University of Athens
Department of Informatics and Telecommunications

E-commerce literature review

Student name: *Dimitris Tsiompikas*
AM: 7115112300036

Course: *E-commerce technologies (M165)*
Semester: *Summer Semester 2024*

Contents

1. Abstract

E-commerce has seen increased usage of AI and ML models in the last few years due to the advantages they provide in product review evaluation, sentiment analysis on product reviews, customer satisfaction and many more fields. This literature review explores the application of Large Language Models (LLMs) in the e-commerce sector by comparing and contrasting six recent studies.

The reviewed papers investigate various aspects of LLM deployment, including the creation of an instruction dataset which can be used for a multitude of e-commerce ML applications, LLM comparison for product review sentiment analysis and creation of product descriptions by LLMs to enhance user engagement. Through a comparative analysis of methodologies, findings, and theoretical frameworks, this review identifies common themes and divergent approaches in current research.

The results highlight the significant potential of LLMs to revolutionize e-commerce operations by increasing efficiency and customer satisfaction. However, the studies also reveal challenges such as ethical considerations and the need for extensive computational resources. This review concludes with a discussion of the implications for practitioners and researchers, and proposes directions for future research to address existing gaps in the literature.

2. Introduction

2.1. Brief overview of the topic

The e-commerce industry has experienced exponential growth over the past decade, driven by advancements in technology and changes in consumer behavior. With the increasing complexity of online retail operations and the demand for personalized customer experiences, businesses are continuously seeking innovative solutions to enhance their services and operational efficiency. One such groundbreaking innovation is the application of Large Language Models (LLMs).

LLMs, exemplified by models like GPT-3, Llama and Gemini, are sophisticated AI systems designed to process and generate human-like text based on vast amounts of data. These models have demonstrated exceptional capabilities in understanding context, generating coherent responses, and performing a variety of language-based tasks. In the context of e-commerce, some fields that LLMs can be employed are: automation of customer support, personalization of marketing efforts, optimization of search functionalities, enhancement of recommendation systems and product review sentiment analysis.

The integration of LLMs in e-commerce promises numerous benefits, including improved customer satisfaction through more responsive and personalized interactions, increased operational efficiency by automating routine tasks, and enhanced decision-making with advanced data analysis. However, the deployment of LLMs also presents challenges, such as ethical considerations related to data privacy and the need for significant computational resources.

2.2. Purpose of the review

This literature review aims to explore the current state of research on LLM usage in e-commerce by comparing six research papers which apply LLMs to different e-commerce ML tasks. Through this analysis, I seek to provide insights into the practical applications, benefits, and limitations of LLMs in the e-commerce domain, contributing to a deeper understanding of their potential impact and guiding future research directions.

To further expand on this topic, the purpose of this literature review is to synthesize and critically evaluate recent research on the application of Large Language Models (LLMs) in the e-commerce industry. By comparing and contrasting these six papers, this review aims to achieve the following objectives:

1. **Understand the Practical Applications**: Examine how LLMs are being utilized in various e-commerce functions such as customer service, personalized marketing, and recommendation systems.
2. **Evaluate Methodologies**: Analyze the different research methodologies employed in these studies to understand their approaches, strengths, and limitations.
3. **Compare Findings and Implications**: Identify common themes and divergences in the findings of these studies, and discuss the implications of these results for the e-commerce industry.
4. **Identify Challenges and Opportunities**: Highlight the challenges faced in the implementation of LLMs in e-commerce, including ethical considerations and technical constraints, and explore the opportunities for further innovation and research in this field.

2.3. Paper overview

In this review, I will compare the findings, methodologies and conclusions of 3 papers regarding the usage of LLMs for e-commerce purposes.

The first paper [13] involves the usage of LLMs for product review sentiment analysis, in order to observe their performance and find out if they would provide useful results for customer satisfaction understanding. The models showed promising results and the authors stated that there is still further research to be done in regards to improving the models' accuracy and get a clearer understanding of the customers' satisfaction.

The second paper [11] studies LLMs being used for generalized e-commerce purposes such as providing valid results for new products or new users (the same problem that applies to recommendation systems). In order to achieve this, the authors created an instruction dataset called ECInstruct and a series of LLMs under the name eCeLLM, which are specifically designed for this purpose. Results showed that these models outperformed the currently used models and there is further work ahead with more data to add in the ECInstruct dataset and more user profiling data.

The third paper [7] uses LLMs for the creation of accurate product descriptions and keywords for e-commerce product listings based on images (image recognition) and metadata analysis with a few examples being item size, color and weight. The authors also created a dataset from a website called Taobao, one of the biggest online shopping platforms which is owned by Alibaba, and they named it TaoDescribe. Results showed very good performance with adequate descriptions for the products and the LLMs being able to properly recognize each product distinctly from each other, while also paving the way for future research in this particular topic by expanding the TaoDescribe dataset and the paper's limitations.

The fourth paper [9] leverages a fine-tuned LLM on Click-Through-rate prediction for product advertisements in e-commerce. This paper is also of important significance since it has been cited 34 times according to Google Scholar. The model consists of two BERT-like encoders (hence the name CTR-BERT) that match specific conditions that will be described more thoroughly in the paper's summary. It surpassed currently used CTR techniques both in offline and online experiments.

The fifth paper [14] utilizes a Llama model from Meta, fine-tuned on an instruction set created by domain experts, for the task of e-commerce authoring. E-commerce authoring refers to the process of creating and managing content for online stores. This involves a range of activities aimed at ensuring that the digital storefront is appealing, informative, and effective in converting visitors into customers. Certain e-commerce authoring tasks which are also used in the instruction set include: advertisement creation, query-enhanced product title rewriting, product classification, purchase intent speculation, and general Q&A. Results showed that this model outperforms currently used techniques in both fine-tuned experiments and zero-shot evaluation.

