Natural Language Processing Journal 11 (2025) 100142

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

LLMs for product classification in e-commerce: A zero-shot comparative study of GPT and claude models

Konstantinos I. Roumeliotis [a,*], Nikolaos D. Tselikas [a], Dimitrios K. Nasiopoulos [b]

[a] Department of Informatics and Telecommunications, University of Peloponnese, 22 131 Tripoli, Greece
[b] Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 118 55 Athens, Greece

ARTICLE INFO

Keywords:
Product classification
Zero-shot classification
LLMs e-commerce
Claude model
GPT model

ABSTRACT

In the rapidly evolving e-commerce landscape, efficient and accurate product classification is essential for enhancing customer experience and streamlining operations. Traditional product classification methods, which depend heavily on labeled data and manual effort, struggle with scalability and adaptability to diverse product categories. This study explores the transformative potential of large language models (LLMs) for zero-shot product classification in e-commerce, addressing the challenge of automating product categorization without prior labeled training data. We evaluate the performance of four state-of-the-art LLMs — GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku — on a diverse dataset of 248 product categories, each containing 20 samples, structured into 8 subsets. Each model performs zero-shot classification, assigning products to predefined categories without prior exposure. Our findings reveal significant variations in classification accuracy across models, with certain LLMs demonstrating superior scalability and adaptability for real-world e-commerce applications. Based on these insights, we developed an API software to integrate the top-performing models into e-commerce systems, enhancing automation and efficiency. This study underscores the transformative role of LLMs in revolutionizing e-commerce workflows and recommends their adoption for scalable, intelligent product classification.

1. Introduction

The rapid expansion of e-commerce has led to increasingly vast and diverse product catalogs, making product classification a crucial yet challenging task. Effective product classification enhances searchability, improves user experience, and optimizes business operations. However, traditional classification methods rely heavily on manual labeling, domain expertise, and extensive training data, making them costly, time-consuming, and difficult to scale. As product catalogs continue to grow, these limitations create bottlenecks in automation and hinder real-time adaptability.

To address these challenges, advancements in artificial intelligence (AI), particularly large language models (LLMs), offer promising solutions. LLMs, such as OpenAI's GPT and Anthropic's Claude models, have demonstrated remarkable capabilities in understanding and generating human-like text. One of their most compelling features is zero-shot learning, wherein a model can perform tasks without prior domain-specific training. By leveraging LLMs for zero-shot product classification, e-commerce platforms can potentially achieve accurate categorization with minimal manual intervention, paving the way for highly automated operations.

This study investigates the potential of large language models (LLMs) for zero-shot product classification—the ability to categorize products without prior exposure to specific classification tasks. The primary objectives are:

Q1: Identify which LLM (GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, Claude 3.5 Haiku) demonstrates superior efficacy in zero-shot product classification.

Q2: Analyze how model performance varies vary across different product categories and dataset subsets.

Q3: Examine the limitations of LLMs in handling ambiguous, short, or domain-specific product title.

Q4: Assess the practicality of integrating an LLM-based API for real-world e-commerce platforms.

To achieve these objectives, we evaluate four state-of-the-art LLMs on a dataset of 248 product categories, each containing 20 product samples, and systematically test their classification accuracy across eight subsets. This ensures a robust comparison of model capabilities under different conditions.

* Corresponding author.
E-mail addresses: k.roumeliotis@uop.gr (K.I. Roumeliotis), dimnas@aua.gr (D.K. Nasiopoulos).

https://doi.org/10.1016/j.nlp.2025.100142
Received 5 December 2024; Received in revised form 21 March 2025; Accepted 22 March 2025

This study is structured to provide a comprehensive analysis of recent advancements in LLMs for e-commerce and their practical applications. Section 2 presents a literature review of the latest developments in LLMs within the e-commerce domain. Section 3 outlines the research methodology employed, while Section 4 delves into the findings and offers a comparative analysis of four models studied. The research culminates in a detailed discussion addressing the core research questions.

The findings of this research contribute to the growing body of knowledge on LLM applications in e-commerce by:

- Conducting a comparative evaluation of four state-of-the-art LLMs (GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku) in zero-shot product classification.
- Establishing a structured methodology to assess LLM performance across diverse product categories and dataset subsets.
- Analyzing the practical challenges and limitations of LLM-based product classification, including handling ambiguous, short, or domain-specific product titles.
- Developing an open-source API that enables automated product categorization, facilitating real-world adoption in e-commerce platforms.

By bridging the gap between AI advancements and industry needs, this study seeks to enhance product classification workflows, making them more efficient, scalable, and adaptable to the evolving demands of the digital marketplaces.

2. Literature review

LLMs have been widely embraced across various industries, empowering businesses to adapt, innovate, and maintain a competitive edge. Among the earliest adopters of LLMs' transformative AI capabilities was the e-commerce sector, where industry leaders have leveraged these technologies to drive automation, improve customer experiences, and optimize operations. This literature review explores the latest advancements in LLM applications within e-commerce, with a focus on their impact and potential for innovation. The review is structured around specific niches, including product recommendations, machine translation, sentiment analysis, and other critical applications, providing a comprehensive understanding of their role in shaping the future of e-commerce. Table 1 presents a summary of studies on LLM applications in e-commerce.

2.1. LLMs in product recommendation

In the rapidly evolving field of product recommendation systems, the integration of LLMs is proving transformative across various studies. Katlariwala and Gupta (2024) propose a novel use of the Llama-2 LLM to enhance product recommendation systems by generating personalized user embeddings. This approach addresses traditional systems' limitations, such as the cold start problem, by improving metrics like click-through and purchase rates. Roumeliotis et al. (2024b) further advance the field with a precision-driven methodology combining unsupervised models and GPT-4 LLM for semantic refining, thereby enhancing recommendation precision and relevance. This integration, implemented through a flask-based API, leverages the advanced language understanding of GPT-4 to significantly enrich the semantic features of product data. Xu et al. (2024a) showcase the LLM's capability to handle complex, multi-dimensional data in e-commerce's recommender systems. Their framework demonstrates marked improvements in precision, recall, F1 score, CTR, and recommendation diversity, emphasizing the LLM's superior understanding of user needs through deep semantic analysis. Together, these studies underscore the profound impact of LLMs on recommender systems, rendering them more effective at capturing nuanced consumer preferences and thereby significantly enhancing user experience and sales in e-commerce domains.

2.2. Generative retrieval and E-commerce search

In addressing the challenges of generative retrieval in e-commerce, Wu et al. (2024) introduce Hi-Gen, a method that improves personalized search systems by leveraging hierarchical encoding–decoding mechanisms. Their approach addresses inefficiencies in existing models, especially concerning docID generation and positional information, which are crucial in large-scale environments. Key innovations include a representation learning model for semantic and efficiency-aware encoding, alongside a hierarchical clustering scheme to enhance docID generation. Conversely, WangHaixun and NaTaesik (2024) propose rethinking e-commerce search by merging structured and unstructured data through a unique framework. They critique traditional methodologies that convert unstructured data into a structured format, which oftentimes is inefficient and low-quality. Instead, they propose a reverse approach by transforming structured data into textual data to be used directly within language models, facilitating search and recommendations through a Q/A mechanism. Xu et al. (2024a) focus on optimizing the technical encoding aspects of generative retrieval addresses practical search efficiency in existing e-commerce structures, while WangHaixun and NaTaesik (2024) propose a conceptual shift by redefining the data integration process for search systems. Both papers, while navigating different aspects of retrieval and search optimization, provide compelling advancements that contribute to the evolving methods in enhancing e-commerce search capabilities.

2.3. Machine translation and product descriptions

The study by Gao et al. (2024) focuses on the shortcomings of LLMs and specialized translation models (STMs) when applied to e-commerce domains, due to the presence of domain-specific terms and complex text structures. The authors introduce an LLMs-based e-commerce machine translation approach termed LEMT. This approach enhances translation quality through the use of specialized resources such as aligned bilingual terms and a parallel corpus derived from real e-commerce scenarios, along with tokenizer optimization and a two-stage fine-tuning process. Their comprehensive evaluations demonstrate that LEMT can surpass existing neural machine translation (NMT) models like NLLB, LLaMA, and even GPT-4 in terms of robustness and translation quality for e-commerce. Zhou et al. (2023) address the challenge of generating high-quality product descriptions in e-commerce, essential for improving search visibility and customer engagement. Utilizing the LLAMA 2.0 7B language model, they offer a scalable solution that reduces human effort and maintains consistency through automation. Both papers highlight the significant potential of LLMs in e-commerce, albeit in different facets—translation and product description. Gao et al. (2024) focus on improving the accuracy and quality of translations in specialized e-commerce languages, whereas Zhou et al. (2023) explore enhancing product descriptions for better customer interaction and business outcomes. Together, these studies suggest a promising trajectory for the application of LLMs in optimizing e-commerce operations, emphasizing domain-specific customization.

2.4. Applications of generative AI in e-commerce

In their research, Ghaffari et al. (2024) provide a comprehensive review of the potential applications of generative AI in e-commerce, highlighting its capacity to enhance both customer experience and merchant productivity. Key use-cases explored include product description generation, sentiment analysis of reviews, and product categorization. The paper also addresses challenges and risks associated with integrating generative AI technologies, particularly through prompt engineering with LLMs. Farfade et al. (2024) address a different aspect of e-commerce in their work, introducing a use-case based shopping (UBS) system. This system aims to improve customer experience by linking product use-cases with seller-provided attributes through LLMs,

Table 1
Summary of recent studies on LLM applications in e-commerce.

