International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Generative AI for Scalable and Explainable E-Commerce Product Title Evaluation: A Prompt-Driven Framework

Priyadarshini Balachandran

Distinguished Software Engineer
Fellow IETE, Forbes Technology Council, San Jose, California, United States - 95122
Email: *writetopriyab[at]gmail.com*

Abstract: ***Product titles are a decisive factor in e-commerce, directly impacting search visibility, customer engagement, and sales conversion. However, practices such as keyword stuffing, excessively long titles, and promotional claims often degrade discoverability and marketplace quality. This paper introduces a generative AI–based framework for automated product title evaluation that classifies titles as "Good" or "Bad" with interpretable reasoning. Unlike traditional machine learning pipelines requiring labeled datasets and feature engineering, our approach leverages large language models (LLMs) with few-shot prompting to deliver transparent, audit-ready outputs. The system architecture comprises of four modular components, Normalizer & Precheck, LLM Judge, LLM Provider Layer, and Policy Aggregator, ensuring scalability, explainability, and adaptability to policy changes. Experiments conducted on a benchmark dataset of 1,000 manually annotated product titles achieved 91% accuracy, 88% precision, and 85% recall, demonstrating strong performance in detecting violations such as keyword stuffing, category mismatch, and irrelevant descriptors. The framework reduces development overhead, supports rapid policy iteration, and provides structured feedback for sellers, thereby enhancing marketplace compliance and buyer trust. Future work will explore multilingual extensions, continuous human-in-the-loop refinement, and integration with complementary metadata such as images and seller reputation to further strengthen title quality evaluation.***

Keywords: Search Relevance Optimization**,** Content Quality Evaluation, Generative AI Evaluation, Large Language Models, AI Orchestration Evaluation

1. Introduction

In modern e-commerce, product titles serve as the primary textual signal that drives product discoverability, search relevance, and customer decision-making. A concise, accurate, and well-structured title not only improves search rankings but also builds customer trust and directly influences sales conversion. However, marketplace data indicates that many sellers adopt suboptimal practices—such as overloading titles with redundant or irrelevant keywords, inserting unverifiable claims, or combining multiple products into a single title. These practices often confuse buyers, reduce search precision, and negatively affect marketplace credibility.

Prior studies, such as Hirsch (2025) [1] and Gupta & Bansal (2023) [2], have shown that title wording significantly affects both retrieval performance and conversion outcomes. Addressing this issue requires scalable solutions that can enforce quality standards while adapting to evolving marketplace policies. Traditional machine learning approaches, while effective in certain contexts, demand extensive labeled datasets, feature engineering, and complex model pipelines that limit adaptability and transparency.

This paper proposes a generative AI–driven framework for product title evaluation. By leveraging the reasoning capabilities of instruction-tuned LLMs, the framework simplifies evaluation pipelines into a single prompt-based process. This design provides interpretable outputs, reduces development overhead, and enables real-time moderation. The remainder of this paper details the methodology, system architecture, evaluation results, and potential future directions for advancing title quality assessment in e-commerce platforms.

2. Literature Review

Product Title Optimization

Product titles are central to e-commerce search visibility and sales. Well-structured titles align with search algorithms and user intent. Hirsch et al. (2025) [1] analyzed e-commerce title syntax, order, and attributes, showing they differ from other web text and recommending concise inclusion of key attributes while avoiding redundancy or keyword stuffing. Gupta and Bansal (2023) [2] demonstrated that integrating top search terms into Amazon titles increased conversion across 2.5M customers, though excessive keywords reduced user trust and risked policy violations. These findings highlight the tension between discoverability and readability, motivating automated quality checks. Beyond wording, researchers have explored title compression for constrained UIs: multimodal and reinforcement-learning approaches shorten long titles while preserving salient attributes [12], and multimodal GANs have shown online gains in CTR/CVR for short-title generation [13]. Recent surveys also catalogue large-scale title-compression datasets used in industry (e.g., Walmart D-com-human) [15].

Machine Learning Approaches

Early methods treated title optimization as supervised learning with engineered features. Camargo de Souza et al.

