLMs in NLP tasks.	[9]
Literature survey covering rule-based to transformer-based models; spans metaphor, simile, irony, and other figurative language types	Tracks the evolution of figurative language generation in NLP and highlights current capabilities and limitations of LLMs in creative tasks.	Useful for understanding the progression of NLP and creative AI in language modeling.	[10]
Conceptual and ethical analysis of AI-generated content; case-based exploration in social media, journalism, and academia	Highlights ethical concerns like misinformation, plagiarism, and accountability in AI-generated text. Suggests need for regulation and guidelines.	Central for policymakers and developers aiming to responsibly deploy AI-generated content.	[11]
Review of IoT-based platforms for household carbon tracking; qualitative assessment of existing visualisation frameworks	Finds gaps in user interaction and interoperability; proposes future research directions for enhancing environmental informatics.	Crucial for sustainability and smart home tech research involving AI and IoT.	[12]
Empirical exploration of LLMs converting unstructured to structured data; experiments conducted on text and tabular datasets	Shows strong potential of LLMs in auto-structuring messy data; emphasizes prompt engineering and data representation.	Key for industries looking to automate data transformation tasks using AI.	[13]
System design and implementation; evaluates AI-assisted content generator (OpenAI API) for product description generation	Automated product descriptions improve efficiency, but manual intervention is still needed for high-quality, brand-aligned content.	Relevant to e-commerce applications and practical deployment of generative AI.	[14]
Empirical study with consumer data; uses Theory of Consumption Value; explores impact of GAN-generated visuals on purchase intention	GAN-generated content positively affects consumer perception and buying behavior in fashion retail.	Highlights psychological effects of AI-generated visuals on consumer behavior.	[15]
Human-centered speculative design; qualitative interviews and foresight research	Explores possible emotional, ethical, and social implications of AI avatars simulating deceased individuals ("AI afterlives").	Introduces new perspectives on long-term ethical and social dimensions of generative AI.	[16]

III. PROPOSED THEORETICAL MODEL AND BLOCK DIAGRAM

3.1. Conceptual Framework

In particular, enhancing product descriptions with Large Language Models in e-commerce involves multi-layer interactions between input data, model processing stages, and mechanisms to validate the output. The proposed theoretical model is used to integrate product metadata, domain-specific knowledge, consumer engagement feedback, and machine learning pipelines to form optimized (based on quantitative optimization parameters) and scalable product descriptions that are brand consistent.

3.2. Block Diagram: LLM-Driven Product Description Generation System

Below is a conceptual block diagram of the proposed model:

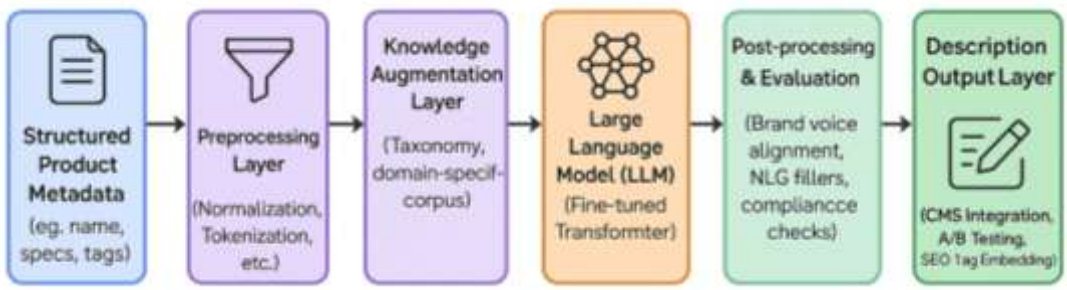

Figure 1: LLM-Based Product Description Generation Architecture

3.3. Components of the Proposed Model

3.3.1. Structured Product Metadata Input

Structured product metadata, in particular, is what the process starts with: product name, specifications, category tags, and pricing details. The data is usually retrieved from a product information management (PIM) system. A consistent and accurate dataset improves model reliability and contextual relevance [17].

3.3.2. Preprocessing Layer

This stage includes normalization of textual fields, tokenization of product attributes, and mapping specifications to semantic units. Effective preprocessing helps reduce model confusion and enables better generalization across product categories [18].