Study	Authors	Methodology	Elements	Results
LLM-enhanced product recommendation	Katlariwala and Gupta (2024)	Llama-2 LLM for personalized user embeddings	User embeddings, cold start problem, CTR, purchase rates	Improved CTR and purchase rates by addressing cold start issues
Precision-driven recommendation	Roumeliotis et al. (2024b)	Unsupervised models + GPT-4 for semantic refining	Flask-based API, semantic enrichment	Enhanced recommendation precision and relevance
Multi-dimensional recommender system	Xu et al. (2024a)	LLM framework for deep semantic analysis	Precision, recall, F1 score, CTR, recommendation diversity	Marked improvements in all metrics, better user preference understanding
Hierarchical generative retrieval (Hi-Gen)	Wu et al. (2024)	Hierarchical encoding–decoding for personalized search	Representation learning, docID generation, clustering	Improved efficiency in personalized search retrieval
E-commerce search restructuring	WangHaixun and NaTaesik (2024)	Transform structured data into text for LLM use	Structured/unstructured data integration	Improved search and recommendation efficiency
E-commerce machine translation (LEMT)	Gao et al. (2024)	LLM-based translation with aligned bilingual terms	Parallel corpus, tokenizer optimization, fine-tuning	Surpasses NMT models (NLLB, LLaMA, GPT-4) in translation quality
LLM-generated product descriptions	Zhou et al. (2023)	LLAMA 2.0 7B model for scalable description generation	Automation, consistency, search visibility	Reduced human effort, improved product descriptions
Generative AI applications in e-commerce	Ghaffari et al. (2024)	Review of AI use cases (product descriptions, sentiment analysis, categorization)	LLM prompt engineering, risks	Highlights potential and challenges of AI integration
Use-case Based Shopping (UBS) System	Farfade et al. (2024)	LLMs for linking use-cases to seller attributes	Product use-cases, attribute matching	Overcomes data limitations in product categorization
Customer trust analysis in e-commerce	Davoodi and Mezei (2024)	LLMs + QCA to analyze customer trust determinants	Selection, post-purchase support	Identifies key trust factors in the customer journey
LLMs vs. traditional ML in categorization	Ihsanoğlu et al.	Comparative analysis of ML and LLMs for product categorization	Operational cost, resource efficiency	LLMs offer no major advantage despite text understanding
Dual-expert product categorization	Cheng et al. (2024)	Domain-specific knowledge + LLMs for classification	In-context learning, self-summarization	Improved accuracy with hybrid approach
Entity resolution with cost-effective prompts	Nananukul et al. (2024)	GPT-3.5 prompt engineering for product matching	Simple vs. complex prompts	Simple prompts can match complex ones in effectiveness
Sentiment analysis in e-commerce	Roumeliotis et al. (2024a)	Comparing LLMs (GPT-3.5, LLaMA-2) with NLP models (BERT, RoBERTa)	Fine-tuning, customer sentiment, AI-driven analysis	Improved customer satisfaction decoding via LLMs
Proactive consumer insights extraction	Shahin et al. (2024)	GPT-3.5 Turbo for VoC analysis	Traditional vs. AI-driven consumer insights	AI methods outperform surveys/interviews in capturing VoC
LLMs in recommender systems	Xu et al. (2024b)	Review of LLMs vs. deep neural networks in recommendations	User/item representations, personalization	LLMs provide better textual and user interest modeling
Conversational recommender systems (CRS) + LLMs	Liu et al. (2023)	CRS and LLM collaboration strategies	Pre-sales dialogues, task performance, domain expertise	Combined approach enhances accuracy in e-commerce dialogues
Relation labeling in product knowledge graphs	Chen et al. (2024)	LLMs (PaLM-2, GPT-3.5, LLaMA-2) for few-shot learning	Complementary/substitutable product relations, KG completion	LLMs match human labelers, surpass traditional KG models
LLMs in broader KG completion	Chen et al. (2023)	Empirical study on few-shot learning for KG tasks	Knowledge graphs, model fairness, bias	LLMs reduce human labor costs, improve accuracy
LLMs for relevance assessment	Soviero et al. (2024)	LLMs replacing human annotators in relevance tasks	Human vs. AI-generated relevance judgments	82% agreement with human annotations
LLM-Ensemble for attribute extraction	Fang et al. (2024)	Multi-LLM approach for product attributes	Walmart data, model weighting optimization	Improved GMV, CTR, CVR, ATC
Instruction-tuned LLMs for e-commerce tasks	Palen-Michel et al. (2024)	Fine-tuning LLMs with e-commerce datasets	Classification, NER, summarization	Smaller, task-specific models outperform few-shot larger LLMs
LLMs + tensor factorization for product discovery	Wang et al. (2024)	Context-aware ranking using LLMs & IR features	User demographics, session behaviors	Improved search precision (NDCG, AP metrics)
LLaSA: Versatile e-commerce shopping assistant	Zhang et al. (2024)	LLMs + EshopInstruct dataset (65k samples)	Task diversity, instruction tuning	Enhanced assistant adaptability & efficiency
EcomGPT: Chain-of-Task learning	Li et al. (2024)	EcomInstruct dataset (2.5M samples) for LLM generalization	Atomic task construction, zero-shot learning	Improved generalization in e-commerce

thereby overcoming the practical challenges of training models across various product categories without extensive datasets. Davoodi and Mezei (2024), in their study focus on analyzing and predicting customer trust using LLMs combined with qualitative comparative analysis

(QCA). They explore how different stages of the customer journey, such as selection and post-purchase support, impact trust, providing a nuanced understanding of trust determinants in e-commerce. Collectively, these papers underscore the vast potential of LLMs in transforming e-commerce through use-case applications, product categorization and recommendation, and trust analysis, while also highlighting the necessity of addressing associated challenges.

2.5. Product categorization and entity resolution

The research article by Ihsanoğlu et al. examines the use of traditional machine learning algorithms versus LLMs for product categorization in e-commerce, concluding that LLMs do not markedly outperform traditional methods despite their capability to understand complex text. This implies that the choice between these methodologies should be informed by their operational costs and resource demands rather than performance. Cheng et al. (2024) propose a dual-expert classification system for e-commerce product categorization that leverages LLMs for improved accuracy by melding domain-specific knowledge with LLMs' general understanding. They demonstrate enhanced performance by incorporating in-context learning and self-summarization. Meanwhile, Nananukul et al. (2024) explore entity resolution, particularly focusing on cost-effective prompt engineering with LLMs like GPT-3.5 for product matching. They find that simple, inexpensive prompts are often just as effective as complex, costly ones. Collectively, these studies highlight the nuanced role of LLMs in e-commerce, suggesting their limited advantage in categorization tasks without significant cost considerations, but revealing their potential in dual-expert systems and cost-sensitive entity resolution.

2.6. Sentiment analysis and customer insights

In their research, Roumeliotis et al. (2024a) delve into the use of LLMs such as GPT-3.5 and LLaMA-2 for sentiment analysis in e-commerce, comparing these with NLP models like BERT and RoBERTa. They aim to assess how effectively these models, especially when fine-tuned, can decode customer satisfaction from product reviews. This study stands out by focusing on the shift toward understanding emotions and customer sentiment through advanced AI models, demonstrating how LLMs can meaningfully inform customer satisfaction strategies in the e-commerce sector—an area critical for the industry's growth post-pandemic. Similarly, Shahin et al. (2024) explore the potential of the GPT-3.5 Turbo model in extracting consumer insights proactively, emphasizing its application in product development. Their research indicates that traditional methods like surveys and interviews have limitations in capturing the true voice of the customer (VoC). Both papers accentuate the transformative potential of LLMs in enhancing customer service and satisfaction understanding, emphasizing the shift from traditional methods to advanced AI-driven approaches in both product review sentiment analysis and proactive VoC extraction. They collectively underline the necessity and ethical implications of adapting these cutting-edge technologies in evolving e-commerce and service industries.

2.7. Personalized recommendations and conversational systems

The research paper by Xu et al. (2024b) provides a comprehensive review of the emerging intersection between LLMs and recommender systems within e-commerce, emphasizing the limitations of traditional deep neural networks in capturing user interests and textual information effectively. A critical aspect of this review is the analysis of user and item representations facilitated by LLMs, providing valuable insights into potential improvements in personalized recommendations. Liu et al. (2023) focus on the synergies between conversational recommender systems (CRSs) and LLMs in e-commerce pre-sales dialogues. While CRSs are designed to understand user representations through dialogue context and require external knowledge for accurate recommendations, LLMs excel in generating natural language responses but lack domain-specific expertise. They identify this complementarity and examine two collaboration strategies: the use of CRS to support LLMs and vice versa. Through empirical experiments on a real-world dataset, they demonstrate the effectiveness of these collaborative approaches in enhancing task performance in e-commerce dialogues. Together, these papers illustrate the potential of leveraging LLMs to address existing limitations in both recommendation accuracy and conversational interaction within e-commerce contexts. While Xu et al. (2024b) shine a light on the broader potential of LLMs in enriching recommendation systems, Liu et al. (2023) provide practical insights into the specific context of pre-sales dialogue, driving home the complementarity of LLMs and CRSs. This sets a promising stage for future exploration and development in personalized recommendation systems powered by LLMs.

2.8. E-commerce knowledge graphs

In their 2024 paper on relation labeling in product knowledge graphs, Chen et al. (2024) explore the potential of LLMs for improving e-commerce systems by structuring and predicting product relations such as complementary or substitutable products. They focus on the challenges posed by the dynamically changing e-commerce domain and the cost of human annotation. The authors utilize LLMs like PaLM-2, GPT-3.5, and Llama-2, assessing their effectiveness in few-shot learning scenarios for relation labeling tasks. Through careful prompt engineering, the paper demonstrates that LLMs perform comparably to human labelers with minimal labeled examples, and significantly surpass traditional knowledge graph (KG) completion models. In a parallel paper, Chen et al. (2023) extend their empirical investigation to KGs more broadly. They show similar performance benefits of few-shot learning capabilities of LLMs when incorporated into KG completion tasks, thus reinforcing the applicability of these advanced models in e-commerce. Both papers highlight the transformative potential of LLMs for reducing human labor costs and improving accuracy in e-commerce KGs, while simultaneously pointing to necessary considerations regarding model fairness and bias.

2.9. LLMs as a tool for relevance assessment

The research by Soviero et al. (2024) explores the potential of LLMs to replace human annotators in relevance assessment tasks specifically within an e-commerce setting. The authors conducted experiments using both open and proprietary datasets to compare LLM-generated relevance judgments with human annotations, finding a strong agreement of 82% between the two. This indicates the promising capability of LLMs as alternatives to human involvement in relevance judgment tasks. On the other hand, Fang et al. (2024) address the challenges inherent in extracting product attribute values using LLMs in e-commerce. Recognizing that different LLMs bring unique strengths and weaknesses to various tasks due to their diverse architectures and hyperparameters, the authors propose a novel algorithm termed LLM-Ensemble. This algorithm leverages the complementary capabilities of multiple LLMs by iteratively learning and optimizing the weights assigned to each model's output, aiming for accurate and efficient predictions of product attribute values. Their findings, based on experiments with Walmart's internal data, indicate that the LLM-Ensemble method not only outperforms single LLMs but also enhances business metrics such as gross merchandise volume (GMV), click-through rate (CTR), conversion rate (CVR), and add-to-cart rate (ATC). Both papers underscore the transformative role of LLMs in e-commerce, from streamlining relevance assessments to enhancing attribute extraction. While Soviero et al. (2024) focus on substituting human judgment with LLMs, Fang et al. (2024) demonstrate the efficacy of combining multiple LLMs for optimized outputs, both contributing significantly to advancing the efficiency and effectiveness of e-commerce applications.