(2018) [3] built a neural text generator and quality predictor that cleaned noisy titles at scale, though requiring separate models and dataset-specific tuning. Classification tasks gained traction through the CIKM AnalytiCup 2017 challenge. Singh and Sunder (2018) [4] achieved state-of-the-art results by combining CNN-LSTM semantic models with LightGBM features such as length and readability. While effective, these pipelines required large, labeled datasets and acted as black boxes, offering little explanation of quality judgments. Additional competition solutions reinforced the strength of tree-based and bagging ensembles over deep models on limited data [14]. Tay et al. [14] provided one of the simplest yet most effective ensemble approaches, while Nguyen et al. [15] extended this with bagging to handle noisy titles. Title refinement using self-supervised multimodal methods has also emerged [16].

LLMs in E-Commerce Classification

Large language models (LLMs) reduce reliance on extensive training data. Palla et al. (2025) [5] proposed "policy-as-prompt," where LLMs apply platform rules directly from textual instructions, offering adaptability without retraining. Roumeliotis et al. (2025) [6] showed GPT-4 and Claude could classify products across 248 categories in zero-shot settings, achieving high accuracy. Cheng et al. (2024) [7] combined fine-tuned classifiers with LLM reasoning, yielding better accuracy than either alone. These works demonstrate LLMs' flexibility and contextual reasoning for classification and moderation.

Research Gap

To date, limited research has applied generative LLMs for auditing product title quality in zero-/few-shot settings. Earlier models relied on retraining with new rules and lacked interpretability. This study applies instruction-tuned LLMs that evaluate titles directly against guidelines in prompts, producing both decisions and explanations. By incorporating category and description context, the system balances relevance, transparency, and adaptability, addressing gaps in scalability and policy compliance for automated title quality assessment.

3. Methodology

Our methodology employs a generative AI–first approach to classify e-commerce product titles as "Good" or "Bad," producing both a decision and an interpretable rationale without requiring traditional model training. Unlike conventional machine learning pipelines, which depend on extensive data collection, manual labeling, and feature engineering, the LLM evaluates raw inputs, product title, category path, and short description through carefully designed prompts. The output is a structured JSON record containing the classification, confidence score, evidence-based reasoning, and checks for category fit and title–description consistency.

This approach offers several advantages over traditional ML. It eliminates the need for labeled training data, provides built-in reasoning and explainability, and allows immediate adaptability through prompt updates. Lightweight safeguards, including objective fact checks (e.g., character count, repeated words) and optional validation prompts, ensure stability and consistency without influencing the model's judgment. A simple decision rule converts outputs into Pass/Fail, with low-confidence cases flagged for human oversight. This architecture supports near–real-time moderation, fast policy iteration, and audit-ready explanations, making the system scalable, transparent, and actionable for marketplace enforcement.

The following section details the System Architecture, describing how the LLM-based evaluation integrates with safeguards, JSON outputs, and decision rules to deliver an end-to-end, explainable title quality control system.

3.1 System Architecture

The proposed system is structured as a modular framework for evaluating e-commerce product titles using a generative AI–first approach. As illustrated in Figure 1, it consists of four primary components, each with a well-defined role:

1) **Normalizer & Precheck:** Prepares inputs by standardizing titles, category paths, and descriptions. This ensures that all inputs are clean, consistent, and interpretable.
2) **LLM Judge (Prompt Orchestrator):** Serves as the central decision-making module. It defines the evaluation criteria based on policies and rules and formulates instructions for the AI system. The Judge interprets outputs and provides reasoning for the classification, ensuring the system's assessments are explainable and aligned with marketplace standards.
3) **LLM Provider Layer:** Acts as the execution engine. It receives instructions from the Judge and performs the evaluation using the AI system. This layer ensures that outputs are consistent, structured, and ready for downstream processing. It does not make decisions; it executes the evaluation.
4) **Policy Aggregator:** Consolidates verified outputs into actionable decisions—Pass, Fail. It preserves rationale and evidence for transparency and auditability.