The sixth and final paper [2] employs a specific form of BERT in order to improve search engines for e-commerce products. Using several techniques that will be discussed later on, the authors managed to obtain very good results on offline and online A/B testing experiments. This is a very recent paper, that leverages the largest B2B e-commerce platform in the world to perform its experiments, Alibaba.

3. Theoretical framework

3.1. E-commerce

E-commerce , since its beginnings in 1995 has generated a huge amount of revenue for large and small businesses alike and has changed the landscape of markets forever. The rise of the internet in the 1990s paved the way for modern e-commerce, with pioneering platforms like Amazon and eBay setting the stage for today's global digital marketplace. (Information obtained from here which can be read for more e-commerce knowledge: [8])

To be more precise, E-commerce, short for electronic commerce, is the buying and selling of goods and services over the internet. It encompasses a wide range of business activities, from retail transactions between businesses and consumers (B2C) to exchanges between businesses (B2B), consumer-to-consumer sales (C2C), and even consumer-to-business transactions (C2B). E-commerce leverages digital technologies such as websites, mobile apps, social media, and electronic payment systems to facilitate transactions and improve the efficiency, convenience, and reach of commerce.

E-commerce works with business models as mentioned above and below I will describe them in a concise manner in order to get a better understanding of them:

Business-to-Consumer (B2C): Direct transactions between businesses and end-users. Examples include online retailers like Amazon.

Business-to-Business (B2B): Transactions between businesses, such as a wholesaler selling to a retailer. Alibaba is a prominent B2B platform.

Consumer-to-Consumer (C2C): Transactions between consumers, facilitated by third-party platforms like eBay.

Consumer-to-Business (C2B): Individuals selling products or services to businesses, such as freelance services on platforms like Upwork.

3.2. LLM

LLMs are a type of artificial intelligence model designed to understand and generate human language. They are trained on vast datasets comprising text from books, articles, websites, and other written materials. The training process involves feeding the model massive amounts of text data and using machine learning algorithms, particularly neural networks, to learn the patterns, structures, and nuances of language.

The journey of LLMs began with simpler models like n-grams and bag-of-words, which relied on statistical methods to understand text. These methods were limited in their ability to capture context and nuances in language.

The introduction of neural networks and recurrent neural networks (RNNs) marked a significant step forward. RNNs, with their ability to process sequences of data, brought better performance in tasks like language modeling and translation.

The game-changer in LLMs came with the introduction of the transformer architecture, introduced in the paper "Attention is All You Need" by Vaswani et al. (2017) [15]. Transformers use a mechanism called self-attention, allowing models to weigh the importance of different words in a sentence relative to each other. This architecture paved the way for models like BERT (Bidirectional Encoder Representations from Transformers) [3] and GPT (Generative Pre-trained Transformer) [16].

4. Paper summaries

4.1. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation

4.1.1. Introduction and objectives.

This paper involves the usage of LLMs in order to extract information and evaluate product reviews. According to the study, product reviews provide businesses, e-commerce platforms and product manufacturers with invaluable information regarding customer satisfaction. Questions like the customer's contentment, encountered issues during shipping or product utilization are among the inquiries e-commerce owners and manufacturers seek answers to, in order to optimize their services. [13]

The goals of the paper are two: the first is to evaluate two state-of-the-art LLMs on product review evaluation, namely GPT 3.5, developed by OpenAI, and Llama 2 which is open source and developed by Meta. The second is to address specific research questions that have not been adequately answered by prior studies, according to the authors, which are the following:

Q1: Which LLM demonstrates superior efficacy in assessing product reviews?

Q2: Among the LLMs, which exhibits superior performance after undergoing fine-tuning for the assessment of product reviews?

Q3: How significant is the process of fine-tuning LLMs for domain-specific tasks?

Q4: What is the impact of both the quantity and quality of the dataset designated for fine-tuning purposes?

Q5: To what extent does an optimized prompting engineering plan and fine-tuning contribute to achieving better and more cost-effective results from LLMs?

Q6: What impact does fine-tuning have on the generated output of LLMs?

Q7: Are LLMs, such as GPT-3.5 and LLaMA-2 models, or NLP models like BERT and RoBERTa, more effective in predicting product review ratings and, more generally, in regression and sentiment analysis tasks?

Q8: Can LLMs be effectively utilized for the evaluation of product reviews, and how can LLMs revolutionize the e-commerce sector?

4.1.2. Methodology.

Figure 1: Fine-tuning JSONL samples.

```
# Llama 2 JsonL Sample
{"prompt": "You are a product reviewer. Assign integer star ratings (between 1 and 5)"
           "to the following product reviews."
           "Return your response in JSON format like this example:"
           "{\"rating1\":integer, \"rating2\":integer, ...}."
           "Do not provide explanations or justifications for the ratings. Reviews\"\n"
           "1. almost perfect. these would be perfect if they had pockets",
 "completion": "{\"rating1\":4}"}

# GPT 3.5 JsonL Sample
{"messages": [{"role": "system", "content": "You are a product reviewer"},
              {"role": "user", "content":
                  "Assign integer star ratings (between 1 and 5)to the following product reviews."
                  "Return your response in json format like this example"
                  "{'rating1':integer,'rating2':integer,...}."
                  "Please avoid providing additional explanations. Reviews:\n"
                  "1. almost perfect. these would be perfect if they had pockets"},
              {"role": "assistant", "content": "{\"rating1\":4}"}]}
```

First of all, the authors begin by comparing papers themselves (conducting a literature review on papers written for the e-commerce field) they proceed as follows:

They employ two more LLMs which are known for good performance already since they were the state-of-the-art before the GPT architecture came through: BERT and RoBERTa.

The authors began by conducting a literature review of e-commerce-related papers. They then utilized two additional state-of-the-art LLMs, BERT and RoBERTa, which were leading models before the emergence of the GPT architecture.