2.10. Comprehensive evaluations of LLMs in E-commerce

The study by Palen-Michel et al. (2024) provides a thorough evaluation of LLMs in the e-commerce domain, focusing on tasks such as classification, generation, summarization, and named entity recognition (NER). By instruction-tuning an open-source LLM with diverse e-commerce datasets, the authors compared its performance against traditional pre-trained models commonly used in the industry. Their findings reveal that task-specific fine-tuning of smaller models often surpasses the performance of few-shot inference with larger LLMs. In contrast, Wang et al. (2024) present a novel approach for improving product discovery in e-commerce by integrating LLMs with tensor factorization techniques. Their method captures user–object–content interactions using a multifaceted context representation, which includes factors like user demographics and session behaviors. The research introduces an innovative context-aware ranking algorithm that melds traditional information retrieval features with LLM-derived semantic signals, yielding substantial improvements in average precision and normalized discounted cumulative gain across large e-commerce datasets. Both papers emphasize the transformative potential of LLMs in e-commerce, yet highlight differing aspects: one through fine-tuning and training method comparisons, and the other via strategic system architecture and integration for improved search outcomes.

2.11. Instruction-tuning and optimization of LLMs for E-commerce

In their paper, Zhang et al. (2024) address the challenges of task-specificity and poor generalization in e-commerce shopping assistants by employing LLMs to construct a versatile assistant named LLaSA. The authors identified that prior assistants required specialized models for different tasks, increasing costs and limiting performance with new product data. To tackle this, they created the EshopInstruct, a dataset with 65,000 samples encompassing diverse tasks, and applied instruction tuning to enhance LLaSA's versatility and performance. Li et al. (2024) also focused on enhancing LLMs for e-commerce applications, proposing EcomGPT and introducing EcomInstruct, a comprehensive instruction dataset of 2.5 million data points. The dataset was designed to enhance the LLM's ability to generalize across datasets and tasks through Chain-of-Task tasks, which involve defining atomic tasks as building blocks for more complex tasks. This approach, leveraging the fundamental semantic understanding capabilities of LLMs, resulted in EcomGPT achieving outstanding zero-shot generalization capabilities in e-commerce settings. While Zhang et al. (2024) emphasize effective model orchestration and resource-efficient inference strategies, Li et al. (2024) focus on enhancing semantic understanding through atomic task construction. Collectively, they illustrate innovative approaches to optimizing LLMs for e-commerce, substantially contributing to the field by blending novel dataset construction with intelligent model application.

The studies examined in this review collectively underscore the e-commerce industry's significant shift toward leveraging the advanced capabilities of LLMs for diverse automation tasks. These advancements not only streamline operational efficiency but also drive enhanced performance and increased sales, solidifying the role of LLMs as a transformative force in the e-commerce landscape.

3. Research methodology

3.1. Dataset preparation

The success of any machine learning (ML) application heavily depends on the quality and structure of the dataset used. For this study, the dataset serves as the foundation for evaluating the performance of LLMs in zero-shot product classification. Given the diversity and complexity of e-commerce product catalogs, it is essential to utilize a dataset that accurately reflects real-world scenarios while ensuring it is manageable for experimental purposes.

This section outlines the dataset preparation process in detail, focusing on its description, segmentation strategy, and preprocessing steps. These steps were carefully designed to create a robust and balanced dataset that facilitates fair and meaningful evaluation of the models (see Fig. 1).

3.1.1. Description of the dataset

For this study, we utilized the Amazon Products Dataset 2023, one of the most comprehensive resources available for e-commerce research (Asaniczka, 2023). This dataset includes approximately 1.4 million products and their corresponding categories, sourced from Amazon, the largest online retailer in the United States, known for its vast catalog of over 12 million products (Chhabra et al., 2024). The dataset offers valuable insights into product trends, category diversity, pricing strategies, and SEO practices. Hosted on Kaggle, it is licensed under the Open Data Commons Attribution License (ODC-By) v1.0 and has received a perfect usability score of 10 from users, reflecting its credibility, completeness, and compatibility. Regular updates further enhance the dataset by incorporating new and diverse product data.

The dataset is provided in two files:

- `amazon_products.csv` — Contains details about products.
- `amazon_categories.csv` — Includes category information linked via a foreign key relationship, where the `category_id` column in `amazon_products` references the `id` column in `amazon_categories`.

To streamline our analysis, we merged these two files into a single dataset, retaining only the columns relevant to our study: `title` (product title) and `category` (category name). This approach omits extraneous columns such as `imgUrl`, `productURL`, `stars`, `reviews`, `price` etc. focusing exclusively on the data needed for product classification.

The processed dataset comprises 248 unique product categories, showcasing remarkable diversity. Examples include PC games accessories, personal care products, and science education supplies, highlighting the dataset's broad scope and confirming its suitability for assessing the generalizability of our models. However, the dataset also presents some variability in category distribution, with certain categories containing as many as 492 products, while others include as few as 40.

The average length of entries in the `title` column was 116.89 characters, while for the `category` column, it was 19.72 characters. These lengths reflect the granularity of the data, providing a robust foundation for exploring zero-shot classification capabilities. The dataset's completeness and representativeness position it as an ideal resource for evaluating LLMs in the context of e-commerce.

3.1.2. Data segmentation and preprocessing

The Amazon Products Dataset 2023, as described earlier, is an excellent resource for classification tasks, sentiment analysis, and price prediction. For this study, we required our four selected models to classify products into predefined categories without prior training (zero-shot classification), relying only on a provided list of category labels. While the complete dataset with its 1.4 million products is highly comprehensive, its sheer size presented challenges for our research objectives. Specifically, the large volume of data would make it difficult to meticulously track model-specific behaviors during the classification task and would significantly increase costs due to API usage fees for LLMs, which are determined by the number of input and output tokens.

To address these concerns, we created a smaller, balanced dataset by randomly selecting 20 products from each of the 248 categories in the original dataset. This approach resulted in a balanced dataset with

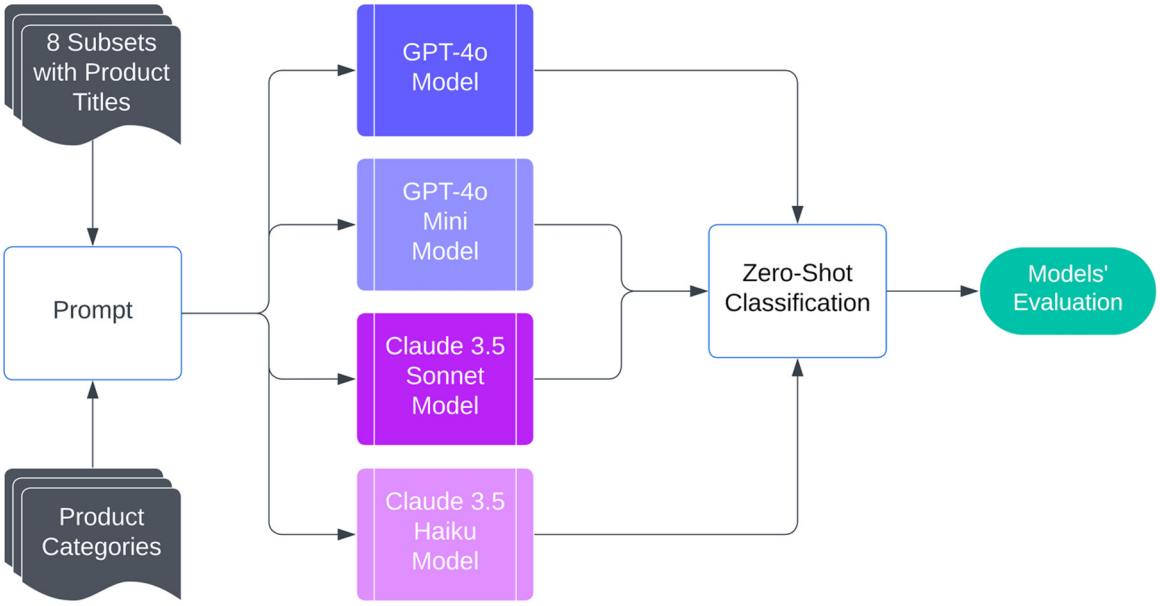

Fig. 1. Zero-shot classification process.

4,960 products—sufficient to evaluate the models' zero-shot classification capabilities while remaining manageable in terms of cost and analysis complexity.

Prior to this selection, the full dataset underwent preprocessing to ensure consistency and quality. The steps included:

- Text Normalization: Standardizing text in the `title` and `category` columns to ensure uniform formatting.
- Tokenization: Breaking down text into manageable units to optimize compatibility with the LLMs.
- Cleaning: Removing empty rows to ensure no incomplete data interfered with the analysis.
- ID Column Addition: Introducing a unique identifier for each product to facilitate referencing.

The resulting processed dataset comprised 248 categories with 20 products each, providing a balanced distribution for evaluating model performance.

Given the constraints of LLMs, such as limits on input and output tokens, we further divided this dataset into 8 subsets, each containing 31 categories and 620 products. This segmentation allowed us to:

- Analyze the predictive capabilities of each model at the subset level.
- Evaluate overall performance across all subsets in a structured manner.

The rationale for segmenting the dataset into 8 subsets is closely linked to the prompt engineering considerations discussed in Section 3.2. By organizing the data in this way, we could better manage computational resources and ensure that the tasks presented to the LLMs were within their token limits, facilitating accurate and efficient zero-shot classification.