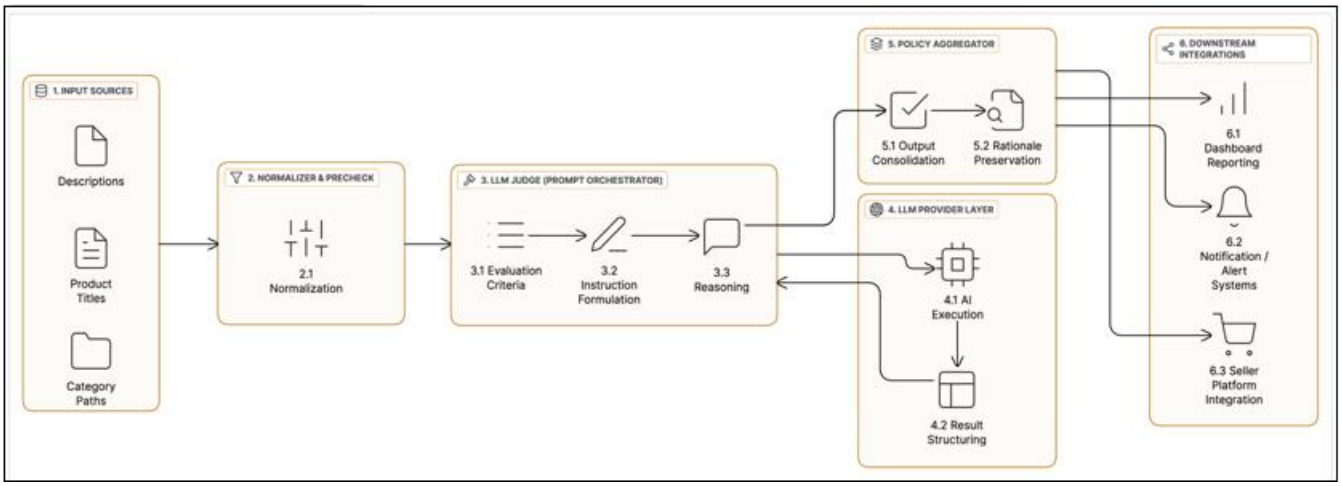

Figure 1: System Architecture for GenAI based Product Title Evaluation

3.2 Data Collection & Preprocessing

In the context of a generative AI–driven methodology, data collection diverges from traditional machine learning practices. Instead of amassing extensive labelled corpora for model training, the focus is on curating a representative, unlabelled dataset to facilitate rigorous evaluation of the LLM's performance. This study utilizes the publicly available Amazon product datasets provided by the Stanford Network Analysis Project (SNAP), which encompass authentic e-commerce listings, including product titles, hierarchical category paths, and short descriptions [8]. These fields constitute the primary inputs for the LLM-based evaluation framework.

To establish a benchmark for quantitative assessment, a carefully selected subset of titles is manually annotated as "Good" or "Bad," forming a gold-standard reference. Preprocessing is deliberately minimal, emphasizing high-level normalization: trimming extraneous whitespace, resolving common Unicode inconsistencies, and standardizing unit representations (e.g., "55-inch" → "55 in"). These measures ensure that input text is clean, consistent, and readily interpretable by the LLM, thereby obviating the need for elaborate feature engineering and preserving the integrity of prompt-driven evaluation.

3.3 Main Methodology

This section describes the methodology for evaluating e-commerce product titles with LLMs. After normalization and pre-check resolve inconsistencies and standardize the dataset, the framework processes three key inputs - product title, category path, and short description. These are transformed into a Good/Bad classification accompanied by clear reasoning, all without relying on traditional machine learning pipelines.

3.3.1 Inputs

The system relies on three fields provided by the seller:

- **TITLE**: the seller-provided product title (raw string).
- **CATEGORY_PATH**: the marketplace taxonomy path rendered as plain text (e.g., *"Electronics › TV & Video › 4K TVs"*).
- **DESCRIPTION**: the short product description supplied by the seller.

No external features, handcrafted signals, or auxiliary metadata are used. All judgments are inferred directly from these three textual fields.