They developed a web application using Python and Flask on Chrome, functioning as a web crawler to extract product reviews from prominent online marketplaces like Amazon and eBay. The extracted dataset underwent standard NLP pre-processing techniques, including the removal of special characters and accents, and converting all text to lowercase. The data was split into 70% training, 15% validation, and 15% testing sets. The code for the dataset, application, and models is available here [12].

For evaluation, the authors initially ran GPT-3.5 Turbo and Llama 2B in their base forms on the test dataset for star rating prediction using specifically created prompts. Subsequently, they fine-tuned all four LLMs with the training data and re-evaluated the models on lexical context understanding and review predictions, producing the final results of the experiment.

4.1.3. Findings.

Figure 2: Performance metrics for the LLMs

Table 1. Model performance metrics comparison.

Model	Accuracy	Precision	Recall	F1
base:`gpt-3.5-turbo-1106`	0.5483	0.553	0.5483	0.5467
base:`llama-2-70b-chat`	0.5086	0.5029	0.5086	0.5038
ft:`gpt-3.5-turbo-1106` **(100%)**	0.6424	0.6416	0.6424	0.6409
ft:`llama-2-70b-chat` **(100%)**	0.6185	0.6162	0.6185	0.6147
ft:bert-adam (100%)	0.5881	0.5881	0.5881	0.5871
ft:bert-adamw (100%)	0.5841	0.5921	0.5841	0.5766
ft:bert-sgd (100%)	0.2225	0.0706	0.2225	0.0888
ft:roberta-adam (100%)	0.6066	0.6157	0.6066	0.6017
ft:roberta-adamw (100%)	0.5536	0.5951	0.5536	0.5325
ft:roberta-sgd (100%)	0.2225	0.0495	0.2225	0.081
ft:`gpt-3.5-turbo-1106` **(50%)**	0.6212	0.6207	0.6212	0.6197
ft:`llama-2-70b-chat` **(50%)**	0.604	0.6009	0.604	0.5997
ft:bert-adam (50%)	0.5523	0.5575	0.5523	0.5532
ft:bert-adamw (50%)	0.5775	0.5741	0.5775	0.5736
ft:bert-sgd (50%)	0.1629	0.0269	0.1629	0.0462
ft:roberta-adam (50%)	0.5152	0.5305	0.5152	0.5181
ft:roberta-adamw (50%)	0.5497	0.5351	0.5497	0.5286
ft:roberta-sgd (50%)	0.1642	0.027	0.1642	0.0463

To begin with, results showed that the fine-tuned GPT 3.5 Turbo outperformed all other models with an accuracy of **64.2%** in the star ratings prediction problem. Authors state that since there are 5 numbers for the rating scale (1 to 5) the probability of a model predicting correctly is 20% so anything above that is a very good result, especially since some models hit 55% and up.

Furthermore, the authors answered all the questions that they created in the introduction. Here are the answers:

Q1: Which LLM demonstrates superior efficacy in assessing product reviews?

GPT 3.5 outperformed Llama 2B in predicting star ratings for product reviews.

Q2: Among the LLMs, which exhibits superior performance after undergoing fine-tuning for the assessment of product reviews?

Following the fine-tuning process, the GPT-3.5 model exhibits superior performance to the LLaMA-2 model by a margin of 2.39% in tasks related to predicting star ratings for product reviews.

Q3: How significant is the process of fine-tuning LLMs for domain-specific tasks?

The fine-tuned LLaMA-2 and GPT-3.5 models demonstrated a 10.99% and 9.41% increase in predictions in comparison to their base models predictions.

Q4: What is the impact of both the quantity and quality of the dataset designated for fine-tuning purposes?

The fine-tuned LLaMA-2 and GPT-3.5 models demonstrated a 1.45% and 2.12% increase in predictions when the dataset was doubled in size.

Q5: To what extent does an optimized prompting engineering plan and fine-tuning contribute to achieving better and more cost-effective results from LLMs?

Effective prompting engineering and fine-tuning of LLMs play a crucial role in achieving improved outputs and greater cost efficiency.

Q6: What impact does fine-tuning have on the generated output of LLMs?

The generated output of LLMs can be enhanced through fine-tuning, with the LLaMA-2 model improving its responses by 100% following the fine-tuning process.

Q7: Are LLMs, such as GPT-3.5 and LLaMA-2 models, or NLP models like BERT and RoBERTa, more effective in predicting product review ratings and, more generally, in regression and sentiment analysis tasks?

Overall, LLMs are more accurate in predicting product review ratings than non-LLMs after fine-tuning.

Q8: Can LLMs be effectively utilized for the evaluation of product reviews, and how can LLMs revolutionize the e-commerce sector?

The authors answer this one in the final chapter which will be discussed in the next section.

4.1.4. Conclusion and future work.

This paper concludes that LLMs are highly effective in predicting user sentiment and identifying lexical context in product reviews, providing valuable insights for managers, data analysts, and marketers to track customer satisfaction. The authors note limitations such as the small dataset size and ethical concerns regarding LLMs' handling of sensitive data, which may impact user privacy. Future work involves expanding the dataset to enable scalable fine-tuning, incorporating segmentation techniques, utilizing multimodal data (combining video, images, and text) for sentiment analysis, and developing robust ethical frameworks to ensure secure AI integration into e-commerce, with a focus on secure data handling.

4.2. eCeLLM: Generalizing Large Language Models for E-commerce from Large-scale, High-quality Instruction Data

4.2.1. Introduction and objectives.

This paper investigates the use of Large Language Models (LLMs) for generalized e-commerce modeling, introducing the ECInstruct dataset to enhance model generalization for new products and users. The authors developed a series of LLMs, named eCeLLM, to evaluate their performance on out-of-domain tasks with unseen products and users. The code for the ECInstruct dataset and eCeLLM models is available here [10].

