A SURVEY ON FAIRNESS OF LARGE LANGUAGE MODELS IN E-COMMERCE: PROGRESS, APPLICATION, AND CHALLENGE

Qingyang Ren[1], Zilin Jiang[2], Jinghan Cao[3], Sijia Li[4], Chiqu Li[5], Yiyang Liu[6], Shuning Huo[7], Tiange He [8], and Yuan Chen[9]

[1]Department of Computer Science, Cornell Univerisity
[2]Carnegie Mellon University
[3]San Francisco State University
[4]Carnegie Mellon University
[5]Columbia University
[6]University of Southern California
[7]Virginia Tech
[8]Khoury College of Computer Sciences, Northeastern University
[9]New York University

ABSTRACT

This survey explores the fairness of large language models (LLMs) in e-commerce, examining their progress, applications, and the challenges they face. LLMs have become pivotal in the e-commerce domain, offering innovative solutions and enhancing customer experiences. This work presents a comprehensive survey on the applications and challenges of LLMs in e-commerce.

The paper begins by introducing the key principles underlying the use of LLMs in e-commerce, detailing the processes of pretraining, fine-tuning, and prompting that tailor these models to specific needs. It then explores the varied applications of LLMs in e-commerce, including product reviews, where they synthesize and analyze customer feedback; product recommendations, where they leverage consumer data to suggest relevant items; product information translation, enhancing global accessibility; and product question and answer sections, where they automate customer support.

The paper critically addresses the fairness challenges in e-commerce, highlighting how biases in training data and algorithms can lead to unfair outcomes, such as reinforcing stereotypes or discriminating against certain groups. These issues not only undermine consumer trust, but also raise ethical and legal concerns.

Finally, the work outlines future research directions, emphasizing the need for more equitable and transparent LLMs in e-commerce. It advocates for ongoing efforts to mitigate biases and improve the fairness of these systems, ensuring they serve diverse global markets effectively and ethically. Through this comprehensive analysis, the survey provides a holistic view of the current landscape of LLMs in e-commerce, offering insights into their potential and limitations, and guiding future endeavors in creating fairer and more inclusive e-commerce environments.

1 Introduction

The rapid advancement of LLMs has initiated a new era of natural language processing (NLP), revolutionizing various fields with their remarkable capabilities. Among these domains, e-commerce has emerged as a promising arena for the application of LLMs, offering innovative solutions and enhancing customer experiences. This survey investigates the fairness of LLMs in the e-commerce landscape, exploring their progress, applications, and the challenges they face.

---

*Author contact Information: qr23@cornell.edu (Qingyang Ren), zilinjia@alumni.cmu.edu (Zilin Jiang), jcao3@alumni.sfsu.edu (Jinghan Cao), sijiali@alumni.cmu.edu (Sijia Li), chiqu.l@columbia.edu (Chiqu Li), ianl@alumni.usc.edu (Yiyang Liu), shuni93@vt.edu (Shuning Huo), he.ti@northeastern.edu (Tiange He)
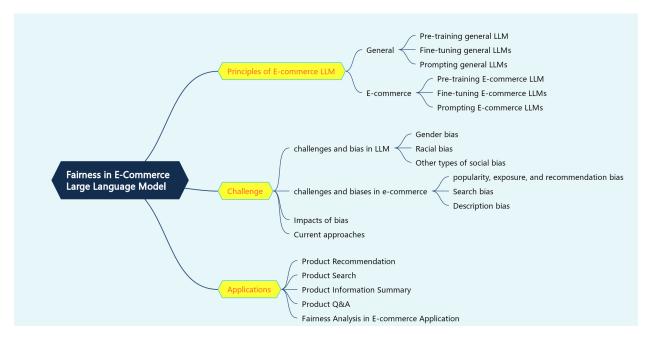
Figure 1: An overview of the fairness of e-commerce LLMs

The emergence of general LLMs, such as LLaMA [1], the GPT series [2,3], and Claude [4,5], has set new benchmarks in NLP tasks, including text classification, summarization, and question answering. Inspired by their remarkable success in general domains, the e-commerce sector has witnessed the rise of specialized LLMs tailored to its unique needs, such as understanding consumer behavior, optimizing search and recommendation systems, and automating content creation for product listings and marketing materials. Notable examples include EComGPT [6] and E-BERT [7], which have gained growing research interest in resolving pain points in both shopping and retailing experiences.

Despite the promising results achieved by existing e-commerce LLMs, several limitations and challenges remain to be addressed. These research gaps motivate the need for a comprehensive review that examines the fairness of LLMs in the e-commerce domain. There are concerns over existing e-commerce LLM regarding their potential to perpetuate harm. These models are typically trained on vast amounts of uncurated data sourced from the Internet, which can result in the inheritance of stereotypes, misrepresentations, derogatory language, and exclusionary behaviors. LLMs not only reflect existing biases but can also amplify them, leading to the automated perpetuation of injustice and the reinforcement of inequitable systems. The presence of biases in LLMs has been extensively documented, encompassing negative sentiment and toxicity towards specific social groups, stereotypical language associations, and a lack of recognition for certain language dialects [8,9].

As shown in Figure 1, this survey is organized as follows: Section 2 summarizes the principles underlying the development of existing e-commerce LLMs, detailing the processes of pretraining, fine-tuning, and prompting that tailor these models to specific e-commerce needs. Section 3 describes the common fairness challenges faced by both general and e-commerce LLMs, shedding light on the potential biases and discriminatory outcomes that can arise from the training data and algorithms employed.

Section 4 outlines the outstanding applications of e-commerce LLMs, showcasing their versatility in areas such as product reviews, where they synthesize and analyze customer feedback; product recommendations, where they leverage consumer data to suggest relevant items; product information translation, enhancing global accessibility; and product question and answer sections, where they automate customer support. However, this section also critically examines the specific fairness challenges that arise within each application domain, highlighting the potential for biases and unfair outcomes that can undermine consumer trust and raise ethical and legal concerns.