3.3.2 LLM-Guided Evaluation Procedure

The end-to-end evaluation procedure consists of five steps, described below.

Step 1 - Prompt Construction

Following normalization, the three inputs are inserted into a structured prompt designed to enforce deterministic and reproducible judgments. The prompt encodes marketplace quality policies (e.g., overly long, irrelevant keywords, generic adjectives, keyword stuffing, misspellings, excessive symbols, promotional claims, multi-product bundling, category mismatch, and description mismatch).

To minimize stochasticity, decoding parameters were fixed at temperature = 0.2, top_p = 0.9, max_new_tokens = 400, and presence/frequency penalties = 0.0. Where supported, JSON/function-calling mode was enabled to guarantee structural compliance. In practice, stability was further enhanced by prepending two compact few-shot examples (one Good, one Bad) adhering to the same JSON schema.

Figure 2 and Figure 3 illustrates the prompt used in the experiment with placeholders substituted by the seller inputs.

Step 2 - Invocation via Provider Layer

The prompt was submitted to the LLM provider layer, which abstracts execution details of the underlying foundation model. For experimentation, GPT-4.0 was used [9]. Regardless of the provider, outputs were constrained to the same schema, ensuring cross-model consistency.

Step 3 - LLM Judge (Prompt Orchestrator)

The returned output was interpreted by the Judge module, which serves as the central orchestrator of policy enforcement. The Judge does not simply label titles as Good or Bad, but also generates reason codes, verbatim evidence spans, and short explanations for category and description alignment. This ensures that each decision is both interpretable and auditable.

Step 4 - Decision Mapping

Parsed outputs were mapped to final decisions through a transparent set of rules:

- **Fail** if label = Bad and confidence ≥ 0.70.
- **Pass** otherwise.

These thresholds were selected for this study as initial operating points and can be adjusted in future deployments based on empirical validation.