4.2.2. Methodology.

The ECInstruct dataset includes 10 real-world e-commerce tasks divided into four categories: product understanding, user understanding, query-product matching, and product question answering. Data quality was ensured by filtering for English content, removing overlaps, selecting well-documented products, and manual processing. The dataset was split into training (92k samples), validation (9.2k samples), in-domain (IND, 1k samples), and out-of-domain (OOD) test sets. Three eCeLLM models were developed using fine-tuning with LoRA and the Huggingface transformers library:

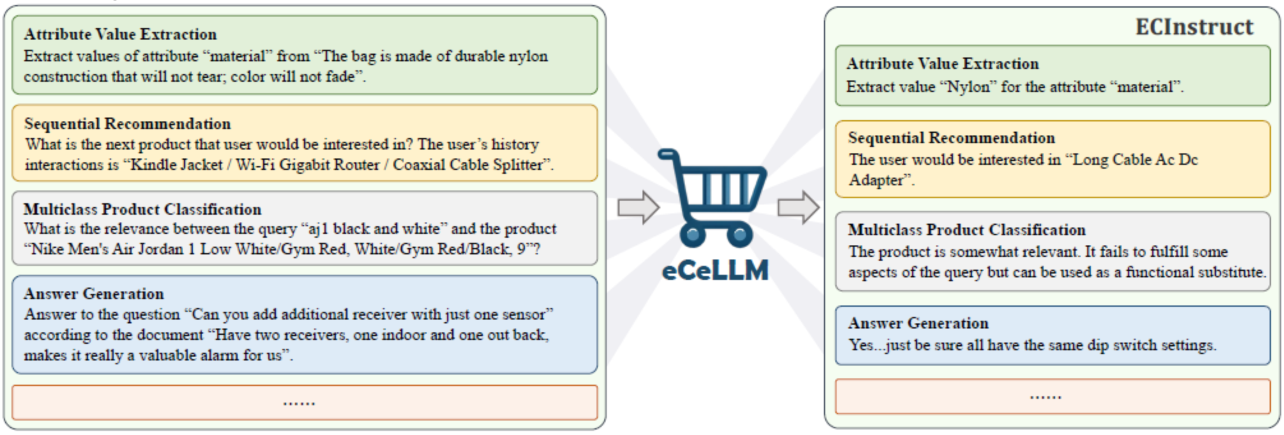
eCeLLM-L: Large model with training from Flan-T5 XXL (11B parameters) and Llama-2-13B.

eCeLLM-M: Trained from medium-sized base models, Llama-2 7B-chat and Mistral-7B Instruct-v0.2

eCeLLM-S: Trained from small base models, Flan-T5 XL (3B) and Phi-2 (3B).

Evaluations included 1-shot testing for general-purpose LLMs (GPT-4 Turbo, Gemini Pro, Claude 2.1), 0-shot and 1-shot testing for e-commerce LLMs (EcomGPT), and specific task evaluations for state-of-the-art task-specific models using ECInstruct data.

Figure 3: Overall scheme of eCeLLM instruction-tuned with ECInstruct



4.2.3. Findings.

Figure 4: Overall performance with the IND dataset

Table 3. Overall Performance in IND Evaluation

Model	AVE	PRP	PM	SA	SR	MPC	PSI	QPR	AP	AG
	F1*	Macro F1	F1	Macro F1	HR@1	Accuracy	F1	NDCG	F1	F_{BERT}
GPT-4 Turbo	0.495	0.326	0.753	0.516	0.387	0.611	0.195	0.875	0.649	**0.858**
Gemini Pro	0.396	0.136	0.867	0.470	0.269	0.584	0.248	0.821	0.506	0.855
Claude 2.1	0.381	0.275	0.523	0.415	0.066	0.655	0.273	0.821	0.280	0.841
Llama-2 13B-chat	0.002	0.333	0.434	0.188	0.056	0.504	0.252	0.815	0.623	0.811
Mistral-7B Instruct-v0.2	0.369	0.324	0.613	0.470	0.164	0.529	0.305	0.842	0.588	0.853
EcomGPT	0.000	0.091	0.648	0.188	0.042	0.540	0.170	0.000	0.086	0.669
SoTA task-specific model	0.546	0.588	**0.995**	0.573	0.265	**0.703**	0.389	0.859	0.830	**0.858**
eCeLLM-L	0.582	**0.611**	**0.995**	**0.648**	0.526	0.684	**0.501**	0.870	**0.851**	0.841
eCeLLM-M	**0.662**	0.558	**0.995**	0.639	**0.542**	0.696	0.305	**0.876**	0.846	0.842
eCeLLM-S	0.509	0.518	0.991	0.596	0.479	0.650	0.392	0.870	0.846	0.842
improvement (%, avg: 10.7)	21.2	3.9	0.0	13.1	40.1	-1.0	28.8	0.1	2.5	-1.9

Results showed that eCeLLM series surpassed the currently used models for generalized tasks (also shown in the figure above for the IND dataset). These models had better performance in both the in-domain and out-of-domain datasets.

4.2.4. Conclusion and future work.

The conclusion of this paper was that the dataset created will prove very useful for future research on generalized tasks for e-commerce, as will the eCeLLM series of models. According to the authors, future work can be seen in two parts: 1) explanation generation can be included to further improve ECInstruct. 2) User profiling could be better enabled once metadata is available and advanced foundation models could be developed to address unique and fundamental challenges and tasks in e-commerce.

4.3. Optimizing ECommerce Listing: LLM Based Description and Keyword Generation from Multimodal Data

4.3.1. Introduction and objectives.

This paper explores using Large Language Models (LLMs) to create descriptions and keywords for e-commerce products through image recognition and metadata. It outlines five key methods for writing effective e-commerce product descriptions: using persuasive language, focusing on benefits over features, employing emotional triggers, optimizing for SEO, and conducting evaluations to identify and fix issues. The goal is to develop a system using GPT-3.5 to produce product descriptions from given input.