Finally, Section 5 explores future research directions, emphasizing the need for more equitable and transparent LLMs in e-commerce. It advocates for ongoing efforts to mitigate biases and improve the fairness of these systems, ensuring they serve diverse global markets effectively and ethically. Through rigorous evaluation and continuous improvement, e-commerce LLMs can foster inclusive and trustworthy online shopping experiences, benefiting both consumers and businesses alike. Through this comprehensive analysis, the survey provides a holistic view of the current landscape of LLMs in e-commerce, offering insights into their potential and limitations, and guiding future endeavors in creating

fairer and more inclusive e-commerce environments. By addressing the fairness challenges directly and promoting responsible development and deployment of LLMs, the e-commerce sector can harness the full potential of these powerful models while upholding ethical principles and safeguarding consumer rights.

2 The Principles of E-commerce LLMs

LLM training comprises three major different approaches: pre-training, fine-tuning and prompting. Given the system complexity in E-commerce domain, the relevant research is shifting from applying a single model training to an integration of multiple LLM model, of which is trained for specific tasks in a larger system.

Table 1: Summary of existing LLMs in E-commerce

Domain	Model Type	Model	Base	Param	Data Source
General	Pre-training	ALBERT [10]	BERT	12M/18M/60M/235M	BooksCorpus, English Wikipedia
General	Pre-training	BERT [11]	-	110M/340M	BooksCorpus, English Wikipedia [11]
General	Pre-training	BART [12]	-	140M/400M	mix of books and Wikipedia data
General	Pre-training	ELECTRA [13]	-	14M/110M/335M	BooksCorpus, English Wikipedia
General	Pre-training	XLNet [14]	-	110M/340M	Wikipedia, BookCorpus
General	Pre-training	ERNIE [15]	-	110M	Wikipedia, other texts
General	Pre-training	Galactica [16]	-	6.7B/30.0B/120.0B	Scientific papers
General	Pre-training	GPT-2 [17]	-	1.5B	WebText
General	Pre-training	DeBERTa [18]	BERT	1.5B	BooksCorpus, English Wikipedia
General	Pre-training	LLaMA [19]	-	7B/13B/33B/65B	Diverse internet data
General	Pre-training	LLaMA-2 [1]	-	7B/13B/34B/70B	Larger dataset
General	Pre-training	GPT-3 [20]	-	6.7B/13B/175B	Extensive internet text
General	Pre-training	PaLM [21]	-	8B/62B/540B	Public and proprietary data
General	Fine-tuning	Alpaca [22]	LLaMA	7B/13B	LLaMA datasets, additional data
General	Fine-tuning	InstructGPT [23]	-	175B	Based on GPT-3
General	Fine-tuning	GPT-4 [3]	-	-	-
General	Fine-tuning	Mixtral [24]	-	8x7B	multilingual data using a context size of 32k tokens
E-commerce	Pre-training	E-BERT [7]	BERT	110M	Amazon Dataset
E-commerce	Pre-training	KG-FLIP [25]	BLIP	224M	Amazon Dataset
E-commerce	Fine-tuning	Ecom-GPT [6]	BLOOMZ	560M	EcomInstruct
E-commerce	Fine-tuning	G2ST [26]	Qwen-14B	14B	Alibaba.com
E-commerce	Fine-tuning	eCeLLM-L [27]	Flan-T5-XXL, Llama-2-13B-chat	11B-13B	ECInstruct [27]
E-commerce	Fine-tuning	eCeLLM-M [27]	Llama-2-7B-chat, Mistral-7B	7B	ECInstruct [27]
E-commerce	Fine-tuning	eCeLLM-S [27]	Flan-T5-XL-3B, Phi-2-3B	3B	ECInstruct [27]
E-commerce	Fine-tuning	GPT4Rec [28]	GPT-2	117M	Amazon Review: Beauty and Electronics [28]
E-commerce	Prompt-tuning	MixPAVE [29]	Pre-training Transformer [30]	0.445M	AE-110K [31] , MAVE [29]
E-commerce	Prompt-tuning	CTM [32]	characterBERT , BERT	-	Huski.ai
E-commerce	Prompt-tuning	Aspect Extraction LLM [33]	GPT-2,BERT	-	Amazon, Yelp, Tripadvisor
E-commerce	Prompt-tuning	CF Recommender Enhancement Model [34]	BERT,RoBERTa	-	Amazon US Reviews
E-commerce	Prompt-tuning	recGPT [34]	pre-trained ChatGPT	-	Amazon reviews and Yelp

2.1 Pre-training

Pre-training involves training a large language model (LLM) from scratch on a substantial corpus of e-commerce texts. This foundational training equips the model with domain-specific knowledge necessary to tackle a wide range of tasks within the e-commerce sector. Utilizing foundational architectures such as the Transformer [35] and subsequent adaptations like BERT [11] and GPT [2], these models are pre-trained on extensive corpora, including product descriptions, customer reviews, and user interactions. These datasets enable the models to capture a variety of linguistic nuances and e-commerce specific terminologies. For instance, BERT's bidirectional training structure is particularly effective for tasks like sentiment analysis and intent recognition, which are crucial for personalized product recommendations and customer service automation [11] . Similarly, GPT's autoregressive features are adept at generating coherent and engaging product descriptions that can significantly enhance search engine optimization (SEO) and user interaction [2].