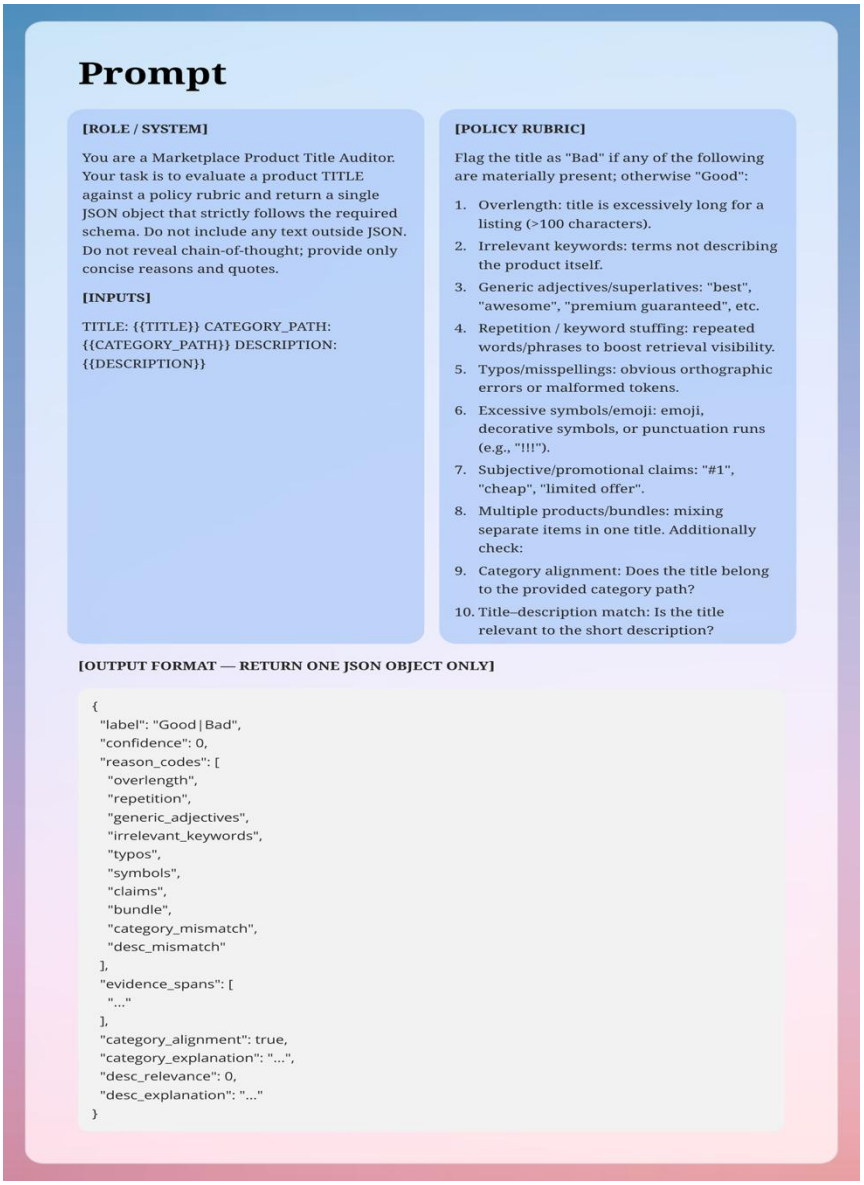

Figure 2: Prompt Construction 1

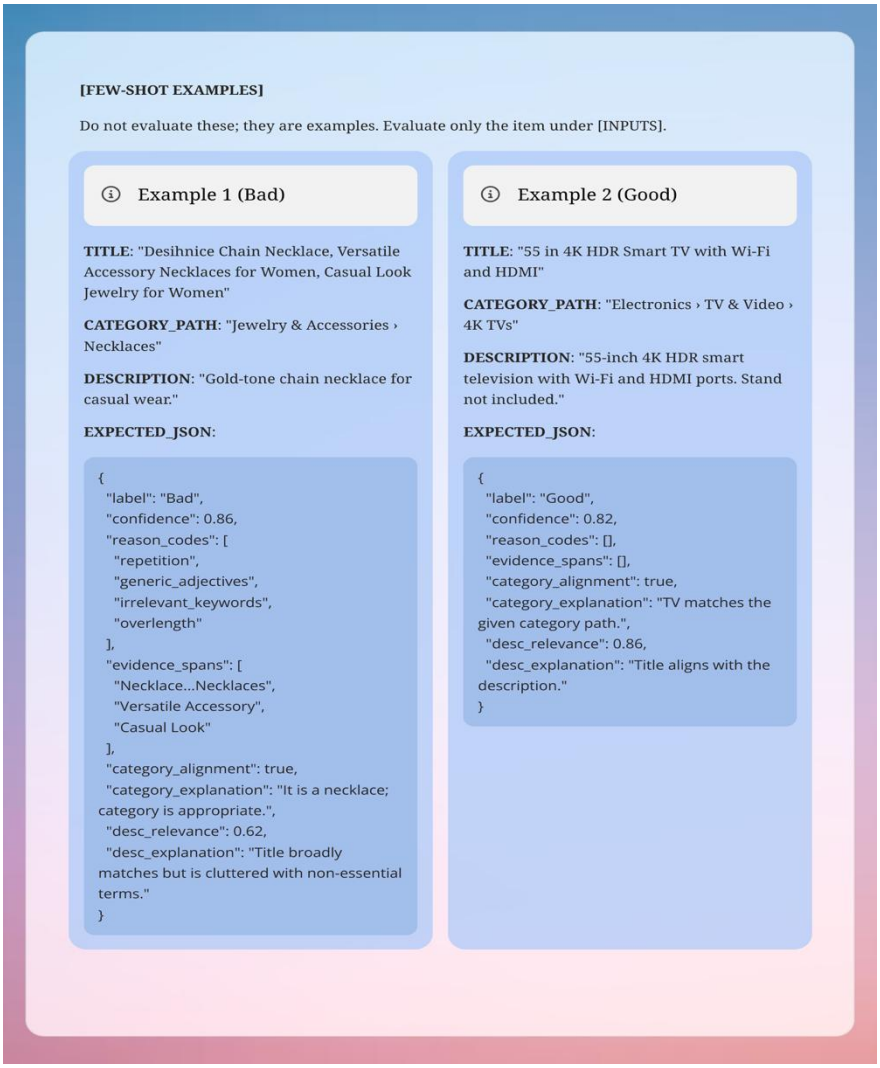

Figure 3: Prompt Construction 2

Step 5 - Result Storage and Transparency

The final structured outputs, including reason codes and evidence spans, were stored for downstream analysis. This not only enables dataset-level quality measurement but also provides sellers and auditors with actionable, token-level evidence of violations.