3.2. Prompt engineering

Prompt engineering is a crucial aspect of working with LLMs to ensure optimal performance in various tasks, including zero-shot classification. It involves crafting well-structured input queries — referred to as prompts — that guide the model to generate the most accurate and relevant output. The quality and clarity of a prompt can significantly influence the model's ability to understand the task at hand and deliver the desired results, making it an essential skill for utilizing LLMs effectively.

In the context of this research, prompt engineering plays a central role in ensuring that the models correctly classify products into predefined categories without prior training on those categories. Given the flexibility of LLMs, prompt design becomes an iterative process where adjustments are made to improve model understanding, reduce ambiguity, and enhance classification accuracy. Factors such as the specificity of instructions, the format of input data, and the structure of the prompt can all impact the model's ability to perform well in real-world scenarios.

To develop a universal prompt applicable to all models, we began by conducting trial-and-error tests using each model's chat interface. After several iterations, we arrived at a prompt structure that all models could understand. Recognizing that LLMs typically perform better with structured data rather than unstructured text, we initially experimented with XML-based formatting to organize the prompt. This approach allowed us to present the product and category information in a more structured manner, making it easier for the models to parse and understand the data. However, while this structure made the prompt more understandable to the models, it also significantly increased the length of the prompt, resulting in a higher number of input tokens and thus more resource-intensive API calls.

To address this issue, we turned to Anthropic's console prompt generator, a tool designed to help users craft effective prompts for their models (Anthropic PBC, 2024a). As outlined in Anthropic's documentation, the prompt generator alleviates the "blank page problem" by providing high-quality prompt templates tailored to specific tasks, incorporating best practices for prompt engineering. After experimenting with this tool, we found that using JSON format for the prompt was more efficient than XML. JSON not only streamlined the structure but also reduced the overall token count, thus making the process more cost-effective and resource-efficient.

The final prompt was refined to ensure compatibility across all four LLMs, balancing clarity and efficiency. An example of the finalized prompt is shown in Fig. 2. This structured approach to prompt engineering allowed us to maximize the performance of the models while managing the constraints of the API calls effectively.

```
prompt = (
    'Your task is to analyze a given product title and assign '
    'it to the most appropriate category from a predefined list.\n'
    '{'
    '    "product_title": "usb c to lightning cable apple mfi certified 3pack",\n'
    '    "categories": [\n'
    '        "abrasive finishing products",\n'
    '        "accessories supplies",\n'
    '        "baby activity entertainment products",\n'
    '        "baby boys clothing shoes",\n'
    '        "baby care products",\n'
    '        ...\n'
    '    ]\n'
    '}\n\n'
    'Provide your final classification in the following JSON format '
    'without explanations: {"category": "chosen_category_name"}'
)
```

Fig. 2. Universal JSON-formatted prompt for zero-shot classification.

Table 2
Comparison of models' internal structure and performance.

Model	Context Window	Performance	Speed	Input Cost per 1M Tokens (USD)	Output Cost per 1M Tokens (USD)	Best Use Case
GPT-4o	128K tokens	High	Moderate	2.5	1.25	Complex, high-stakes classification
GPT-4o Mini	128K tokens	Moderate	High	0.15	0.08	Real-time, cost-efficient classification
Claude 3.5 Sonnet	200K tokens	High	Moderate	3	15	Complex problem-solving, deep analysis
Claude 3.5 Haiku	200K tokens	Moderate	High	0.25	1.25	Fast, high-volume classification tasks

3.3. Model selection

The choice of models is a critical component of this research, as it directly impacts the ability to achieve accurate and scalable product classification. In this study, we focus on state-of-the-art LLMs that leverage advanced natural language processing capabilities. These models have demonstrated exceptional performance across various tasks, making them promising candidates for addressing the challenges of zero-shot product classification.

Traditional classification models often require extensive task-specific training to perform effectively. However, LLMs, being pre-trained on vast amounts of text data, possess a deep understanding of language and context. This pre-training enables them to interpret and process the nuances of product titles and categories without needing additional fine-tuning. In essence, their pre-trained nature allows them to understand the underlying context behind each word, whether in the product title or category, making them exceptionally suited for zero-shot classification tasks.

This section introduces the selected LLMs — GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku — and their relevance to the research objectives. It also elaborates on the zero-shot learning paradigm, a key aspect of this study, which enables these models to classify products into predefined categories without the need for additional fine-tuning or task-specific training. In Table 2, a brief comparison of the models' internal structure and performance is presented.

3.3.1. GPT models

The GPT-4o and GPT-4o mini models are key components of OpenAI's efforts to refine and optimize LLMs for enhanced performance, efficiency, and cost-effectiveness. GPT-4o, with its 128K-token context window (OpenAI, Inc., 2024a), excels in complex, high-stakes classification tasks, offering high performance at moderate efficiency and a higher cost ($2.50 per million input tokens, $1.25 for output) (Roumeliotis et al., 2024a). GPT-4o mini, also supporting 128K tokens, is a cost-efficient alternative optimized for real-time classification, delivering moderate performance with high efficiency at a lower cost ($0.15 per million input tokens, $0.08 for output) (OpenAI, Inc., 2024b,c). These models are particularly relevant to this research due to their advanced NLP capabilities, which make them suitable for tasks such as zero-shot classification in e-commerce product classification.

3.3.2. Claude models

Claude 3.5 Sonnet and Claude 3.5 Haiku, part of Anthropic's Claude family, are advanced NLP models suited for zero-shot product classification. Claude 3.5 Sonnet, with a 200K-token context window, excels in complex problem-solving (Oh et al., 2024), coding abilities (Rahman et al., 2024) and deep analysis, making it ideal for high-stakes classification tasks. It offers high performance at moderate efficiency but comes with a significantly higher output cost ($3 per million input tokens, $15 for output). Claude 3.5 Haiku, optimized for speed and affordability, supports fast, high-volume classification with moderate performance and high efficiency (Anthropic PBC, 2024b). While more expensive in terms of output cost than GPT models, both Claude models provide strong AI capabilities (Nguyen et al., 2024), serving as valuable benchmarks in this research.

3.3.3. Zero-shot learning

Zero-shot learning is an advanced paradigm in NLP where models classify data into predefined categories without requiring any task-specific training or fine-tuning (López Espejel et al., 2023). This approach leverages the pre-trained knowledge of LLMs to infer relationships and make predictions based solely on the input provided (D'Asaro et al., 2024). In the context of e-commerce product categorization, zero-shot learning is particularly advantageous as it eliminates the need for costly and time-consuming model fine-tuning processes (Roumeliotis et al., 2023). By providing a prompt with a list of potential categories and a product description, LLMs can deduce the most appropriate classification based on their extensive pre-trained knowledge. This capability makes zero-shot learning highly scalable, efficient, and adaptable for real-world applications, allowing for rapid implementation across diverse product catalogs with minimal setup. Moreover, it provides a practical solution for managing dynamic and extensive datasets like those in e-commerce, where categories and products are frequently updated.

3.4. Evaluation metrics

To assess the performance of the selected models in the zero-shot classification task, we utilized standard evaluation metrics: accuracy, precision, recall, and F1 score. These metrics were applied across the 8 subsets of the dataset, each containing 31 categories with 20 products per category. Accuracy provided an overall measure of correct classifications, while precision and recall offered deeper insights into the models' ability to correctly identify products within a category and minimize false positives and negatives (Munkova et al., 2020). The F1 score, as the harmonic mean of precision and recall, was particularly relevant for balancing these metrics, ensuring a comprehensive evaluation of the models' effectiveness (Yao and Shepperd, 2021). These metrics are well-suited for e-commerce classification tasks where both the correctness and reliability of predictions are crucial for enhancing user experience and operational efficiency.

In addition to these quantitative metrics, we employed the top-influential words method to qualitatively analyze why each model made its predictions for specific categories. This method helped us identify the key words or phrases within product titles that influenced the models' decisions, providing valuable insights into their interpretability and alignment with human reasoning (Huang et al., 2003). This dual approach of quantitative and qualitative evaluation ensured a robust and meaningful assessment of the models' performance in real-world e-commerce scenarios.

3.5. API development and integration

Transforming research findings into practical solutions is a vital step in bridging the gap between academic innovation and real-world application. In this study, the development of an API operationalizes the best-performing LLMs for zero-shot product classification in e-commerce platforms. The API is designed to automate product categorization with high efficiency and scalability, enabling seamless integration into existing e-commerce systems. By acting as an intermediary, it eliminates the complexities of prompt crafting, processing model outputs, and managing APIs, simplifying the adoption of advanced AI capabilities for e-commerce businesses. Notably, all components of this project — including the models' deployment scripts, the Flask API, and the evaluation source code — are made available under an open-source MIT license via a GitHub repository, promoting accessibility and collaboration (Roumeliotis, 2024b).

This section details the API's architectural design, implementation workflow, and key operational considerations, emphasizing its functionality, adaptability, and performance in real-world scenarios.

3.5.1. Implementation environment

The API was developed using Flask, a lightweight Python framework, and deployed on the PythonAnywhere web hosting service, which is tailored for Python projects (PythonAnywhere LLP, 2024). This API accepts requests in JSON format, parses the input, crafts an optimized prompt, and communicates with OpenAI's and Anthropic's APIs to classify products. It returns the processed results in a structured format, ensuring usability across diverse e-commerce systems.

The choice of Flask and PythonAnywhere ensures a reliable, scalable, and cost-effective environment for hosting the API. Additionally, leveraging OpenAI's and Anthropic's APIs ensures that the API harnesses state-of-the-art language models without requiring extensive infrastructure from the user.

3.5.2. Implementation workflow

The API processes product classification requests as follows:

- Input: The user submits a JSON payload containing the API key (OpenAI or Anthropic), the product title, and an array of product categories.
- Prompt Crafting: Using the provided information, the API dynamically constructs the most effective prompt tailored to the specified language model.
- Model Interaction: The API communicates with the specified model via its official API, sending the crafted prompt and retrieving the classification results.
- Processing Results: The raw model output is cleaned, parsed, and formatted to ensure clarity and precision.
- Output: The processed results are returned to the user in JSON format, indicating the most appropriate category for the given product title.

This streamlined workflow minimizes latency and computational overhead, with prompt crafting and result processing optimized for efficiency. The API's intermediate role ensures that users, such as e-commerce business owners, can leverage cutting-edge LLMs without needing specialized knowledge of AI models, prompt engineering, or output processing.