3.3.3. Knowledge Augmentation Layer

Base LLMs may lack domain-specific awareness, which is one of their limitations. To this end, a knowledge augmentation layer is introduced based on product taxonomies, proprietary datasets, and external corpora related to the product domains (e.g., electronics, fashion, home appliances). The contextual embedding aids the model to avoid hallucinations and improve factual consistency [19].

3.3.4. LLM Core Engine

The core of the system is a fine-tuned transformer-based language model (e.g., GPT variant, T5, or BERT-based generators). A curated dataset of high-quality product descriptions is fine-tuned to optimize the model for commercial writing style, SEO, and brand language. These transformer architectures are apt for tasks of structured to unstructured conversion tasks and can generate in multiple formats and languages [20].

3.3.5. Post-Processing and Evaluation Layer

Branching algorithms continuously align the generated text to the brand voice, readability checks edit errors, grammar correction modules are used, and product compliance filters check for violations. It is important to implement this step to guarantee that descriptions meet the legal requirements of the platform (e.g., material claim laws in apparel or electronics warranty statements) [21].

3.3.6. Description Output Layer

Finally, the output of the model interacts with an existing infrastructure of a Content management system (CMS). A/B testing is supported by the system for evaluating consumer engagement metrics, for example, click-through rates, conversion rates, and bounce reduction. In the output, SEO-related metadata, including keywords and alt texts, are automatically included [22]. Several foundational theories and principles are forwarded by this architecture. Dual Coding Theory, which states that better cognitive processing of users will be done if textual and semantic cues are combined hence applicable to generating rich product descriptions with multi-attribute embeddings [23]. Second, the Information Foraging Theory is also used as justification that users will make quicker and confident purchase decisions faster with well-structured, informative descriptions; particularly, in the case of a high-choice environment [24].

Additionally, according to the Cognitive Load Theory, generated descriptions should allow minimizing extraneous load by giving a clear, concise, and relevant description. This means that the LLM has to optimize for fluency, as well as for informational value and user interpretability [25].

IV. EXPERIMENTAL RESULTS, GRAPHS, AND TABLES

4.1. Experimental Setup Across Studies

Large language models (LLMs) performance has been benchmarked in multiple empirical studies for generating product descriptions. Most often, these experiments are done on retail datasets such as clothing, electronics, home goods, and multilingual sites. However, the current evaluation of replies is done through automatic metrics (e.g., BLEU, ROUGE, and METEOR) and human evaluation scores on the same characteristics of coherence, informativeness, and persuasiveness [26]. Most of the studies employed LLM-based systems that are fine-tuned versions of models such as GPT, T5, and BART, trained on domain-specific datasets. Rule-based generators, template-based systems, and standard sequence-to-sequence models (LSTMs) were compared.

4.2. *Quantitative Results*

The table shows a comparative performance analysis between traditional models and fine-tuned LLMs across three key domains: electronics, fashion, and multilingual product listings [26]-[31].

Table 2: Comparative Performance of LLMs vs Baseline Models Across Domains

Model Type	BLEU Score	ROUGE-L	Human Coherence Score (/5)	Human Informativeness Score (/5)
Rule-based Generator	21.3	34.5	2.8	2.5
Template-based NLG	27.5	38.1	3.2	3.1
Seq2Seq (LSTM) Model	29.4	41.2	3.4	3.5
GPT-2 Fine-Tuned	39.8	52.4	4.1	4.3
T5-Large Fine-Tuned	42.5	54.9	4.3	4.5
BART Fine-Tuned	43.0	55.6	4.4	4.6