In e-commerce field, specialized models such as E-BERT [7] and KG-FLIP [25] further refine these capabilities. E-BERT, a derivative of BERT, is re-pre-trained on the Amazon Dataset to boost its efficacy in product-related tasks, thereby enhancing customer interaction quality and the accuracy of sentiment analysis. On the other hand, KG-FLIP extends this by integrating knowledge graphs, which enrich the model's understanding of structured product information and customer data, leading to improved contextual awareness and precision in search functionalities.

Future research is likely to focus on refining pre-training approaches to better handle the informal and varied nature of e-commerce text, expand multilingual support, and enhance context-aware systems, potentially incorporating newer models such as GPT-4, LLaMA-3 for even more robust applications.

2.2 Fine-tuning

There's a lack of consensus on the precise definition of "fine-tuning" (also known as model tuning [11]) within the industry as it emerges with the iteration of researchers experimenting on pre-trained models. Fine-tuning is based on an
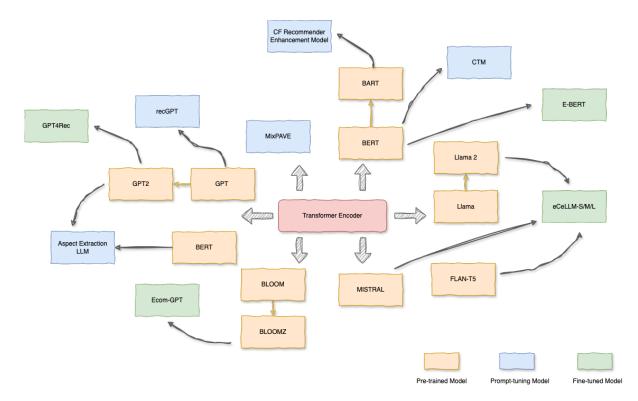
Figure 2: LLM Dependency in E-commerce

existing model and then further trained with specific datasets of samples and parameter-efficient tuning approaches such as Lora [36], Prefix-tuning [37] and full parameter tuning [38]. In E-commerce, precedent researches have emphasized the feasibility of applying fine-tuned LLMs to address specific tasks. Li et al. [6] (2023) proposed a Ecom-GPT model which was trained based on BLOOMZ [39] with instruct datasets. In zero-shot scenarios, this model shows outperformed metrics than other general LLMs in terms of attribution extraction, customer topic classification and product title generation. Chen et al. [26](2024), targeting the translation tasks in E-commerce, offered a general-to-specialized paradigm based on Neural Machine Translation models. Two-step fine-tuning approaches are incorporated into the experiment with 14 billion parameters trained on bilingual datasets. The ROUGE-N and ROUGE-L metrics reveals a better results in translation tasks compared with LLaMA, Qwen and GTP. Peng et al. [27] (2024) has proposed a set of E-commerce models (eCeLLM) to strengthen the generalization abilities including product understanding, user request understanding, product matching and question answering. Particularly, three different model sizes are developed and compared with general-purpose LLMs, e-commerce LLMs and task-specific models given a comprehensive set of individual tasks. It is noted that the models demonstrate higher F1 scores and better generalization ability in out-of-domain test cases. Li et al. [28] (2023) fine-tuned GPT-2 with a 2-step training process and then integrate it with a search engine. This framework aims to leverage LLM to generate recommended products given the customers previous purchase history.

Apart from relevant works on improving single model's performance for either specific tasks or generalization ability in E-commerce, researchers also proposed novel systematic integration of multiple trained LLMs to handle customer requests in real-world applications. Zhou et al. [40] (2023) proposed an approach to synthesize fine-tuned BERT and Llama 2 in a system to efficiently extract product attributes for customer queries in Walmart search functionality. When customers query for specific products, the most likely matched results will be returned. This system employs BERT to generate contextual embedding as the encoding phase and Conditional Random Field(CRFs) [41] layer to decode the tags. In parallel the encoding from BERT is utilized and trained to construct neural network providing decorative relation correction scrutinizing on the returned responses. This system not only incorporate a fine-tuned BERT as base model but also leverage LLAMA 2.0 with prompts to retrieve additional product attributes for customers. Another practice proposed by Zhao et al. [42] (2024) also utilizes BERT-CRF model in encoding/decoding for entity extraction. The difference is that Zhao et al. [42] (2024) builds a complementary graph (Entity Dict) to recommend the next products for customers. Cloude 2, as the pre-trained model, is fine-tuned to discern the complementary relationships in the graph construction.

2.3 Prompt-tuning

Prompt-tuning, as Lester et al. [43] explained, is a mechanism of freezing language models and tuning model with task-specific prompt for each task. The whole process is effective in terms of only a few tun-able tokens prepend per downstream task and grouping multiple adaptations of a pre-trained language model to achieve similar or even better performance than traditional fine-tuning approach (See Figure 3).
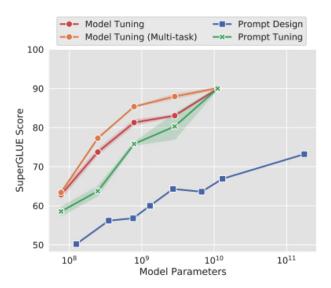

Figure 3: Prompt tuning of T5 performance as size increases [43]

Lester et al. [43]'s work reveals the lightweight parameter footprint and multi-task serving in prompt-tuning. Based on such findings, there have been growing interests in applying prompt-tuning in E-commerce. Yang et al. [29] (2023) proposed a mix-prompt trained model for product attribute value extraction(MixPAVE). The base pre-trained transformer model is frozen except the extraction head [44] to be trained with textual and key-value prompts. Additionally, two datasets: MAVE [45] with 3 million attribute-value annotations and AE-110K collected from AliExpress are used for training and partially for testing in few-shot scenarios. The experiment result shows that MixPAVE outperforms fine-tuning models regarding certain attributes extraction with fewer parameters trained over the process. In a similar study on prompt-tuning, Wang et al. [32] observed a significant rise in the number of new entities emerging in the E-commerce domain. To overcome limitations of existing LLM's ability to handle emerging product entities and titles, Wang et al. [32] proposes a textual entailment model with continuous prompt-tuning approach to better classify entity types. The experiment result shows higher average F1 score in both addition (0.30%) and concatenation (0.38%) as fusion methods.