4. Research findings: Comparative analysis and model evaluation

This chapter provides an in-depth evaluation of the performance of four state-of-the-art LLMs — GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku — on eight datasets specifically designed for zero-shot product classification. The 8 datasets, each containing 31 predefined product categories with 20 product samples per category, serve as the basis for comparing the models' capabilities in terms of accuracy, precision, recall, and F1-score. The analysis not only highlights the strengths and weaknesses of each model but also identifies trends across datasets that offer insights into their scalability and suitability for e-commerce automation tasks.

4.1. Overall model performance across datasets

The comparative analysis of the selected LLMs — GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku — across eight dataset splits highlights notable variations in their zero-shot product classification capabilities.

On average, the models achieved an accuracy of 74.62%, underscoring their proficiency in handling complex classification tasks. The lowest accuracy, 63.71%, was observed with the Claude 3.5 Haiku model on the first dataset split, while the highest accuracy, 85.81%, was achieved by the Claude 3.5 Sonnet model on the fifth split. These results illustrate the robustness of LLMs in classification tasks, a domain often considered challenging for models primarily designed for text generation and other general-purpose applications.

Table 3
Zero-shot product classification: evaluation metrics for models on subset 2.

Model	Accuracy	Precision	Recall	F1
gpt_4o_prediction	0,7677	0,8026	0,7677	0,7608
gpt_4o_mini_prediction	0,7419	0,7873	0,7419	0,7368
claude_3.5_sonnet_prediction	0,7435	0,7861	0,7435	0,7328
claude_3.5_haiku_prediction	0,7371	0,7908	0,7371	0,7314

Table 4
Zero-shot product classification: evaluation metrics for models on subset 5.

Model	Accuracy	Precision	Recall	F1
gpt_4o_prediction	0,8371	0,8538	0,8371	0,8361
gpt_4o_mini_prediction	0,771	0,8239	0,771	0,7577
claude_3.5_sonnet_prediction	0,8581	0,8822	0,8581	0,8569
claude_3.5_haiku_prediction	0,7774	0,806	0,7774	0,7608

It is worth noting that the task involved classifying products into 31 distinct categories based solely on their titles. With a random chance of selecting the correct category being just 3.21% (1/31), the models' performance — ranging from 63.71% to 85.81% — demonstrates their remarkable capacity for zero-shot inference. This performance underscores the impact of the models' pre-training, enabling them to excel at complex tasks without additional fine-tuning.

Performance differences among the models can be attributed to variations in architecture, training objectives, and the trade-offs between accuracy, computational speed, and cost-efficiency. Each model exhibited distinct strengths and weaknesses:

- Some models excelled in scenarios requiring nuanced reasoning and deeper contextual understanding.
- Others performed more consistently on tasks involving straightforward categorizations.

These findings emphasize the importance of selecting a model based on specific use cases, especially when considering constraints such as budget or response time requirements.

The evaluation metrics — accuracy, precision, recall, and F1-score — offered a well-rounded perspective on performance. Metrics for each dataset split were computed to analyze how the models handled diverse product categories. Detailed results for subset 2 and subset 5, provided in Table 3 and Table 4, exemplify the models' predictive capabilities for these representative splits.

For a comprehensive review of performance across all dataset splits, the full evaluation report is accessible in the accompanying GitHub repository. This ensures transparency and facilitates further analysis and validation (Roumeliotis, 2024a).

The comparative analysis of the models across eight dataset splits highlights notable differences in their capabilities for zero-shot product classification. Each model demonstrated unique strengths and weaknesses, reflecting their varied design and tuning for specific tasks.

GPT-4o consistently achieves high accuracy and F1-scores across 75% of the dataset splits, often outperforming other models, including Claude 3.5 Sonnet, in key metrics. It performs particularly well in datasets that include products with complex or ambiguous attributes. Its high accuracy and competitive F1-scores across splits demonstrate its capacity for generalization. This model appears to be particularly resilient when dealing with challenging classification tasks, maintaining competitive precision while achieving strong recall. This balance makes GPT-4o a robust choice for scenarios that demand consistent performance across diverse product categories.

Claude 3.5 Sonnet performed competitively across many datasets, showcasing a strong balance of accuracy, precision, recall, and F1-scores. In some datasets (1 and 5), it outperformed other models, demonstrating its ability to classify products reliably while maintaining a low rate of false positives and false negatives. While not consistently the best, its performance suggests a solid ability to generalize across varied datasets, an important trait for e-commerce contexts with nuanced product categories.

GPT-4o mini, while slightly less effective than its larger counterpart, offered a commendable trade-off between performance and computational efficiency. The reduced scale of this model translated into slightly lower accuracy and F1-scores, particularly in datasets with greater category diversity. However, its performance was still strong enough to make it a viable option for platforms prioritizing speed and resource efficiency over marginal gains in classification accuracy.

Claude 3.5 Haiku, in contrast, showed more variability in its results. While its precision was often on par with or higher than other models, its lower recall in some datasets indicates difficulty in capturing a broader range of product categories. This imbalance suggests that Claude 3.5 Haiku Prediction may be better suited for use cases where minimizing false positives is more critical than achieving high recall. However, its inconsistent performance across datasets highlights potential limitations in handling diverse or complex product types.

Overall, these distinctions underline the importance of aligning model selection with specific operational needs. Claude 3.5 Sonnet and GPT-4o are well-suited for high-stakes scenarios requiring accuracy and consistency, while GPT-4o mini and Claude 3.5 Haiku cater to more specialized or resource-constrained applications. These findings emphasize the varied capabilities of LLMs and their potential for driving innovation in e-commerce automation.

4.2. Dataset-by-dataset analysis

GPT-4o consistently delivered strong results, often achieving the highest accuracy and F1-scores, demonstrating a robust balance between precision and recall. Claude 3.5 Sonnet also performed competitively, particularly in datasets where minimizing false positives was critical. Meanwhile, GPT-4o mini and Claude 3.5 Haiku generally lagged behind in both precision and recall, although they displayed specific strengths in certain contexts. An analysis of the models' performance across individual dataset splits is presented below.

In the first dataset split, Claude 3.5 Sonnet achieved the best performance, with an accuracy of 67.58% and an F1-score of 0.6687. Its high precision of 0.7385 made it reliable for correctly identifying categories with minimal false positives, although its recall of 0.6758 indicated some missed classifications. GPT-4o closely followed with an accuracy of 65.16% and a balanced F1-score of 0.6369, reflecting a good compromise between precision (0.6738) and recall (0.6516). In contrast, Claude 3.5 Haiku struggled in this split, with the lowest recall (0.6371) impacting its F1-score.

The second split revealed GPT-4o as the top performer, achieving the highest accuracy of 76.77% and an F1-score of 0.7608. Its exceptional precision of 0.8026, combined with a strong recall of 0.7677, highlights its ability to generalize effectively across diverse categories while maintaining a low false positive rate. Claude 3.5 Sonnet, while delivering a solid accuracy of 74.35%, achieved a lower F1-score of 0.7328, largely due to its slightly reduced recall (0.7435), indicating some challenges in identifying all relevant categories. Interestingly, GPT-4o Mini demonstrated better performance in terms of F1-score (0.7368) compared to Claude 3.5 Sonnet, despite achieving a slightly lower accuracy of 74.19%. Its comparable precision of 0.7873 to Claude 3.5 Sonnet suggests a strong ability to reduce false positives, but its superior balance between precision and recall allowed GPT-4o Mini to edge out in F1 performance.

In the third split, GPT-4o again led the models with an impressive accuracy of 79.84% and an F1-score of 0.7948. Its high precision (0.8346) and balanced recall (0.7984) underscored its strong generalization capabilities. Claude 3.5 Sonnet followed closely, achieving an F1-score of 0.7827 and high precision (0.8183), but with a slightly reduced recall of 0.7952. Claude 3.5 Haiku and GPT-4o mini both demonstrated solid accuracy and precision but showed relatively weaker recall, leading to lower F1-scores.

Table 5
MAE evaluation across subsets and models.

Subsets	gpt-4o	gpt-4o-mini	claude-3.5-sonnet	claude-3.5-haiku
1	2.41	2.40	1.89	2.47
2	1.18	1.9	1.46	1.31
3	2.93	2.49	2.94	2.45
4	2.71	2.89	3.38	2.88
5	1.89	2.88	1.70	3.44
6	1.89	2.43	2.09	2.80
7	2.11	2.10	2.44	2.29
8	1.74	2.39	2.34	2.26

The fourth split highlighted a more modest performance across models. GPT-4o achieved the highest accuracy at 73.55% and an F1-score of 0.7284. Claude 3.5 Sonnet showed competitive precision (0.7917), but its recall of 0.7274 affected its ability to classify all relevant categories effectively. Both Claude 3.5 Haiku and GPT-4o mini experienced notable drops in recall, which impacted their F1-scores.

In the fifth split, Claude 3.5 Sonnet stood out with the best accuracy of 85.81% and an F1-score of 0.8569. Its precision of 0.8822 and recall of 0.8581 highlighted its strong ability to minimize false positives while identifying most relevant instances. GPT-4o also performed well, with an accuracy of 83.71% and an F1-score of 0.8361, reflecting a balanced performance. The other models, while showing respectable precision, demonstrated reduced recall, leading to lower F1-scores.

In the sixth split, GPT-4o maintained its strong performance, achieving an accuracy of 79.35% and an F1-score of 0.7891. Its precision (0.8289) and recall (0.7935) confirmed its consistent reliability. Claude 3.5 Sonnet and Claude 3.5 Haiku also performed well but displayed lower recall compared to GPT-4o Prediction, affecting their overall scores.

In the seventh split, GPT-4o again led with an accuracy of 76.77% and an F1-score of 0.7560. Claude 3.5 Sonnet delivered similar accuracy but showed slightly reduced precision, which impacted its F1-score. Claude 3.5 Haiku struggled in this split, with a lower recall of 0.7032, making it the weakest performer.

Finally, in the eighth split, GPT-4o achieved the highest accuracy (78.06%) and an F1-score of 0.7827, reinforcing its status as the top model for most datasets. Claude 3.5 Sonnet showed competitive performance, with an accuracy of 74.84% and an F1-score of 0.7549, driven by strong precision but slightly lower recall. The other models, while demonstrating decent performance, continued to lag behind GPT-4o and Claude 3.5 Sonnet in overall metrics.