4.3.2. Methodology.

Figure 5: Table with each of the LLMs used in the experiment

Table 1: Detailed list of LLMs, with each model's name, parameter count, and API or GPU access for local execution

LLM	Exact Name	Model Size	API/ GPUs
GPT - 3.5	gpt - 3.5 - turbo - 0613	175B	API
GPT - 4	gpt - 4 - 0613	~1.8T	API
Beluga2	StableBeluga2	70B	4
Beluga - 7B	StableBeluga - 7B	7B	1
Solar	SOLAR - 0 - 70b - 16bit	70B	3

Data was collected from Taobao transactions, though the dataset is not shared. Five LLMs were tested: GPT-3.5 and GPT-4 (via paid APIs) and three open-source models (ran on local GPUs). They're shown in the figure above. The models generated keywords and descriptions from product names. Performance was measured using custom precision, F1, recall scores, BLEU score [4], lexical diversity, and attribute capturing. The KOBE model was also tested against baseline models.

4.3.3. Findings.

Results indicated that GPT models outperformed open-source ones, with GPT-3.5 generally performing better than GPT-4, except in JSON format where GPT-4 excelled. The KOBE model, which is a custom model developed by the authors, outperformed all baseline models across metrics such as BLEU score, lexical diversity, and text quality.

4.3.4. Conclusion and future work.

The KOBE model showed promising results, suggesting its potential for future e–commerce applications. The study's limitations included a phrasally limited Taobao dataset and a lack of user data. Future work should explore performance improvements and real-time applications. While GPT-4 outperformed GPT-3.5 overall, GPT-3.5 offered significant cost savings. The KOBE model consistently outperformed state-of-the-art models in BLEU score. The authors recommend further research using their TaoDescribe dataset.

4.4. CTR-BERT: Cost-effective knowledge distillation for billion-parameter teacher models

4.4.1. Introduction and objectives.

This paper focuses on applying pre-trained LLMs for Click-through-rate (CTR) [5] prediction in e-commerce product advertisements, which is challenging due to the need to learn from both language and tabular data, maintain low latency (<5 ms) during inference, and adapt to changing advertisement distributions. A BERT-like model with twin encoders, called CTR-BERT, was scaled to 1.5 Billion parameters which showed a considerable increase in performance compared to currently used techniques. The goal of this paper is to enhance CTR prediction for e-commerce product advertisements by leveraging a novel LLM-based approach that integrates both language and tabular data, maintains low-latency performance, and adapts to changing advertisement distributions.

4.4.2. Methodology.

CTR-BERT is a lightweight, cache-friendly model with separate arms for Page and Ad products, using late fusion of text and tabular features. It reduces inference latency to less than 3 ms by caching precomputed text embeddings. The model, with 70 million parameters, can be trained on 8 GPUs in under a day, making it cost-effective compared to a 1.5 billion parameter teacher model. The BERT backbone processes long-term textual features, while the fusion MLP layer handles frequently changing features, minimizing the need for frequent re-training. Knowledge distillation from the teacher model is used to train CTR-BERT, combining cross-entropy and KL divergence losses. The student model is initialized, distilled, fine-tuned and tested using OOD data from 2020 and ID data from 2021 to optimize ID performance.

4.4.3. Findings.

The dataset is sampled from online traffic with both text and numeric/categorical features. The OOD dataset from 2020 has 1 billion training points and 25 million test/validation points. The ID dataset from 2021 has 200 million training points and 25 million test/validation points. Text data is preprocessed using the Sentencepiece tokenizer with a 32,000-token vocabulary. Metrics are reported using ROC-AUC on imbalanced binary classification tasks. The CTR-BERT student model, although 25 times smaller, experiences only a 0.30% ROC-AUC drop compared to the 1.5 billion parameter teacher. MLM pre-training followed by OOD distillation and self-training provides the best performance gains, combining benefits from pre-training and distillation. Distillation significantly boosts performance over supervised learning alone on OOD datasets.

4.4.4. Conclusion and future work.

In an online experiment within a major e-commerce platform, CTR-BERT outperformed the MLP-based baseline, resulting in an average CTR improvement of 2%. Notably, there was a significant 5+% CTR lift on tail traffic, indicating superior generalization of the CTR-BERT model compared to conventional methods. In the future, the authors aim to enhance representation learning by incorporating image data, further refining the performance and applicability of LLMs in CTR prediction.

4.5. LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following

4.5.1. Introduction and objectives.

The topic of this paper is to introduce LLM usage on e-commerce authoring which involves creating compelling promotional content tailored to drive product sales. While large language models (LLMs) offer a promising solution for various authoring tasks in this domain, mainstream LLMs trained on general corpora have limitations in addressing the unique and complex features of e-commerce products and customers. Additionally, concerns arise regarding the remote accessibility of LLMs like GPT-3.5 and the protection of customer privacy during data transmission. This paper proposes LLaMA-E, a unified and customized instruction-following language model specifically designed for diverse e-commerce authoring tasks. The goal is to develop models that comprehensively understand e-commerce authoring knowledge by incorporating features relevant to customers, sellers, and platforms.

4.5.2. Methodology.

Domain experts create a seed instruction set focusing on five e-commerce authoring task which are Ads Generation, Query-enhanced Title Rewriting, Product Classification, Purchase Intent Speculation, General Q&A, product title creation, explicit and implicit product preferences, product taxonomy and platform background knowledge. GPT-3.5-turbo-0301 expands the seed instructions, ensuring semantic consistency and then they're used to fine-tune the Llama models using LoRA [6] to effectively incorporate e-commerce knowledge for authoring tasks. Preprocessing included filtering out data marked "no action" and removing emojis and interfering characters. Cross–validation was also used with a test set of 19,367 unseen product instances, including additional product description data and a validation set with 30 Q&A pairs from the platform's "Help Center" not included in training. LLaMA-E models were compared with GPT-2, BART, T5-base, GPT-Neo, and LLaMA models.