In some research, prompt-tuning is jointly applied with fine-tuning approach. For instance, Li et al. [33] (2023) fine-tuned GPT-2 with local offline datasets and then feed the model with soft prompts concatenated with embeddings from customer review texts to generate a list of aspect terms. These terms successively are fed into a neural network to generate aspect-based recommendations. With a chain of fine-tuned LLMs and prompt-tuning method connected to aspect-based recommender systems, the frameworks shows better metrics than state-of-the-art baseline methods in providing more meaningful recommendations for users.

Moreover, some researcher intentionally blurs the boundary of prompt-tuning and fine-tuning to serve specific purpose. For example, Dang et al. [34] fine-tuned BERT and RoBERTa with prompt-based learning paradigm to generate more sentiment data in order to tackle the insufficient rating data and data sparsity issues in collaborative filtering recommendation systems.

3 Bias challenges

3.1 Introduction to Fairness and Bias in LLM, E-commerce

LLMs and AI systems have revolutionized the e-commerce industry, enabling personalized experiences and efficient decision-making processes. However, these advanced technologies also introduce challenges related to fairness and

bias. LLMs can perpetuate and amplify societal biases present in their training data, leading to discriminatory outcomes and unfair treatment of certain demographics. In the e-commerce domain, biases can manifest in various forms, such as popularity bias in product recommendations, exposure imbalance among sellers, and skewed search results. These biases not only impact user experiences but also raise ethical concerns regarding transparency, accountability, and equity. Addressing fairness and bias in LLMs and e-commerce requires a multifaceted approach, including the development of fairness-aware algorithms, diverse and representative training data, and rigorous evaluation frameworks. Ongoing research efforts and interdisciplinary collaboration are crucial to mitigate biases, ensure fair outcomes, and build trust in AI-driven e-commerce systems. As the integration of LLMs and AI continues to shape the future of e-commerce, prioritizing fairness and addressing bias remains paramount for creating inclusive and equitable online marketplaces.

3.2 What is fairness and bias?

Fairness and bias are two interrelated concepts that are crucial in the context of AI and machine learning systems. Fairness refers to the principle of ensuring equitable treatment and outcomes for all individuals or groups, regardless of their protected attributes such as race, gender, age, or socioeconomic status. It involves the absence of discrimination or unjustified disparities in the decisions or outputs generated by AI algorithms. On the other hand, bias refers to the systematic errors or prejudices that can be present in data, algorithms, or models, leading to skewed or unfair results. Bias can arise from various sources, including biased training data, flawed data collection processes, or the inherent limitations of the algorithms themselves. Biases can manifest in different forms, such as demographic biases, measurement biases, or representation biases, and can perpetuate or amplify existing societal inequalities. Ensuring fairness and mitigating bias in AI systems is essential to prevent discriminatory outcomes, promote equal opportunities, and build trust in the technology. It requires a proactive approach that involves careful data curation, algorithmic fairness techniques, rigorous testing and evaluation, and ongoing monitoring to identify and address any potential biases throughout the AI lifecycle.

3.3 The challenges and biases in LLM

LLMs have made remarkable advancements in natural language processing, but they also face significant challenges related to fairness and bias. One major challenge is the presence of various types of biases in LLMs, including gender bias, racial bias, religious bias, age bias, sexuality bias, country bias, and disease bias. These biases can manifest in the model's outputs, leading to stereotypical or discriminatory associations. Another challenge lies in the sources of bias in LLMs, which can stem from biased training data, sampling biases, semantic biases encoded in the model's representations, and the amplification of biases during the learning process. Addressing these challenges requires a comprehensive approach that involves careful data curation, bias mitigation techniques, and rigorous evaluation frameworks. Additionally, the development of explainable and interpretable LLMs is crucial to understand and mitigate biases effectively. Researchers and practitioners must also consider the ethical implications of deploying LLMs in real-world applications and ensure that the models align with principles of fairness, transparency, and accountability. Overcoming these challenges is essential to harness the full potential of LLMs while promoting fairness and reducing the risk of perpetuating societal biases.

3.3.1 Gender bias

LLMs can exhibit various types of biases that pose significant challenges to their fairness and reliability. Among them, gender bias [46] is a prominent issue, where LLMs may associate certain occupations or attributes with specific genders, perpetuating stereotypical assumptions. Gender bias has been demonstrated to be present in word embeddings, as well as in a wide range of models designed for diverse NLP tasks, including machine translation, sentiment analysis, auto-captioning, toxicity detection, and beyond. Since LLMs often failed to acknowledge the ambiguity in pronoun references unless explicitly prompted, LLMs often provided explanations that appeared logical but were factually inaccurate, potentially masking the biases. One significant source that introduces gender bias is labelling [47], which occurs when the training data contains biased or subjective labels provided by annotators, leading the model to learn and perpetuate those biases. If the training data for sentiment analysis predominantly associates certain genders with specific sentiments, such as associating women with emotions like "sensitivity" and men with "strength", an LLM might learn and reinforce these stereotypes. For instance, it may consistently associate pronouns referring to women with negative sentiments or pronouns referring to men with positive ones.

One potential solution to mitigate gender bias in LLM is to ensure that training datasets are diverse and representative of different genders, races, cultures, and backgrounds. This involves collecting data from a wide range of sources and demographics to minimize biases present in the data.