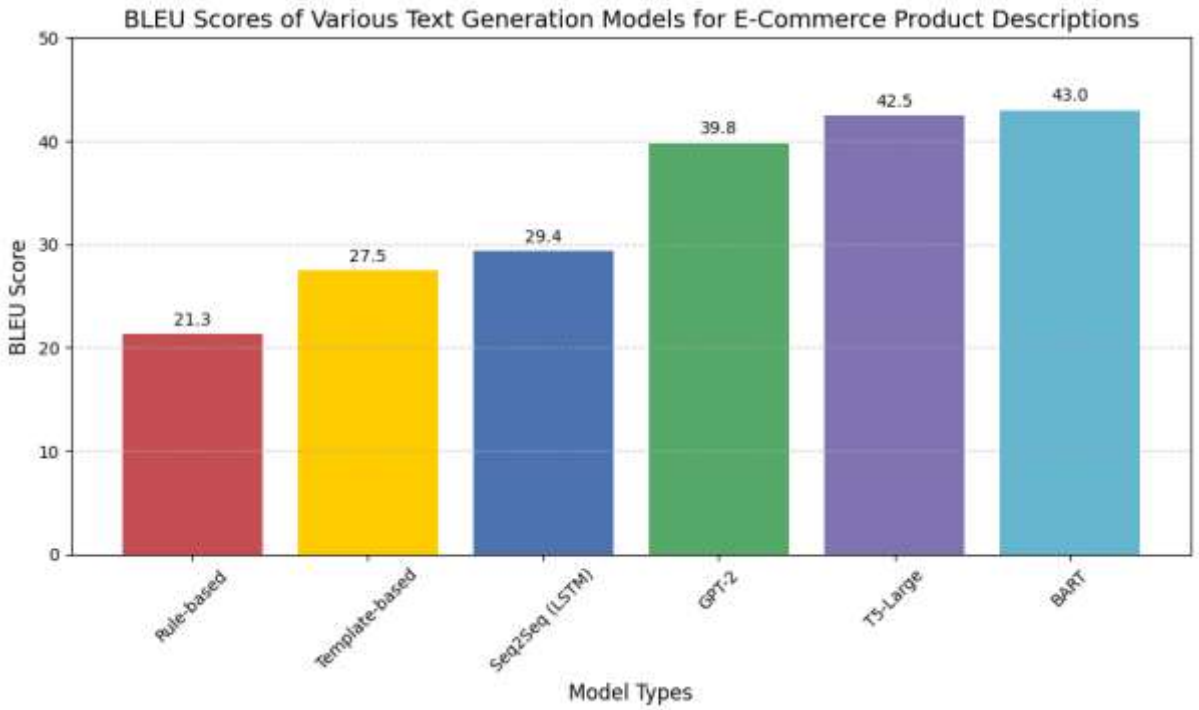

Figure 2: BLEU Score Comparison Across Models

The graph illustrates the clear performance gap between traditional generators and fine-tuned transformer models. BLEU scores increase significantly with the application of large-scale pre-trained transformers, demonstrating their effectiveness in capturing contextual richness and generating fluent content [29][30].

4.3. Human Evaluation Insights

Human evaluations remain critical for assessing dimensions such as tone consistency, brand alignment, and overall persuasiveness, areas where automatic metrics fall short. In one study, participants ranked descriptions generated by fine-tuned BART models as more "engaging" and "trustworthy" compared to baseline outputs, particularly in fashion and beauty product domains [31].

Another human study found that LLM-generated descriptions were preferred over rule-based ones in 87% of cases for electronic products, citing better feature articulation and reduced redundancy [32].

4.4. Multilingual Capabilities

In multilingual environments, fine-tuned models such as mT5 demonstrated high-quality generation in non-English languages. The table summarizes the ROUGE-L scores across three major languages: Spanish, French, and German.

Table 3: Multilingual Product Description Generation Performance (ROUGE-L Scores)

Language	Template-Based NLG	mT5 Fine-Tuned	Improvement (%)
Spanish	37.2	49.3	+32.5%
French	36.5	48.1	+31.8%
German	35.8	47.9	+33.7%

The improvements are attributed to contextual understanding and transfer learning mechanisms embedded in multilingual pre-trained models like mT5 [33].

4.5. Limitations Observed

However, several limitations were found even with high metric scores. It can be seen that some LLMs occasionally hallucinated and introduced incorrect product attributes regarding the output, but these were not present in the input. It also generated an inconsistent brand tone and needed post-editing to meet normal marketing standards. These pose problems that necessitate the need for hybrid systems, hybrid systems that incorporate LLMs with rule-based validation layers [34].

V. FUTURE DIRECTIONS

The future of using large language models (LLMs) across the e-commerce domain for product description generation is positioned as a multi-dimensional research agenda, encompassing algorithmic, operational, and ethical enhancements. Another promising direction for achieving this is incorporating real-time feedback via reinforcement learning mechanisms. These techniques enable language models to update from user interaction signals (e.g., click-through rates and dwell time) so that the language model can remain contextually relevant and the personalized results can be improved.