Overall, the GPT-4o model consistently demonstrated robust performance across diverse datasets, balancing precision and recall effectively. Claude 3.5 Sonnet was a close contender in many splits, particularly excelling in precision, while Claude 3.5 Haiku and GPT-4o mini showed moderate but less consistent performance.

4.3. Evaluation of model performance using MAE

The mean absolute error (MAE) is a metric used to measure the average magnitude of errors in predictions, without considering their direction. It calculates the absolute difference between predicted and actual values, then averages these differences across all predictions (Haque et al., 2024). In classification tasks, MAE can be a useful proxy for evaluating model performance by treating class labels as ordinal or numeric. A lower MAE indicates that the model's predictions are closer to the actual values, reflecting better accuracy and reliability (see Figs. 3 and 4).

As depicted in Table 5, which presents the MAEs for each subset and each model, the results reveal distinct patterns in model performance. GPT-4o demonstrates relatively consistent accuracy, with MAEs ranging from 1.18 to 2.93, achieving its best performance on Subset 2 (1.18). GPT-4o-mini, while similar in behavior to GPT-4o, generally exhibits higher MAEs (1.9 to 2.89). In contrast, Claude-3.5-sonnet shows a broader range of performance (1.7 to 3.38), excelling on Subset 5 (1.7) but performing less consistently overall. Claude-3.5-haiku tends to produce higher MAEs, peaking at 3.44 on Subset 5. These results reaffirm the consistent edge of GPT-4o over other models in the majority of subsets. GPT-4o achieves the lowest MAEs in several cases, such as Subset 2 (1.18) and Subset 8 (1.74), demonstrating superior accuracy and reliability across diverse data. While Claude-3.5-sonnet occasionally matches or outperforms GPT-4o in specific subsets (e.g., Subset 5), its overall performance is less consistent. GPT-4o-mini and Claude-3.5-haiku exhibit higher error rates, further emphasizing GPT-4o's advantage in zero-shot predictions for product categorization.

4.4. Key influential factors in model predictions

In our analysis of how each model generates predictions, we concentrated on identifying the most influential words that drive the model's decision-making for each product category. These influential words are key elements in determining how the models interpret product titles and categorize them accordingly. By examining these words, we aimed to uncover the underlying factors that shape each model's predictions, such as specific keywords or linguistic patterns that the models prioritize when making their classification decisions. Fig. 5 illustrates an example of how the influential words differ between GPT-4o and Claude-3.5 Sonnet for a single category, highlighting the variations in how each model weighs certain terms and the resulting impact on their predictions. This comparison provides valuable insights into the models' distinct approaches to zero-shot product categorization.

Based on the findings from the top-influential words technique applied to each of the subsets and models, we draw several valuable conclusions. To maintain brevity, we focus on presenting the results for dataset splits 2 and 5 in the following analysis.

4.4.1. Analysis of top influential words in subset 2

The GPT-4o, Claude 3.5 Sonnet, GPT-4o Mini, and Claude 3.5 Haiku models showed remarkable consistency in their predictions, particularly in product categorization, though each model emphasized slightly different terms. Across several product categories, the models displayed a strong ability to recognize key classification terms. For example, in the "beauty tools accessories" category, all models highlighted words like "makeup", "travel", and "brush", suggesting a shared understanding of essential product characteristics. However, there were slight nuances in the models' predictions. The Claude models (Sonnet and Haiku) also included "bottle" among their top influential words, whereas the GPT models placed more emphasis on terms like "extension" and "eyelash".

In categories with more technical products, such as "computer components", the models consistently identified key technical terms like "fan", "case", and "thermal", demonstrating their ability to interpret complex product specifications. Furthermore, when classifying products in the "dog supplies" category, the models consistently identified terms like "dog", "pet", and "toy" as influential, showing a strong alignment across the models. Overall, despite these subtle differences, the models exhibited a similar approach to word weighting, indicating a robust classification methodology capable of accurately categorizing products based on their most distinctive features.

4.4.2. Analysis of top influential words in subset 5

In the analysis of the influential words for Subset 5, the models demonstrated different strengths and strategies in their predictions.

- GPT-4o showed a balanced approach, emphasizing a diverse vocabulary across categories. For example, it highlighted terms such as "travel" (for luggage), "men" (for accessories), and "toy" (for educational toys), showcasing a strong contextual understanding and the ability to recognize category-specific descriptors.

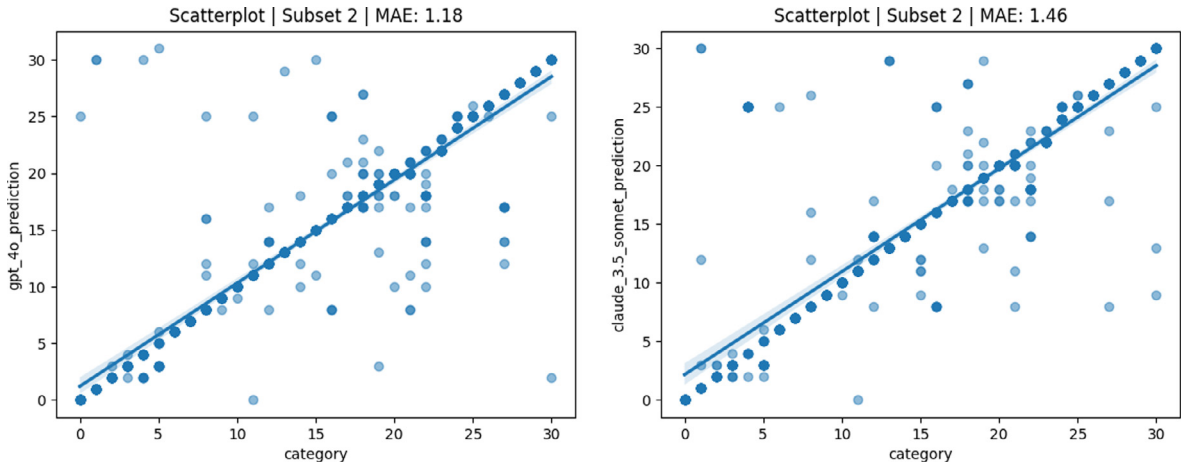

Fig. 3. Scatterplot comparing GPT-4o and Claude-3.5-Sonnet predictions for subset 2.

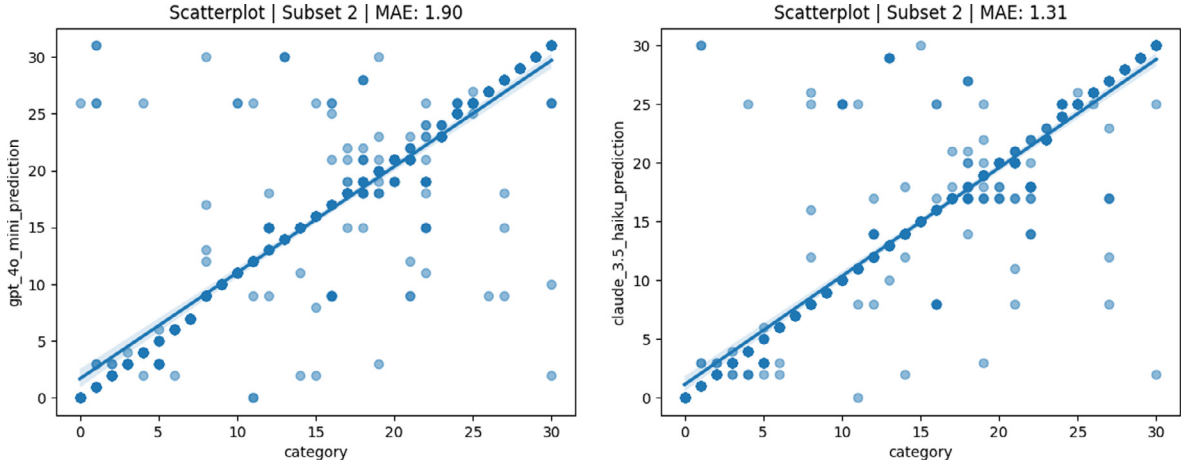

Fig. 4. Scatterplot comparing GPT-4o-mini and Claude-3.5-Haiku predictions for subset 2.

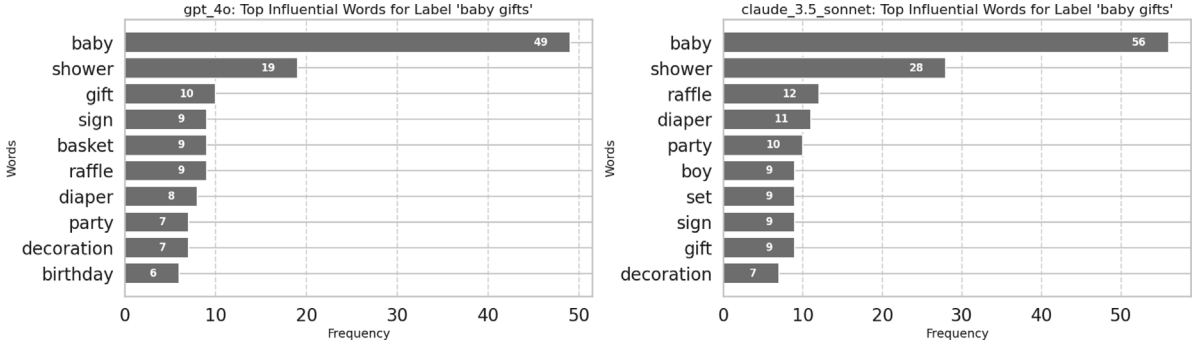

Fig. 5. Comparison of influential words in GPT-4o and Claude-3.5-Sonnet predictions.

- Claude 3.5 Sonnet displayed a more refined prediction strategy, with a greater focus on precise, detailed terms. It excelled at identifying specific product features, such as "waterproof" for watches, "expandable" for luggage sets, and "joycon" for Nintendo Switch accessories. These words suggest that Claude 3.5 Sonnet is particularly adept at capturing nuanced product details and technical specifications, making it well-suited for more complex categorizations.
- Claude 3.5 Haiku, in contrast, took a slightly broader approach in its word selection, with a focus on both functional and descriptive terms. Words like "antitheft" (for laptop bags) and "renewed" (for gaming consoles) stood out, suggesting the model prioritizes a mix of product functionality and descriptive features when making predictions.
- GPT-4o Mini displayed a more generalized prediction approach, often focusing on broader, less specific terms like "black", "compatible", and "replacement". While still relevant to the respective categories, these words indicated a more flexible, yet less specialized, prediction strategy. This generalized approach may make GPT-4o Mini adaptable across different categories but somewhat less precise compared to the other models.