4.5.3. Findings.

Quantitative evaluation results demonstrate the superiority of LLaMA-E models over baseline models across most metrics. The LLaMA-E-7b model particularly excels in overall performance. For qualitative evaluation, Llama-E models outperform the baselines on the following tasks: Ads Generation, Query-enhanced Title Rewriting, Purchase Intent Speculation and General e-commerce Q&A. Human evaluation results indicate that LLaMA-E models achieve competitive rating scores compared to GPT-3.5 and outperform other baselines.

4.5.4. Conclusion and future work.

The authors conclude with saying that by establishing correlations between key features and task requirements, LLaMA-E efficiently utilizes various service agents, offering a reliable and privacy-conscious solution. Compared to other baselines, LLaMA-E models achieve state-of-the-art results in both quantitative and qualitative evaluations. Extending these models to cover a broader range of authoring tasks is a valuable avenue for future research. Additionally, considering the multilingual aspect and incorporating user personalization features into e-commerce authoring tasks could further enhance the models' capabilities and offer more customized content-based recommendation services.

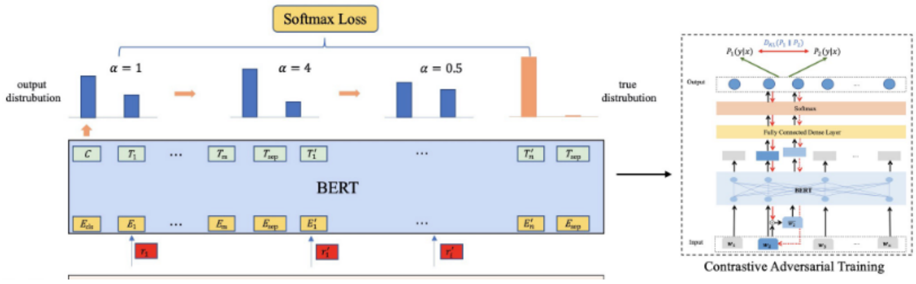
4.6. Robust Interaction-based Relevance Modeling for Online E-Commerce and LLM-based Retrieval

4.6.1. Introduction and objectives.

In this paper, the authors present 3 methods to improve search result relevance. First one is dynamic length representation scheme which facilitates expedited inference by adapting to varying lengths of input text.Next up, we have professional terms recognition method which identifies subjects and core attributes from complex sentence structures, enhancing the model's ability to understand and process specialized terminology. Last method covered is contrastive adversarial training protocol which aims to bolster the model's robustness and matching capabilities by exposing it to adversarial examples during training. Goal of the paper is to leverage an LLM architecture to improve semantic relevance calculation in e-commerce search engines, crucial for ensuring that selected items closely match customer intent.

4.6.2. Methodology.

Figure 6: BERT architecture used for Contrastive Adversarial Training



The methodology, abbreviated as ei-SRC, introduces an interaction-based semantic relevance calculation method for e-commerce with the three components mentioned above. Dynamic-length Representation Scheme dynamically adjusts token lengths by batch-dropping zero-padding columns, reducing computational overhead, while pre–computed tokens reduce computational cost and latency. Professional Terms Recognition Strategy extends the vocabulary with top-frequency words from e-commerce queries and titles, enhancing model representation for specific terms. It also employs Named Entity Recognition (NER) to recognize object and core keywords, improving input embeddings. Contrastive Adversarial Training minimizes the model's sensitivity to perturbed input embeddings, improving robustness. Dataset consisted of 80 million (!!!) pairs of queries and keywords from alibaba's search logs. Authors pre-trained models with a learning rate of 5e-5 and batch size of 32. For fine-tuning, the learning rate was set to 2e-5, with a batch size of 1024. Dynamic-length representation (DRS), online caching, vocabulary expansion, and contrastive adversarial training (CAT) were applied.

4.6.3. Findings.

Offline evaluation included AUC, F1 scores, Spearman's, and Pearson's correlation coefficients where ei-SRC surpassed all baselines. For the online evaluation, ei-SRC was integrated to alibaba's search engine. A/B tests showed that ei-SRC contributed greatly to CTR, conversion rate (CVR) and earnings rate (PAY). Manual evaluation from experts showed incremental improvements in semantic relevance with each strategy implementation, with CAT yielding the most significant enhancement.

4.6.4. Conclusion and future work.

The interaction-based method substantially improved user search experience, leading to increased clicks, conversions, and revenue. Additional evaluation results and source codes are publicly available here [1], with plans to integrate ei-SRC into ranking models and explore its application in recommendation systems and advertising. By integrating dynamic length representation, professional terms recognition, and contrastive adversarial training, the authors aimed to improve relevance matching. Extensive experiments, both offline on annotated query-item pairs and rigorous online A/B tests, have confirmed its effectiveness in enhancing the search experience and increasing industry revenue. Notably, this approach has been consistently operational on www.alibaba.com, managing its entire search traffic for over 12 months and undergoing multiple iterative improvements.

5. Comparative Analysis

In this section, I will be comparing all 6 papers' contribution to two e-commerce themes and assessing some important aspects of each of them that I consider strengths and weaknesses.

5.1. Themes

Below, I will examine some themes for e-commerce that the papers improve through the research provided:

5.1.1. Customer Satisfaction.

the first paper, which uses LLMs for product review evaluation, enhances this theme by allowing businesses, e-commerce platforms and product manufacturers know their customers' sentiment, fix problems that occur with various procedures related to e-commerce such as shipping, increase product quality through feedback and overall improve the buying experience for all people in the future.

The research in the second paper helps customer satisfaction by improving LLMs for generalizing new users and new products, thus alleviating a problem that recommendation systems also have. In this manner, users will have better information on new products that might interest them, despite never looking for them through targeted advertising or explicit searching, and have a better customer experience overall.