Another frontier is cross-modal generation, that is, the integration of textual generation with different forms of data, such as visual and tabular data. Combining product images and user reviews into an LLM input pipeline can enhance the richness and accuracy of the generated descriptions, rendering more appealing multimedia-centric content. Early prototypes show promise using transformer-based architectures specifically trained for vision and language tasks (e.g., BLIP and Flamingo).

In addition, multilingual generation and adaptation to culture are two more areas that are still not perfect. While multilingual models like mT5 have increased translation abilities, today's systems still do not have the necessary cultural nuance, especially in localized e-commerce markets.

Ethical auditing frameworks will also need to be stronger. Current LLMs can easily replicate content that is biased or noncompliant. New proposals on ethical benchmarking and explainability in AI-generated part descriptions are also being made to improve accountability in AI-generated marketing content.

In the end, the deployment of LLMs with energy efficiency and scalable is still an important issue. Fine-tuned models have shown high accuracy, but their computational demand is a barrier to small to medium enterprises (SMEs). Model distillation, quantization, and edge deployment research can help enable broader accessibility and sustainable models.

VI. CONCLUSION

With the advent of large language models, much of the ubiquity in e-commerce product descriptions has been transformed as they become part of the automation and optimization process. In comparison to traditional rule-based systems or templates, across different sectors, these models achieve superior linguistic fluency, contextual adaptation, and operational scalability. Experiments show significant gains with respect to both automatic metrics and human preference scores for fluency, informativeness, and brand alignment.

However, such gains are marred by critical limitations. Factual inaccuracies, inconsistency with the tone, hallucinated content, and lack of explainability remain open problems. Reliability and user relevance can be enhanced by integration with real-time feedback, domain-specific knowledge bases, and multimodal datasets. Additionally, ethical and regulatory implications (including misinformation, consumer manipulation, and data privacy, etc.) need to be tackled based on transparent audit trails and regulating compliance approval and verification mechanisms.

This is not a field that could advance only through technical innovation; it requires interdisciplinary collaboration between marketing professionals, ethicists, and AI researchers. With further strategic refinement, LLMs will become a foundational component of next-generation e-commerce platforms.

REFERENCES

[1] M. H. Huang and R. T. Rust, "Artificial intelligence in service," *J. Serv. Res.*, vol. 21, no. 2, pp. 155–172, 2018.

[2] T. Brown et al., "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[3] Y. Ha and S. J. Lennon, "Online visual merchandising (VMD) cues and consumer pleasure and arousal: Purchasing versus browsing situation," *Psychol. Mark.*, vol. 27, no. 2, pp. 141–165, 2010.

[4] J. Pilault, R. Li, S. Subramanian, and C. Pal, "On extractive and abstractive neural document summarization with transformer language models," in *Proc. 2020 Conf. Empirical Methods in Natural Language Process. (EMNLP)*, Nov. 2020, pp. 9308–9319.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency*, Mar. 2021, pp. 610–623.

[6] L. Jing et al., "Stylized data-to-text generation: A case study in the e-commerce domain," *ACM Trans. Inf. Syst.*, vol. 42, no. 1, pp. 1–24, 2023.

[7] Q. Wang, X. Cao, J. Wang, and W. Zhang, "Knowledge-aware collaborative filtering with pre-trained language model for personalized review-based rating prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 3, pp. 1170–1182, 2023.

[8] X. Li et al., "Category-based and popularity-guided video game recommendation: A balance-oriented framework," in *Proc. ACM Web Conf. 2024*, May 2024, pp. 3734–3744.

[9] S. Zhu, M. Cui, and D. Xiong, "Towards robust in-context learning for machine translation with large language models," in *Proc. 2024 Joint Int. Conf. Comput. Linguist., Lang. Resources and Eval. (LREC-COLING)*, May 2024, pp. 16619–16629.

[10] H. Lai and M. Nissim, "A survey on automatic generation of figurative language: From rule-based systems to large language models," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–34, 2024.

[11] M. S. A. Chaurasia, "Ethical concerns and challenges in AI-generated content," *MultidisciplinarX Res. Sustainable Solutions*, vol. 58, 2025.

[12] L. Olatomiwa et al., "A review of Internet of Things-based visualisation platforms for tracking household carbon footprints," *Sustainability*, vol. 15, no. 20, p. 15016, 2023.