3.4 Evaluation Results

To quantitatively assess the performance of the proposed LLM-based title evaluation methodology, a subset of 1,000 product titles was stratified randomly selected from the dataset and manually annotated as "Good" or "Bad" to create a gold-standard reference. The evaluation metrics include accuracy, precision, recall, and F1-score [10,11], computed as follows:

- **Accuracy**: proportion of titles correctly classified by the LLM.
- **Precision**: proportion of titles predicted as "Bad" that are "Bad".
- **Recall**: proportion of true "Bad" titles correctly identified by the LLM.
- **F1-score**: harmonic mean of precision and recall.

3.4.1 Quantitative Results

Table 1: Quantitative Results

Metric	*Value*
Accuracy	0.91
Precision	0.88
Recall	0.85
F1-Score	0.86

3.4.2 Qualitative Results

3.4.2.1 Example Outputs

Table 2: Qualitative Results

Case Type	*Input*	*Output*	*Illustrative LLM Output*
Bad	**TITLE:** 55 Inch Smart LED TV 4K UHD HDR10 Flat Screen Television with HDMI, WiFi, TV Stand Compatible **CATEGORY_PATH:** Electronics › TV & Video › 4K TVs **DESCRIPTION:** 55-inch 4K HDR10 smart TV with Wi-Fi and HDMI ports. Stand not included.	**Decision**: Fail (label=Bad, confidence=0.83 ≥ 0.70). **Evidence to reviewer**: Highlight "TV Stand Compatible" and over-descriptive phrase.	{ "label": "Bad", "confidence": 0.83, "reason_codes": ["irrelevant_keywords","generic_adjectives","overlength"], "evidence_spans": ["TV Stand Compatible", "4K UHD HDR10 Flat Screen"], "category_alignment": true, "category_explanation": "The item is a TV; category path is appropriate.", "desc_relevance": 0.75, "desc_explanation": "Most terms match the description; 'stand compatible' is not part of the product." }
Good	**TITLE**: Samsung 55 Inch 4K UHD HDR10 Smart LED TV with Wi-Fi and HDMI **CATEGORY_PATH:** Electronics › TV & Video › 4K TVs DESCRIPTION: 55-inch 4K HDR10 smart TV with Wi-Fi and HDMI ports. Stand not included.	**Decision**: Pass (label=Good, confidence=0.91 ≥ 0.70). **Evidence to reviewer:** Title is concise, descriptive, and free of irrelevant/overly generic terms.	{ "label": "Good", "confidence": 0.91, "reason_codes": [], "evidence_spans": [], "category_alignment": true, "category_explanation": "The product title aligns well with its category path (4K TV).", "desc_relevance": 0.95, "desc_explanation": "Title terms (55-inch, 4K UHD, HDR10, Smart, Wi-Fi, HDMI) are consistent with the description." }

4. Conclusion

The proposed LLM-based, prompt-driven methodology provides a robust, scalable, and explainable framework for e-commerce product title evaluation. By leveraging structured prompts with few-shot examples, the system effectively detects policy violations such as keyword stuffing, overlength titles, and irrelevant promotional content. Structured JSON outputs, including reason codes and evidence spans, enable transparent auditing and support human review in ambiguous cases, improving accountability and interpretability. The approach reduces reliance on costly model training and allows rapid adaptation to evolving marketplace policies, making it readily adoptable by online retailers to enhance title quality, boost search relevance, increase user engagement, and improve operational efficiency.