The third paper enhances customer satisfaction by creating better descriptions and keywords for products. Customers have an easier time understanding a product if it has a good description and they save time and energy by being provided with an easier search through improved keywords. This leads them to have a better shopping experience online and increased chances of buying said product.

The fourth paper boosts customer satisfaction by improving CTR. A high CTR indicates that users are finding their desired products faster through search queries, navigation is easier, user experience is optimal and the trust of the platform or website is increased in the customers' eyes, since they will believe that with a single click they'll be able to find what they're looking for.

The fifth paper augments this theme by improving e-commerce authoring tasks. This means that users will have better advertisements, targeted to their likings, clearer product descriptions, high-quality images and videos of products and a much more responsive customer support through the Q&A.

The sixth paper improves customer satisfaction by enhancing search result relevance. This leads to higher engagement, users are more inclined to click on products that interest them, better personalization because users will obtain targeted products from the search results. Furthermore, it will also increase trust and credibility of the platform or website since users will know that they will get what they want with the proper query in the search bar.

5.1.2. Technical Implementation and Scalability.

In the first paper, the authors created a web application which functioned as a web crawler to obtain the product reviews from various prominent online marketplaces such as Amazon and ebay. They also used GPT-3.5, Llama-2B, BERT and RoBERTa for the experiment. The scalability in this paper lies in the web crawler app, which can be coded to handle more online marketplace websites to get a larger amount of product reviews to be used as data and the LLMs can be further fine-tuned with new data or better hyperparameters.

Furthermore, the second paper had two implementations for the experiment, the ECInstruct dataset and the eCeLLM series of models. The dataset, as mentioned in the paper as well, can be further enhanced to have more data and the model series can have more data for fine-tuning, better hyperparameters and of course more models can be added to the series if deemed necessary.

In addition, the third paper contained a dataset from Taobao with a large amount of data, along with 5 LLMs and the KOBE model that were used to conduct the experiment. The dataset itself is scalable, as the authors remarked in the paper, in order to acommodate more data for the LLMs.

To continue, the fourth paper employed CTR-BERT scaled to 1.5 billion parameters, evaluated against other models on CTR prediction. The dataset can be scaled further with more data obtained, more parameters can be added and as the authors mentioned, image data can be added to refine performance of the LLMs.

Moving forward, the fifth paper involved the use of Llama-E, a series of models trained for e-commerce authoring tasks, evaluated against other baseline models. Here, the dataset can also be scaled further with more data from multiple languages and more e-commerce authoring tasks can be added.

Lastly, the sixth paper harnesses LLMs in a different way, creating a BERT-like architecture with the 3-method interaction-based semantic relevance. Here, the only thing that can be scaled further is probably parameters for ei-SRC because the dataset is already very large.

5.2. Critical Evaluation

All papers provided very good results for the e-commerce field of study and they certainly paved the way for future research with their datasets and by using LLMs for customer satisfaction, better and more relevant search results for products. However, they all had some negative aspects that must be pointed out. This sub-section will cover the positive and negative parts of each paper and reasons behind each of them.

5.2.1. First paper positives and negatives.

First of all, this paper [13] is very easy to understand. The experiment is straightforward, a classification task applied to e-commerce that uses LLMs as the model and the classic ML metrics (F1, recall, precision, accuracy) for evaluation along with answers to 8 research questions. The web application that was created to crawl the data is also a huge plus, further showing the expertise of the authors in the field and providing the experiment with a very robust dataset. The questions that were set at the introduction were also a challenge and they were all addressed at the end of the paper very thorougly and in great detail, paving the way for future research. The small literature review at the beginning further helps the reader understand the context of the paper, and provides a quick introduction to the background needed. Dataset and code of the paper are shared to the public.

One negative aspect this paper has is that the dataset is randomly selected (due to crawling) and it is not representative of the diverse e-commerce landscape that exists currently, diminishing some of the potential that the LLMs might have had with a more properly refined dataset. Another weakness this paper has is that there is no data privacy check for the usage of customer sentiment in LLM training.

5.2.2. Second paper positives and negatives.

Positive aspects for this paper include the creation of a whole new dataset used explicitly for the purpose of e-commerce generalization tasks, ECInstruct. Also the series of LLM models called eCeLLM , fine-tuned for generalized tasks too is a remarkable feat of the authors. The cross validation in the pre-processing part sets the stage for two evaluation sets, IND and OOD which are both used to evaluate eCeLLM. This helps the model understand new user and new product text data much better. The experiment is well-evaluated with the eCeLLM series being tested against several state-of-the-art and specialized LLMs for generalization tasks and outperforming them in both in-domain and out-of-domain datasets. Dataset and code of the paper are shared to the public as well.

Negative parts of this paper were that the ECInstruct dataset simply appears in the paper's introduction like a deus ex machina, without any background of how it was created, if there was any pre-processing for it. The generalized tasks are also not explained at all, for example a specific task is not given throughout the whole paper, just the term generalized tasks. Not much metadata is available to enable user profilling.

5.2.3. Third paper positives and negatives.

The third paper also has a dataset of its own called TaoDescribe, containing large amounts of data from Taobao. This helps a lot with the experiment as it provides a multitude of products for the LLMs to create keywords and descriptions from. KOBE is also another positive of this paper as it is specialized for the paper's purpose and helps the authors prove their goal in a more robust manner. The models are evaluated through a variety of experiments, making the result more credible for the reader.

Negative parts of this paper were that there are a lot of metrics used for the experiment, leading to a confusion for the reader, the dataset and code of this paper are also not shared to the public. TaoDescribe has no details on its creation or pre-processing and there is no mention of data privacy safeguards for the users of TaoBao that exist in the dataset's transactions.

5.2.4. Fourth paper positives and negatives.

The fourth paper brings forth a very robust model for the experiment, CTR-BERT is lightweight for the computational cost it requires, the dataset is also pretty large, with a substantial amount of data, techniques used help CTR-BERT outperform baseline models in this task. The paper has also been cited 34 times while being recent, indicating its appreciation by other researchers and the amount of future work it has produced.