3.3.2 Racial bias

Racial bias [48] can also be present, leading to biased outputs or decisions based on race-related information. The models tended to generate biased content for certain racial groups, including unwarranted details based on race. The models exhibited favoritism and has racially-skewed socio-economic projection towards a certain racial group in content recommendations. A important origin of this racial bias is sampling. The issue arises when the distribution of samples from different demographic groups in the training data differs from the actual population distribution, causing the model to exhibit biased behavior. Pre-existing racial prejudices and inequalities within the data can be reflected in the outputs of the language models. Additionally, the vulnerability of the models to prompt manipulation with malicious intent can lead to biased responses.

One potential solution to mitigate racial bias in LLM is to analyze the distribution of racial groups within the data and adjusting the sampling process to ensure equal representation. Random sampling techniques could be adopted to select data for training the language models, which helps reduce the risk of bias by ensuring that each data point has an equal chance of being selected, regardless of racial characteristics. Stratified sampling can be employed to ensure proportional representation of different racial groups in the training data, which involves dividing the dataset into strata based on race and then sampling proportionally from each stratum to ensure balanced representation.

3.3.3 Other types of social bias

Religious bias occurs when LLM demonstrates favoritism or discrimination towards individuals or groups based on their religious beliefs or affiliations. It may originate from the generation of text that stereotypes or stigmatizes certain religions, promotes one religion over others, or misrepresents religious practices and beliefs. Similarly, LLM could cause and even amplify other social bias including age bias, sexuality bias, and country bias. Potential source of these social bias could be semantic bias, which can emerge during the language model encoding process, resulting in biased semantic representations that capture stereotypical associations. These social bias could also be amplified, where the model not only learns the biases present in the training data but also amplifies them during the learning process. They can persist and even intensify further when the model is fine-tuned for downstream tasks.

To mitigate the bias in LLMs requires careful attention to data quality and representative sampling during both pre-training and fine-tuning stages. It also involves developing robust evaluation frameworks to detect and quantify biases, enabling researchers to identify and address them effectively. By understanding and tackling the sources of bias, we can work towards building more fair and unbiased LLMs that provide reliable and equitable outputs.

3.4 What are the challenges and biases in e-commerce?

E-commerce platforms face several challenges and biases that can impact the fairness and equity of the online marketplace. One significant challenge is the presence of various types of biases, such as popularity bias, where popular items or sellers receive disproportionate exposure in recommendation systems, hindering the visibility of less popular offerings. Exposure bias refers to the skewed distribution of visibility and opportunities among sellers, with a small percentage of popular sellers receiving the majority of user attention. This can lead to unfair competition and limit the growth potential of smaller or newer sellers. Recommendation bias can also arise, where the algorithms used to suggest products to users are influenced by factors beyond relevance or user preferences, leading to the promotion of certain products or sellers over others. Search bias can further compound these issues, as the search results on e-commerce platforms may be skewed towards certain products or sellers due to factors such as search optimization techniques or paid placements.Moreover, e-commerce platforms must grapple with description bias, where the textual metadata like product tags and descriptions provided by sellers may not accurately or comprehensively reflect the offerings. Addressing these challenges and biases, which also include ensuring seller-side fairness and providing fair and unbiased product reviews, requires the implementation of fair and transparent algorithms, robust evaluation frameworks, and a commitment to creating an equitable online marketplace. By promoting fairness and mitigating these varied biases, e-commerce platforms can build a more trustworthy and inclusive digital environment that benefits both sellers and consumers.

3.4.1 Visibility and accessibility: popularity, exposure, and recommendation bias

Visibility and accessibility are crucial to products on e-commerce platforms. One prominent issue is popularity bias, where popular items or sellers receive disproportionate exposure and visibility in recommendation systems, overshadowing less popular offerings. This bias can limit the discoverability of new or niche products and hinder the growth of smaller sellers [49]. Another type of bias is exposure bias, which refers to the skewed distribution of visibility and opportunities among sellers, with a small group of popular sellers receiving the majority of user attention and sales. Less exposed items pose the challenge of inaccurate reward function prediction in our e-commerce setting [50].

This bias can create an uneven playing field and stifle competition. As discussed in prior research, the concept of higher-ranked items in recommendation lists commonly receiving more exposure and user attention, and being more likely to be consumed, was also addressed [51, 52].

Recommendation bias occurs when the algorithms used to suggest products to users are influenced by factors beyond relevance or user preferences, such as promotional partnerships or business objectives [49]. This bias can lead to the promotion of certain products or sellers over others, potentially compromising the integrity of the recommendations and creating an uneven playing field [49, 51, 52].

To address recommendation and other biases related to visibility and accessibility, recent research has proposed bias mitigation strategies that go beyond relying solely on binary rating matrices [53]. These more advanced techniques require complex model adjustments, expensive sampling methods, or heuristic propensity scores, and can struggle when users accept or reject multiple recommendations for the same item [53]. An alternative approach, as suggested in the literature, is a multi-process fusion method that combines pre-processing, in-processing, and post-processing techniques to alleviate popularity bias in recommendations [54]. This approach embeds consumer preferences and product popularity information directly into the recommendation model, while also making adjustments to the dataset and recommendation lists, without imposing specific requirements on the underlying recommendation algorithm [54]. This multi-faceted debiasing strategy has been shown to improve recommendation accuracy and consumer interest, making it a promising solution for addressing recommendation bias in e-commerce and LLM-powered systems.

3.4.2 Search bias

Platforms facilitating digital commerce have long grappled with the challenge of search bias, where search results are skewed towards certain products or sellers due to factors such as search optimization techniques or paid placements. This bias in search can adversely impact the visibility and discoverability of products, thereby affecting consumer choice and fair competition. Addressing such biases necessitates the implementation of fair and transparent search algorithms, regular audits and evaluations, and a steadfast commitment to fostering an equitable e-commerce ecosystem for all participants.