Overall, the differences in word selection across these models reveal their unique strategies for understanding and categorizing product titles. While some models emphasize technical details and category-specific terms, others focus on broader descriptors, offering valuable insights into how each model interprets and classifies product features.

4.5. Performance vs. Cost efficiency

When evaluating the models' performance, the cost of generating predictions for 8 datasets, each with 31 categories across 20 labeled products, provides critical insights into their cost-effectiveness. GPT-4o, the top-performing model, incurs a cost of $4.38, reflecting a balance between performance and affordability. In contrast, GPT-4o Mini, while slightly trailing in accuracy, demonstrates exceptional cost efficiency at just $0.25, making it an attractive option for large-scale deployments. Claude 3.5 Sonnet, despite its commendable accuracy, is the most expensive model at $6.18, raising concerns about scalability for cost-sensitive applications. Claude 3.5 Haiku strikes a middle ground at $2.06, offering a more economical alternative with moderate performance. These variations in cost emphasize the trade-offs between performance, scalability, and budget considerations when selecting a model for production.

5. Discussion

5.1. Implications of the study for E-commerce automation

The results from the dataset analysis suggest that LLMs like GPT-4o, GPT-4o mini, and Claude 3.5 Sonnet can significantly enhance the automation of product categorization in e-commerce. The ability to automatically classify products into categories is crucial for e-commerce platforms, as it facilitates better inventory management, improved customer search experiences, and more accurate product recommendations. However, different platforms and applications have unique needs, and the choice of model will depend on the specific operational priorities of the platform.

Addressing both research questions 1 and 2 presented in the introduction, GPT-4o stands out as the optimal choice for e-commerce platforms that prioritize high accuracy and consistency. This model consistently delivers exceptional accuracy and F1-scores across various datasets, making it an ideal solution for platforms that require precise and reliable product categorization. Accurate categorization is crucial not only for enhancing the user experience — ensuring customers can effortlessly find what they are looking for — but also for streamlining backend operations like reporting and inventory management. With its robust performance across diverse product categories, GPT-4o can efficiently handle a wide range of products, from electronics to apparel. By integrating GPT-4o, these platforms can minimize errors and significantly boost the overall efficiency of their product categorization processes.

In contrast, for e-commerce systems that face resource constraints — whether due to limited budget or the need for faster processing — GPT-4o mini provides an optimal solution. Although this model delivers slightly lower performance compared to the full GPT-4o, it offers a favorable balance of efficiency and effectiveness. It achieves competitive accuracy and precision while requiring less computational overhead, making it ideal for platforms that need to process large amounts of data quickly without the high costs associated with more powerful models. For example, smaller or mid-sized platforms that handle a high volume of products but may not have the budget to support the full GPT-4o model can use GPT-4o mini to maintain a good level of performance without overburdening their systems.

Platforms or applications that specifically aim to minimize false positives, such as those where incorrect categorization could lead to customer dissatisfaction or inventory errors, should consider integrating Claude 3.5 Sonnet model. This model excels in precision, meaning it is highly effective at correctly classifying products while minimizing the risk of misclassifying items as irrelevant or incorrectly categorized. However, as noted, it slightly lags in recall, meaning it may miss some products that belong to a particular category. For applications where it is more critical to avoid false positives than to maximize recall, this model provides a tailored solution. For example, in a scenario where products must be grouped into highly specialized categories, or when the cost of a misclassification is high (e.g., premium products), Claude 3.5 Sonnet would be advantageous.

In summary, each model has distinct strengths and weaknesses. GPT-4o shines with its robustness, high accuracy, and well-balanced precision and recall, making it the top performer overall. GPT-4o mini offers a more efficient alternative but sacrifices a bit of recall and F1-score in exchange for reduced computational overhead. Claude 3.5 Sonnet excels in precision and consistency but falls slightly short in recall compared to GPT-4o. Meanwhile, Claude 3.5 Haiku performs well in certain cases but struggles with recall and shows more variability across different datasets.

5.2. Challenges and limitations of LLMs in zero-shot classification

Addressing research question 3 presented in the introduction, the results demonstrated the impressive capabilities of LLMs in zero-shot product classification. However, several challenges and limitations remain, particularly when scaling these models for e-commerce applications. One key constraint encountered during our evaluation was token limitations, especially with Claude models, which impose restrictions of 50 requests per minute (RPM) and a daily cap of 3,500,000 tokens under Tier 1 usage. These operational constraints can hinder high-throughput tasks, necessitating careful resource management or tier upgrades for large-scale deployments.

Beyond operational limitations, challenges inherent to zero-shot classification also warrant attention. LLMs can struggle with nuanced distinctions between product categories, especially in domains requiring specialized knowledge or when dealing with overlapping or ambiguous category definitions. Additionally, while no significant inconsistencies were observed in this evaluation, the potential for biases in model predictions remains a concern, particularly if the pre-training data used by the models does not adequately represent diverse product types or cultural contexts.

Overall, while LLMs have proven to be powerful and versatile tools for zero-shot classification, their deployment in real-world applications requires careful and strategic consideration. These models offer the advantage of being able to classify data without needing task-specific training, which is especially valuable in dynamic environments where new categories and data types frequently emerge. However, their practical use also entails navigating several operational constraints and inherent challenges that can impact their effectiveness. These include issues such as processing limitations, the need for efficient resource management, and the complexity of ensuring consistent performance across various domains and use cases.

Moreover, fairness remains a crucial consideration, as LLMs can inadvertently inherit biases present in the training data, leading to skewed or inequitable results in certain applications. Therefore, it is essential to ensure that the model's output is both reliable and unbiased, particularly in contexts where decisions may affect individuals or communities. To overcome these challenges, a nuanced deployment strategy is necessary, one that accounts for both the limitations of the model and the specific requirements of the application.

In addition to these foundational considerations, there is little doubt that fine-tuning the model with targeted, labeled data can substantially improve its accuracy and effectiveness for specific tasks. By adjusting the model to better align with the unique characteristics and needs of a particular dataset or application, fine-tuning can enhance the model's ability to deliver highly accurate and relevant classifications. This tailored approach is particularly beneficial for specialized scenarios,

such as those involving industry-specific terminology, complex product categorization, or nuanced sentiment analysis, where generic zero-shot capabilities might fall short. Ultimately, fine-tuning can help bridge the gap between the broad capabilities of LLMs and the specific, high-performance needs of real-world tasks, leading to more reliable and efficient outcomes.

5.3. Practical considerations for LLM-based API integration in E-commerce

Addressing research question 4, integrating these insights into API-based software solutions can allow e-commerce platforms to select and implement the model that best aligns with their operational goals. For instance, a platform focused on accuracy may configure the API to default to GPT-4o for its categorization tasks. Meanwhile, a platform with limited budget could be configured to leverage GPT-4o mini, ensuring that it operates efficiently without compromising too much on performance. By providing access to different models, the API can offer flexibility, allowing e-commerce businesses to optimize product categorization according to their specific needs. This approach ensures scalable, adaptable, and practical solutions that cater to a wide range of real-world e-commerce requirements, from large-scale platforms to resource-conscious enterprises. In this way, the deployment of LLMs for e-commerce automation becomes a key enabler of operational efficiency and customer satisfaction.

6. Conclusion

This study highlights the transformative potential of LLMs for zero-shot product classification in e-commerce, offering an innovative solution to the challenges of traditional, manual classification methods. By evaluating four state-of-the-art LLMs — GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku — across a diverse dataset of 248 product categories, we demonstrated that LLMs can achieve impressive classification accuracy without prior fine-tuning. The results showed that LLMs are capable of accurately categorizing products into predefined classes across various datasets, with models like GPT-4o and Claude 3.5 Sonnet leading the way in performance.

The findings of this study aim to fill a critical knowledge gap in the domain of e-commerce automation. While previous research has explored LLMs in general-purpose natural language processing tasks, their utility in specialized domains, such as e-commerce product classification, remains underexplored. By focusing on zero-shot classification, this research eliminates the reliance on labeled training data, making the findings applicable to platforms with limited resources for data annotation.

The ability of LLMs to classify products in a zero-shot setting opens new possibilities for e-commerce automation, streamlining workflows and reducing the need for extensive labeled data. Our integration of the top-performing models into an API software further emphasizes their potential for real-world applications, enhancing operational efficiency and scalability. In conclusion, LLMs represent a powerful tool for revolutionizing e-commerce product classification, offering significant automation benefits and scalability for future e-commerce systems.

CRediT authorship contribution statement

Konstantinos I. Roumeliotis: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nikolaos D. Tselikas:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dimitrios K. Nasiopoulos:** Writing – review & editing, Validation, Supervision, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Anthropic PBC, 2024a. Automatically generate first draft prompt templates - Anthropic. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator.

Anthropic PBC, 2024b. Models - Anthropic. https://docs.anthropic.com/en/docs/about-claude/models#model-comparison-table.

Asaniczka, 2023. Amazon products dataset 2023 (1.4M products). https://www.kaggle.com/datasets/asaniczka/amazon-products-dataset-2023-1-4m-products.

Chen, J., Ma, L., Li, X., Thakurdesai, N., Xu, J., Cho, J.H.D., Nag, K., Korpeoglu, E., Kumar, S., Achan, K., 2023. Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in E-commerce with LLMs. https://arxiv.org/abs/2305.09858v1.

Chen, J., Ma, L., Li, X., Xu, J., Cho, J.H.D., Nag, K., Korpeoglu, E., Kumar, S., Achan, K., 2024. Relation labeling in product knowledge graphs with large language models for e-commerce. Int. J. Mach. Learn. Cybern. 15 (12), 5725–5743. http://dx.doi.org/10.1007/S13042-024-02274-5/METRICS.