Negative aspects include no sharing of the dataset. Few inputs for future work from authors, despite the paper being cited a lot of times. No image data present.

5.2.5. Fifth paper positives and negatives.

Positives for the fifth paper are that this is a very thorough experiment with lots of models being used as a baseline for evaluation of Llama-E. The dataset is hand-crafted by domain experts, leading to adequate instructions for the model and also there is an ethical consideration made by the authors. All data in this experiment is protected and no user privacy data has been used.

Negative parts are once again that this paper does not share the dataset or the source code publicly. There are also no training times for the baseline models.

5.2.6. Sixth paper positives and negatives.

A Positive part of this paper that should definitely be mentioned is the very large amount of data that the dataset contained. 80 million is an extraordinary number, although given the fact that the data is from Alibaba, this is to be expected. The experiment is also very robust, providing very good results overall and leading to further research on recommendation systems. Source code is also available to the public here [1]

Negative parts include that there is no mention of privacy in the paper. Also, in order to understand that the experiment utilizes BERT for a part of the architecture,one has to see the image that shows it in the paper. There is no mention of ei-SRC using it directly. There are indirect references only such as token embeddings and bi-directional encoding.

6. Conclusion

This literature review has explored the application of Large Language Models (LLMs) in the e-commerce sector by comparing six pivotal studies. Through this comparative analysis, several key insights and themes have emerged.

The reviewed papers collectively highlight the transformative potential of LLMs in improving customer satisfaction. These models excel in handling product review sentiment, keyword-description generation, generalized task completion, CTR-prediction, e-commerce authoring and search result relevance improvement leading to increased customer satisfaction and engagement. The studies demonstrate that LLMs provide better results overall than the classic ML algorithms used until now. There are also clear benefits in creating specialized datasets which can be used for fine-tuning said LLMs and get even better results.

In addition, these six studies paved the way for future research in LLM usage for e-commerce. The hand-made datasets can be used for a variety of other experiments in the field such as recommendation systems for products, personalized marketing and search optimization.

Despite the clear benefits, the deployment of LLMs in e-commerce is not without challenges. Ethical concerns, such as data privacy and potential biases in AI-generated content, are significant issues that need to be addressed. Additionally, the implementation of LLMs requires substantial computational resources, which can be a barrier for smaller enterprises.

Overall, the papers reviewed offer valuable insights into the current applications and potential of LLMs in e-commerce. They collectively underscore the importance of leveraging advanced AI technologies to enhance various aspects of the e-commerce experience. By addressing the identified challenges and exploring new research avenues, the field can continue to evolve, driving innovation and efficiency in e-commerce.

This review has provided a detailed comparison of six key studies, shedding light on the strengths and gaps in current research. It aims to show results, benefits and disadvantages of LLMs in e-commerce, guiding future developments in this rapidly evolving field.

7. Bibliography

References

[1] Ben Chen. Github repo for eiSRC source code and methods. https://github.com/benchen4395/ei-src4search, 2023.

[2] Ben Chen, Huangyu Dai, Xiang Ma, Wen Jiang, and Wei Ning. Robust Interaction-based Relevance Modeling for Online E-Commerce and LLM-based Retrieval. https://arxiv.org/abs/2406.02135, 4 Jun 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/2305.10435, 24 May 2019.

[4] Ketan Doshi. Foundations of NLP Explained — Bleu Score and WER Metrics. https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b, 9 May 2021.

[5] Adam Hayes. Click-Through Rate (CTR): Definition, Formula, and Analysis. https://www.investopedia.com/terms/c/clickthroughrates.asp, 22 December 2022.

[6] HuggingFace. LoRA guide. https://huggingface.co/docs/peft/main/en/conceptual_guides/lora, 2023.

[7] Satish Kathiriya, Mahidhar Mullapudi, and Rajath Karangara. Optimizing ECommerce Listing: LLM Based Description and Keyword Generation from Multimodal Data. https://www.researchgate.net/publication/379954342_Optimizing_ECommerce_Listing_LLM_Based_Description_and_Keyword_Generation_from_Multimodal_Data, October 2023.

[8] Kenneth C. Laudon and Carol Guercio Traver. E-commerce, Business. Technology. Society. 10th edition. https://eclass.emt.ihu.gr/modules/document/file.php/AD183/e-books/E-Commerce%202014%20%2810E%29%20by%20Kenneth%20Laudon_9780133024449.pdf, 2014.

[9] Aashiq Muhamed, Iman Keivanloo, Sujan Perera, James Mracek, Yi Xu, Qingjun Cui, Santosh Rajagopalan, Belinda Zeng, and Trishul Chilimb. CTR-BERT: Cost-effective knowledge distillation for billion-parameter teacher models. https://neurips2021-nlp.github.io/papers/20/CameraReady/camera_ready_final.pdf, 2021.

[10] Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. eCeLLM and ECInstruct dataset github repository. https://github.com/ninglab/eCeLLM, 13 February 2024.

[11] Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. eCeLLM: Generalizing Large Language Models for E-commerce from Large-scale, High-quality Instruction Data. https://arxiv.org/abs/2402.08831, 13 February 2024.

[12] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. First paper github repository. https://github.com/kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews, 16 January 2024.

[13] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. https://www.sciencedirect.com/science/article/pii/S2949719124000049#sec1, 16 January 2024.

[14] Kaize Shi, Xueyao Sun, Dingxian Wang, Yinlin Fu, Guandong Xu, and Qing Li. LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following. https://arxiv.org/abs/2308.04913, 9 Aug 2023.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. https://arxiv.org/pdf/1706.03762, 12 June 2017.

[16] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. https://arxiv.org/pdf/2305.10435, 21 May 2023.

[13] [11] [7] [9] [14] [2] [8] [16] [3] [15] [12] [10] [4] [5] [6] [1]