To combat the issue of search bias, a novel model training framework dubbed "TripleLearn" was proposed [55]. The cornerstone of this solution is that TripleLearn iteratively learns from three distinct training datasets, deviating from the traditional approach of employing a single training set. By harnessing this iterative learning process, the authors were able to substantially enhance the model's performance, boosting the F1 score from 69.5 to an impressive 93.3 on the holdout test data [55]. This remarkable improvement underscores the efficacy of the TripleLearn approach in mitigating search bias and delivering high-quality search results for e-commerce platforms, thereby fostering a more equitable and transparent search experience.

3.4.3 Description bias

E-commerce platforms have long been grappled with the challenge of description bias, which stems from the manner in which sellers provide textual metadata, such as product tags, to characterize their offerings. As these digital marketplaces facilitate active user participation in the creation and categorization of product-related content, the textual features (e.g., titles, descriptions, tags) generated by sellers may not always be of sufficient quality or accurately reflect the nuances of the products [56]. Sellers, lacking the comprehensive training or domain expertise required to meticulously describe their wares, may instead resort to the use of "tag spam" – the employment of irrelevant yet popular keywords in a misguided attempt to promote their products [56]. This seller-generated bias in product descriptions can have detrimental impacts on the efficacy of crucial e-commerce services, such as search and recommendation systems, ultimately frustrating consumers' ability to effectively locate their desired items [56]. Addressing this description bias is of paramount importance for enhancing the quality of textual product descriptors and, correspondingly, improving the overall e-commerce user experience.

In an effort to alleviate the description bias introduced by seller-provided product tags, the scholarly work under consideration proposes the leveraging of automated tag recommendation techniques that harness search query and click data [56]. The central hypothesis posits that "the set of queries collectively issued by the consumers of the e-marketplace, along with corresponding clicks, reflect a more trustworthy view of the products; thus those queries and clicks can be exploited as a source of high-quality (e.g., more diverse) tags to describe the products" [56]. Guided by this principle, the authors develop novel tag recommendation solutions, including deep learning-based approaches, that generate tags based on the insights gleaned from this search data [56]. Rigorous evaluations revealed that the seller-provided tags often contain significant noise and bias, while the proposed search-boosted tag recommenders were able to substantially outperform the state-of-the-art, improving recommendation effectiveness by over 16 per cent [56]. The authors contend that these recommended tags, borne of the collective consumer search experience, can provide a

more reliable data source for e-commerce search and information services than the original seller-provided descriptions, thereby helping to overcome the inherent biases [56].

3.5 What are the challenges and biases in the application of LLM in the e-commerce field?

The application of LLMs in the e-commerce field presents several challenges and biases that need to be addressed to ensure fairness and equity. One significant challenge is the potential for LLMs to perpetuate and amplify biases present in the training data, which can lead to discriminatory outcomes in e-commerce recommendations, search results, and customer interactions. For example, if an LLM is trained on biased product descriptions or customer reviews, it may generate biased outputs that favor certain products or sellers over others. One example is the Modern collaborative filtering algorithms seek to provide personalized product recommendations by uncovering patterns in consumer product interactions: Addressing Marketing Bias in Product Recommendations [57]. Additionally, LLMs may struggle to capture the nuances and context-specific meanings of e-commerce terminology, leading to misinterpretations or inaccurate recommendations. Another challenge is the lack of transparency and explainability in LLM-based e-commerce systems, making it difficult to identify and mitigate biases. Moreover, the application of LLMs in e-commerce may raise privacy concerns, as these models require vast amounts of user data for training and operation. Ensuring the responsible and ethical use of user data while maintaining the benefits of personalization is a delicate balance. To address these challenges, e-commerce platforms must prioritize the development of fair and unbiased LLMs, incorporate diversity and inclusivity in training data, and implement robust evaluation and auditing mechanisms. Collaboration between e-commerce practitioners, researchers, and ethicists is crucial to navigate the ethical implications and ensure the responsible deployment of LLMs in the e-commerce field [57].

3.6 Impacts of bias

Bias in AI systems and e-commerce platforms can have far-reaching and detrimental impacts on individuals, businesses, and society as a whole. One significant impact is the perpetuation and amplification of societal inequalities and discrimination. Biased algorithms can lead to unfair treatment of certain demographics, limiting their access to opportunities, resources, and services. This can result in a widening of the digital divide and the reinforcement of historical biases [57]. Moreover, biased AI systems in e-commerce can lead to discriminatory outcomes, such as skewed product recommendations, unfair pricing, or biased search results. This can harm the reputation and trust in e-commerce platforms, as consumers may feel misled or unfairly treated. Bias can also have economic consequences, stifling competition and innovation by favoring established or popular brands over newer or niche offerings. This can create barriers to entry for small businesses and limit consumer choice [57]. Additionally, biased AI systems can perpetuate stereotypes and contribute to the spread of misinformation, influencing public opinion and decision-making. The impacts of bias extend beyond the individual level, affecting society's collective values, beliefs, and behaviors. Addressing the impacts of bias requires a proactive and multifaceted approach, including the development of fair and transparent AI systems, regular audits and assessments, and the promotion of diversity and inclusivity in the design and deployment of AI technologies.