4.1 Key Findings

- The LLM-based approach achieved 91% accuracy, demonstrating strong capability in detecting policy violations.
- Precision of 88% and recall of 85% indicate a balanced performance between identifying bad titles and minimizing false positives.
- Structured outputs allow transparent auditing, highlighting reason codes and evidence spans for human verification.
- Few-shot examples in prompts improved evaluation stability and robustness across diverse product categories, eliminating the need for large, labeled training datasets.

4.2 Limitations

- The evaluation relies solely on textual inputs (title, category path, description) and does not consider additional product metadata such as images or ratings.
- Borderline cases, such as slightly overlength titles or subtle promotional adjectives, remain challenging and may require iterative prompt refinement or human intervention.

4.3 Future Work

Future research may explore several directions to enhance the proposed LLM-based methodology. Extending the system to support multi-lingual and cross-marketplace scenarios would enable evaluation of product titles in different languages and adaptation to diverse e-commerce taxonomies. Incorporating continuous learning with human-in-the-loop feedback could improve performance on borderline cases and allow dynamic updating of the evaluation rubric to reflect emerging trends or policy changes. Additionally, integrating complementary product metadata, such as images, ratings, or seller reputation, alongside textual analysis could enable a more comprehensive and nuanced assessment of title quality, further improving marketplace search relevance and operational efficiency.

References

[1] S. Hirsch, "What's in a title? Characterizing product titles in e-commerce," *Computers in Industry*, vol. 145, p. 103758, 2025. Available:

https://www.sciencedirect.com/science/article/pii/S0957417425013247.

[2] T. Gupta and S. Bansal, "Impact of keywords in product title on product sales conversion," *SSRN Electronic Journal*, 2023. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4625354.

[3] A. Camargo de Souza, Y. Liu, F. N. Ribeiro, E. S. de Moura, C. A. Davis, T. A. Almeida, and M. A. Gonçalves, "Product title quality prediction for e-commerce," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, 2018, pp. 337–346.

[4] A. Singh and P. Sunder, "Hybrid deep and shallow learning for product title quality classification," in *CIKM AnalytiCup*, 2018.

[5] A. Palla, J. Lee, H. Wang, and J. Chen, "Policy-as-prompt: Flexible content moderation with large language models," *arXiv preprint arXiv:2502.12345*, 2025.

[6] K. Roumeliotis, D. Smith, and F. Zhang, "Evaluating large language models for zero-shot e-commerce product categorization," *Journal of Retail Analytics*, vol. 12, no. 3, pp. 201–215, 2025.

[7] Y. Cheng, A. Kumar, and J. Liu, "Dual-expert framework for product categorization using large language models," *Amazon Science Blog/Conference Paper*, 2024.

[8] SNAP. Stanford Network Analysis Project. Amazon product data. Available from: https://snap.stanford.edu/data/amazon-meta.html

[9] OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. 2023.

[10] Powers D. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies. 2011;2(1):37–63.

[11] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.

[12] Miao L, Cao D, Li J, Guan W. Multi-modal product title compression. Information Processing & Management. 2020;57(1):102123. doi:10.1016/j.ipm.2019.102123.

[13] Zhang JG, Zou P, Li Z, Wan Y, Pan X, Gong Y, Yu PS. Multi-Modal Generative Adversarial Network for Short Product Title Generation in Mobile E-Commerce. In: Proc. NAACL-HLT 2019 (Industry Track). 2019. Available from: https://aclanthology.org/N19-2009/

[14] Tay MCPY, Ong MP, Tang X, et al. A straightforward and effective solution for predicting the quality of a product title. In: CIKM AnalytiCup 2017 – Lazada Product Title Quality Challenge. 2017. Available from: https://cikm2017.org/download/analytiCup/session3/CIKMAnalytiCup2017_LazadaProductTitleQuality_T5.pdf

[15] Ren Z, He X, Yin D, de Rijke M. Information discovery in e-commerce. arXiv preprint arXiv:2410.05763. 2024. doi:10.48550/arXiv.2410.05763.

[16] Deng J, Liu X, Wen J-R. Homogeneous-listing-augmented self-supervised multimodal title refinement (HLATR). 2024. doi:10.1145/3626772.3661347.