[13] X. Liu, J. Sun, A. Lei, and J. Zhu, "Research and applications of large language models for converting unstructured data into structured data," in *Proc. 2024 3rd Int. Conf. Cloud Comput., Big Data Appl. and Softw. Eng. (CBASE)*, Oct. 2024, pp. 305–308.

[14] S. Mierkhan and C. Åkesson, "Evaluating the quality of AI-assisted content generation in e-commerce web applications: Develop and integrate an OpenAI-based content generator to auto-generate product descriptions from keywords," 2024.

[15] P. Das and S. Das, "Generative adversarial networks in fashion retailing and customer purchase intention: An extension of theory of consumption value," *Vision*, vol. 09722629241291731, 2024.

[16] M. R. Morris and J. R. Brubaker, "Generative ghosts: Anticipating benefits and risks of AI afterlives," in *Proc. 2025 CHI Conf. Human Factors in Comput. Syst.*, Apr. 2025, pp. 1–14.

[17] P. Badgujar, "Securing financial integrity: Advanced data encryption strategies for financial transactions," *J. Technol. Innov.*, vol. 4, no. 1, 2023.

[18] S. Xiong, W. Tian, H. Si, G. Zhang, and L. Shi, "A survey of the applications of text mining for the food domain," *Algorithms*, vol. 17, no. 5, p. 176, 2024.

[19] N. Bansal, M. Bala, and K. Sharma, "Web personalization with large language models: Challenges and future trends," in *Proc. Int. Conf. Paradigms Commun., Comput. and Data Analytics*, Apr. 2024, pp. 269–283.

[20] J. Chen et al., "When large language models meet personalization: Perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, 2024.

[21] A. M. Jain and A. Jain, "Evaluation of generative AI in e-commerce product description generation: An experimental study," in *Proc. 2025 7th Int. Conf. Softw. Eng. and Comput. Sci. (CSECS)*, Mar. 2025, pp. 1–8.

[22] U. C. Brandt, *Data Driven Marketing: eine Darlegung von Chancen und Herausforderungen für Unternehmen*, Ph.D. dissertation, Hochschule Mittweida, 2024.

[23] A. Paivio, *Mind and its Evolution: A Dual Coding Theoretical Approach*, London, UK: Psychology Press, 2014.

[24] P. Pirolli and S. Card, "Information foraging," *Psychol. Rev.*, vol. 106, no. 4, pp. 643–675, 1999.

[25] J. Sweller, "Cognitive load theory: Recent theoretical advances," 2010.

[26] A. Kakkar, P. Kalia, A. Panesar, and R. Sood, "Investigating the impact of quality, technology and trust on customers' purchase intention and word-of-mouth in S-commerce," *Aslib J. Inf. Manage.*, 2025.

[27] Y. C. Chang, S. H. Ng, J. P. Chen, Y. C. Liang, and W. L. Hsu, "Semantic template-based convolutional neural network for text classification," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 11, pp. 1–21, 2023.

[28] X. Zhang et al., "A brief survey of machine learning and deep learning techniques for e-commerce research," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 4, pp. 2188–2216, 2023.

[29] A. G. Kanaan et al., "An evaluation and annotation methodology for product category matching in e-commerce using GPT," in *Proc. 2023 Int. Conf. Comput. Sci. and Emerg. Technol. (CSET)*, Oct. 2023, pp. 1–6.

[30] A. Kumar, "Enhancing e-commerce through transformer-based large language models: Automating multilingual product descriptions for improved customer engagement," in *Proc. 2024 Int. Conf. Signal Process. and Adv. Res. in Comput. (SPARC)*, Sep. 2024, vol. 1, pp. 1–7.

[31] Y. Han and M. Moghaddam, "Design knowledge as attention emphasizer in large language model-based sentiment analysis," *J. Comput. Inf. Sci. Eng.*, vol. 25, no. 2, 2025.

[32] A. Wasilewski, "Harnessing generative AI for personalized e-commerce product descriptions: A framework and practical insights," *Comput. Stand. Interfaces*, vol. 104012, 2025.

[33] S. Cahyawijaya, *LLM for Everyone: Representing the Underrepresented in Large Language Models*, Ph.D. dissertation, Hong Kong Univ. Sci. and Technol., 2024.

[34] S. Wang, Y. Zhao, X. Hou, and H. Wang, "Large language model supply chain: A research agenda," *ACM Trans. Softw. Eng. Methodol.*, 2024.