3.7 Current approaches

Addressing the challenges of fairness and bias in LLMs and e-commerce platforms requires a multifaceted approach. Current efforts focus on developing fairness-aware algorithms that incorporate fairness metrics and constraints into the training and evaluation processes [58]. These algorithms aim to mitigate biases by ensuring equitable treatment of different groups and promoting diversity in the model's outputs. Another approach is the use of adversarial debiasing techniques, which involve training the model to be invariant to sensitive attributes, such as gender or race, while still maintaining its predictive performance [58]. Researchers are also exploring the use of counterfactual fairness frameworks, which assess the fairness of a model by considering the outcomes under different hypothetical scenarios. In the e-commerce domain, current approaches include the development of fair ranking algorithms that ensure equitable exposure for all sellers and products, regardless of their popularity or historical performance. Collaborative filtering techniques are being adapted to incorporate fairness constraints and promote diversity in recommendations. Additionally, there is a growing emphasis on transparency and explainability in e-commerce algorithms, allowing stakeholders to understand and audit the decision-making processes [58]. Efforts are also being made to curate diverse and representative training data to reduce the impact of historical biases. Overall, the current approaches to tackling fairness and bias in LLMs and e-commerce involve a combination of algorithmic innovations, data curation strategies, and transparency initiatives to ensure equitable outcomes for all participants in the digital marketplace [58].
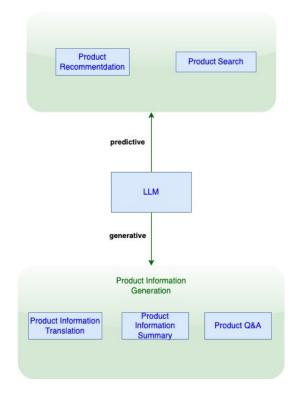
Figure 4: intergration of LLMs in e-commerce workflow

4 E-commerce Application

The integration of language models such as ChatGPT has transformed the customer-business interaction within e-commerce applications. By harnessing the extensive knowledge and linguistic capabilities of these models, e-commerce platforms can deliver personalized and interactive experiences to users. Language models facilitate natural language understanding, empowering customers to ask questions, receive product recommendations, and obtain detailed information in a conversational manner. These models contribute to product search functionalities, generate accurate summaries, and even facilitate translation to cater to a global user base. Moreover, language models can analyze customer sentiment expressed in reviews and feedback [59], providing businesses with valuable insights into customer preferences and enhancing their offerings. By emulating human-like text generation and comprehension, language models elevate customer engagement, streamline the shopping experience, and ultimately drive sales in the dynamic landscape of e-commerce. In this section, we demonstrate real-world applications within e-commerce and discuss the potential fairness concern.

To further illustrate the impact of language models on e-commerce applications, the integration into various aspects of the e-commerce workflow can be visualized as Figure 4. The following graph showcases how language models, such as ChatGPT, enhance customer-business interactions by enabling personalized experiences, natural language understanding, and conversational interactions. Through the graph, it explores how language models contribute to product recommendation, search functionalities, information summarization, translation services, sentiment analysis, and customer engagement within the e-commerce landscape. This visualization serves to highlight the transformative role of language models in optimizing the customer journey, improving user experience, and driving sales in the dynamic realm of e-commerce. Detail workflows can be illustrated by Figure 5

4.1 Product Recommendation

Product recommendation systems play a critical role in assisting users in finding relevant and personalized items or content. With the emergence of LLMs in Natural Language Processing (NLP), there has been a growing interest in harnessing the power of these models to enhance recommendation systems. Different from traditional recommendation systems, the LLM-based models excel in capturing contextual information, comprehending user queries, item descriptions, and other textual data more effectively [60]. By understanding the context, LLM-based RS can improve the accuracy and relevance of recommendations, leading to enhanced user satisfaction. Meanwhile, facing the common
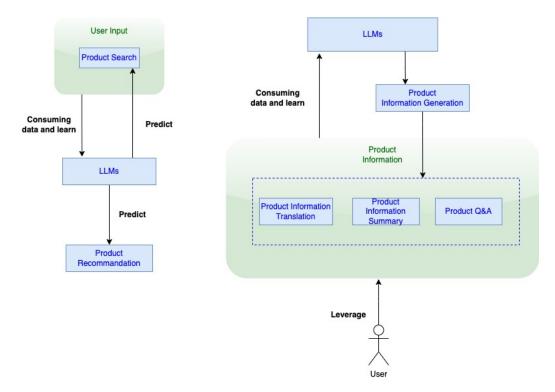
Figure 5: details of LLMs in e-commerce workflow

data sparsity issue of limited historical interactions [61], LLMs also bring new possibilities to recommendation systems through zero/few-shot recommendation capabilities [62]. These models can generalize to unseen candidates due to the extensive pre-training with factual information, domain expertise, and common-sense reasoning, enabling them to provide reasonable recommendations even without prior exposure to specific items or users [63].

4.2 Product Search

In e-commerce, product search involves retrieving catalog items that are semantically related to a customer's query. The search algorithm evolved from relying primarily on lexical matching to semantic matching [65]. With the generalization capability of LLM, the use of language models like ChatGPT can greatly boost the search performance. Leveraging the power of a large language model, a product search system can understand and interpret natural language queries, making the search process more intuitive and efficient [66]. For instance, users can simply describe the product they are looking for in plain language, and the language model can analyze their query to identify relevant products [67].

4.3 Product Information Summary

Language models like ChatGPT have emerged as valuable tools in the e-commerce industry, particularly in generating concise and informative summaries of product information. By utilizing the model's comprehensive knowledge and language processing capabilities, e-commerce platforms can effortlessly condense crucial product details into easily understandable summaries. These summaries encompass essential information such as product features, specifications, pricing, customer reviews, and availability [68]. With their aptitude for comprehending and analyzing textual data, language models can extract pertinent details from various sources, including product descriptions and reviews, to provide comprehensive and accurate summaries. This enables shoppers to swiftly evaluate the suitability of a product based on their specific requirements and preferences, without the need to sift through overwhelming amounts of information.