Cheng, Z., Zhang, W., Chou, C.-C., Jau, Y.-Y., Pathak, A., Gao, P., Batur, U., 2024. E-commerce product categorization with LLM-based dual-expert classification paradigm. In: Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, Or Individual (CustomNLP4U). pp. 294–304. http://dx.doi.org/10.18653/V1/2024.CUSTOMNLP4U-1.22.

Chhabra, D., Sah, S., Rani, R., Bisla, N., Tomar, A., Sharma, A., 2024. Deep learning-based sentiment analysis of amazon product reviews. In: 2024 1st International Conference on Trends in Engineering Systems and Technologies. ICTEST 2024, http://dx.doi.org/10.1109/ICTEST60614.2024.10576180.

D'Asaro, F., De Luca, S., Bongiovanni, L., Rizzo, G., Papadopoulos, S., Schinas, M., Koutlis, C., 2024. Zero-shot content-based crossmodal recommendation system. Expert Syst. Appl. 258, 125108. http://dx.doi.org/10.1016/J.ESWA.2024.125108.

Davoodi, L., Mezei, J., 2024. A large language model and qualitative comparative analysis-based study of trust in E-commerce. Appl. Sci. (Switzerland) 14 (21), 10069. http://dx.doi.org/10.3390/APP142110069/S1.

Fang, C., Li, X., Fan, Z., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., Achan, K., 2024. LLM-ensemble: Optimal large language model ensemble method for E-commerce product attribute value extraction. In: SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2910–2914. http://dx.doi.org/10.1145/3626772.3661357.

Farfade, S., Vernekar, S., Chaoji, V., Mukherjee, R., 2024. Scaling use-case based shopping using LLMs. In: WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining. pp. 1165–1166. http://dx.doi.org/10.1145/3616855.3635748.

Gao, D., Chen, K., Chen, B., Dai, H., Jin, L., Jiang, W., Ning, W., Yu, S., Xuan, Q., Cai, X., Yang, L., Wang, Z., 2024. LLMs-based machine translation for E-commerce. Expert Syst. Appl. 258, 125087. http://dx.doi.org/10.1016/J.ESWA.2024.125087.

Ghaffari, S., Yousefimehr, B., Ghatee, M., 2024. Generative-AI in E-commerce: Use-cases and implementations. In: 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing. AISP 2024, http://dx.doi.org/10.1109/AISP61396.2024.10475266.

Haque, M.A., Nahin, K.H., Nirob, J.H., Ahmed, M.K., Sawaran Singh, N.S., Paul, L.C., Algarni, A.D., ElAffendi, M., Ateya, A.A., 2024. Machine learning-based technique for directivity prediction of a compact and highly efficient 4-port MIMO antenna for 5G millimeter wave applications. Results Eng. 24, 103106. http://dx.doi.org/10.1016/J.RINENG.2024.103106.

Huang, X., Peng, F., Schuurmans, D., Cercone, N., Robertson, S.E., 2003. Applying machine learning to text segmentation for information retrieval. Inf. Retr. 6 (3–4), 333–362. http://dx.doi.org/10.1023/A:1026028229881/METRICS.

Ihsanoğlu, A., Zaval, M., Yıldız, O.T., Comparison of machine learning algorithms and large language models for product categorization. In: 32nd IEEE Conference on Signal Processing and Communications Applications, SIU 2024 - Proceedings. http://dx.doi.org/10.1109/SIU61531.2024.10600809.

Katlariwala, M.Z., Gupta, A., 2024. Product recommendation system using large language model: Llama-2. In: 2024 IEEE 5th World AI IoT Congress. AIIoT, pp. 491–494. http://dx.doi.org/10.1109/AIIOT61789.2024.10579009.

Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H.T., Xie, P., Huang, F., Jiang, Y., 2024. EcomGPT: Instruction-tuning large language models with chain-of-task tasks for E-commerce. Proc. AAAI Conf. Artif. Intell. 38 (17), 18582–18590. http://dx.doi.org/10.1609/AAAI.V38I17.29820.

Liu, Y., Zhang, W.N., Chen, Y., Zhang, Y., Bai, H., Feng, F., Cui, H., Li, Y., Che, W., 2023. Conversational recommender system and large language model are made for each other in E-commerce pre-sales dialogue. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 9587–9605. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.643.

López Espejel, J., Ettifouri, E.H., Yahaya Alassan, M.S., Chouham, E.M., Dahhane, W., 2023. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. Nat. Lang. Process. J. 5, 100032. http://dx.doi.org/10.1016/J.NLP.2023.100032.

Munkova, D., Hajek, P., Munk, M., Skalka, J., 2020. Evaluation of machine translation quality through the metrics of error rate and accuracy. Procedia Comput. Sci. 171, 1327–1336. http://dx.doi.org/10.1016/J.PROCS.2020.04.142.

Nananukul, N., Sisaengsuwanchai, K., Kejriwal, M., 2024. Cost-efficient prompt engineering for unsupervised entity resolution in the product matching domain. Discov. Artif. Intell. 4 (1), 1–21. http://dx.doi.org/10.1007/S44163-024-00159-8/FIGURES/12.

Nguyen, C., Carrion, D., Badawy, M., 2024. Comparative performance of claude and GPT models in basic radiological imaging tasks. MedRxiv http://dx.doi.org/10.1101/2024.11.16.24317414, 2024.11.16.24317414.

Oh, Sejun, Yoon, Jungeun, Chung, Yoojin, Cho, Yoonjoo, Shim, Hyosup, Kwon, Oh Nam, 2024. Analysis of generative AI's mathematical problem-solving performance: Focusing on ChatGPT 4, Claude 3 Opus, and Gemini Advanced. Math. Educ. 63 (3), 549–571. http://dx.doi.org/10.7468/MATHEDU.2024.63.3.549.

OpenAI, Inc., 2024a. Models - OpenAI API - GPT-4o. https://platform.openai.com/docs/models/gpt-4o#gpt-4o.

OpenAI, Inc., 2024b. Models - OpenAI API - GPT-4o-mini. https://platform.openai.com/docs/models/gpt-4o#gpt-4o-mini.

OpenAI, Inc., 2024c. OpenAI's GPT-4o mini offers big performance at a small price. https://www.deeplearning.ai/the-batch/openais-gpt-4o-mini-offers-big-performance-at-a-small-price/.

Palen-Michel, C., Wang, R., Zhang, Y., Yu, D., Xu, C., Wu, Z., 2024. Investigating LLM applications in E-commerce. https://arxiv.org/abs/2408.12779v1.

PythonAnywhere LLP, 2024. Host, run, and code python in the cloud: PythonAnywhere. https://www.pythonanywhere.com/.

Rahman, M., Khatoonabadi, S., Abdellatif, A., Shihab, E., 2024. Automatic detection of LLM-generated code: A case study of claude 3 haiku. https://arxiv.org/abs/2409.01382v1.

Roumeliotis, 2024a. evaluation-results.json · GitHub. https://github.com/Applied-AI-Research-Lab/Battle-of-LLM-Giants-GPT-vs.-Claude-Models-for-Zero-Shot-Classification-in-Ecommerce-Automation/blob/main/Datasets/split_datasets/evaluation-results.json.

Roumeliotis, 2024b. GitHub - applied-AI-research-lab/battle-of-LLM-giants. https://github.com/Applied-AI-Research-Lab/Battle-of-LLM-Giants-GPT-vs.-Claude-Models-for-Zero-Shot-Classification-in-Ecommerce-Automation.

Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K., 2023. Unveiling sustainability in ecommerce: GPT-powered software for identifying sustainable product features. Sustainability 15 (15), 12015. http://dx.doi.org/10.3390/SU151512015.

Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K., 2024a. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. Nat. Lang. Process. J. 6, 100056. http://dx.doi.org/10.1016/J.NLP.2024.100056.

Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K., 2024b. Precision-driven product recommendation software: Unsupervised models, evaluated by GPT-4 LLM for enhanced recommender systems. Software 3 (1), 62–80. http://dx.doi.org/10.3390/SOFTWARE3010004.

Shahin, M., Chen, F.F., Maghanaki, M., Hosseinzadeh, A., 2024. Adapting the GPT engine for proactive customer insight extraction in product development. Manuf. Lett. 41, 1376–1385. http://dx.doi.org/10.1016/J.MFGLET.2024.09.164.

Soviero, B., Kuhn, D., Salle, A., Moreira, V.P., 2024. ChatGPT goes shopping: LLMs can predict relevance in eCommerce search. LNCS, In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14611, pp. 3–11. http://dx.doi.org/10.1007/978-3-031-56066-8_1.

Wang, G., Ni, X., Shen, Q., Yang, M., 2024. Leveraging large language models for context-aware product discovery in E-commerce search systems. J. Knowl. Learn. Sci. Technol. (ISSN: 2959-6386) 3 (4), 300–312. http://dx.doi.org/10.60087/JKLST.V3.N4.P300, (Online).

WangHaixun, NaTaesik, 2024. Rethinking E-commerce search. ACM SIGIR Forum 57 (2), 1–19. http://dx.doi.org/10.1145/3642979.3643007.

Wu, Y., Feng, Y., Wang, J., Zhou, W., Ye, Y., Xiao, R., Xiao, J., 2024. Hi-Gen: Generative retrieval for large-scale personalized E-commerce search. https://arxiv.org/abs/2404.15675v2.

Xu, X., Wu, Y., Liang, P., He, Y., Wang, H., 2024b. Emerging synergies between large language models and machine learning in ecommerce recommendations. https://arxiv.org/abs/2403.02760v2.

Xu, W., Xiao, J., Chen, J., 2024a. Leveraging large language models to enhance personalized recommendations in E-commerce. https://arxiv.org/abs/2410.12829v1.

Yao, J., Shepperd, M., 2021. The impact of using biased performance metrics on software defect prediction research. Inf. Softw. Technol. 139, 106664. http://dx.doi.org/10.1016/J.INFSOF.2021.106664.

Zhang, S., Peng, B., Zhao, X., Hu, B., Zhu, Y., Zeng, Y., Hu, X., 2024. LLaSA: Large language and E-commerce shopping assistant. https://arxiv.org/abs/2408.02006v1.

Zhou, J., Liu, B., Hong, J.N.A.Y., Lee, K., Wen, M., 2023. Leveraging large language models for enhanced product descriptions in ecommerce. https://arxiv.org/abs/2310.18357v1.