4.4 Product Information Translation

Transformer-based Machine Translation (MT) models have achieved significant process in the general domain, with more training parameters and full richer bilingual parallel corpora [69, 70]. Especially for LLMs [39], peculiar
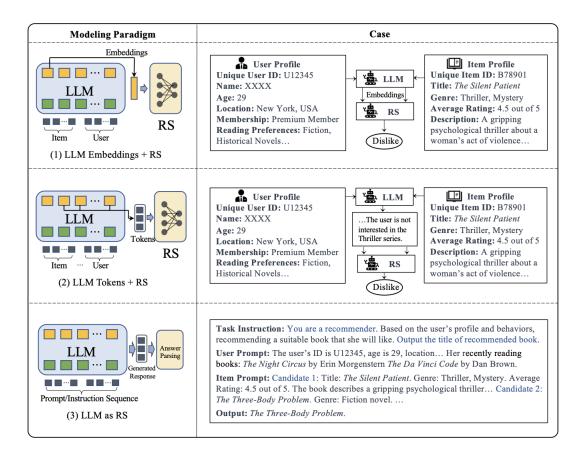
Figure 6: Three modeling paradigms of the research for LLMs on recommendation systems [64].
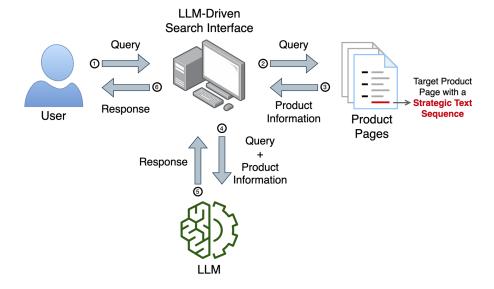

Figure 7: LLM Search: Given a user query, it extracts relevant product information from the internet and passes it to the LLM along with the query. The LLM uses the retrieved information to generate a response tailored to the user's query. The circled numbers indicate the order of the steps. STS: The strategic text sequence is added to the target product's information page to increase its chances of being recommended to the user. [66].

emergence greatly improves their generalization for precise text translation in various sources. Efforts are made into adapting LLM to e-commerce domain by creating linguistic pairs and introduce contrastive learning [26].

LLMs offer a transformative solution for e-commerce product information translation, leveraging their ability to comprehend vast datasets and multilingual proficiency [71]. Through sophisticated contextual understanding, LLMs accurately translate product descriptions, specifications, and reviews across different languages, catering to diverse global markets [35]. Their adeptness in handling technical terminology and customization for specific e-commerce domains ensures precise translations that maintain brand voice and style [11]. Integrated into quality assurance workflows, LLMs facilitate rapid, scalable translation processes while continuously improving through user feedback integration [2], ultimately enabling businesses to reach and engage with global audiences effectively and efficiently.

4.5 Product Information Generation

In the evolving landscape of e-commerce, the integration of LLMs has demonstrated significant potential in enhancing user experience and product visibility. The work by Shanu Vashishtha and colleagues at Walmart Inc. [72] highlights an innovative approach to generating personalized e-commerce banners using LLMs combined with text-to-image technologies like Stable Diffusion. This method effectively transforms user interaction data into visually appealing banners, validated through image quality metrics and human evaluations.

Concurrently, research by Aounon Kumar and Himabindu Lakkaraju [66] investigates manipulating LLMs to prioritize certain products in search results. By embedding strategic text sequences into product descriptions, they show that search algorithms can be influenced to favor these entries, enhancing product visibility and potentially skewing market dynamics. This raises important ethical questions about the manipulation of AI-driven tools in commercial settings.

Both studies exemplify the dual use of LLMs in e-commerce—improving user engagement and challenging the fairness of AI applications. They collectively underscore the need for ethical guidelines and safeguards to ensure that these technologies are used responsibly in enhancing the digital marketplace.

4.6 Product Q&A

LLMs can also positively influence the process of answering user queries. One research [73] demonstrates the utility of LLMs in the domain of product question and answer (Q&A) systems on e-commerce platforms. Specifically, the research focuses on utilizing models like XLNet and BERT to directly answer queries based on product specifications, rather than user reviews. The researchers developed a semi-supervised approach to create a large training dataset for fine-tuning these models, which significantly outperformed the baseline Siamese model in identifying relevant product specifications across various product categories. This method enhances the accuracy of product Q&A systems by leveraging structured product information, showcasing the adaptability of LLMs to different data types within e-commerce. Another research on LLM and Conversational recommender systems (CRS) [74] has conducted experiments on a real-world dataset. It suggests that such collaborations significantly enhance the performance of pre-sales dialogues, offering a promising approach to refining customer interaction and satisfaction in e-commerce settings.

4.7 Fairness Analysis in E-commerce Application

While language models (LLMs) have the potential to enhance e-commerce applications and improve user experience, multiple studies have confirmed that these models can inherit societal biases from the raw training data. Previous work has shown that LLMs tend to reinforce social biases in their generation outputs due to the bias in the large pre-training corpus, leading to unfair treatment of vulnerable groups [75, 76]. Specifically, an increasing concerns about the negative social impact of recommendation systems [77], unfairness issues in recommendation have received significant attention in recent years [64, 78]. Researchers conducted analysis over fairness in recommendation system, including defining the group/individual difference in recommendation results/qualities across different sensitive groups. [78]

Researchers have been actively working on developing quantitative metrics to assess the importance of fairness.

4.7.1 Intrinsic Bias Evaluation Metrics

Similarity-based metrics, such as WEAT [79], SEAT [80], and CEAT [81], employ semantically bleached sentence templates to measure similarities between various demographic groups.

The WEAT [79] metric, quantifies the association between two sets of attribute words (e.g.gender pronouns) and two sets of target words (e.g., career). Formally, the sets of attribute words are indicated by $\mathcal{A}$ and $\mathcal{B}$, and the sets of target
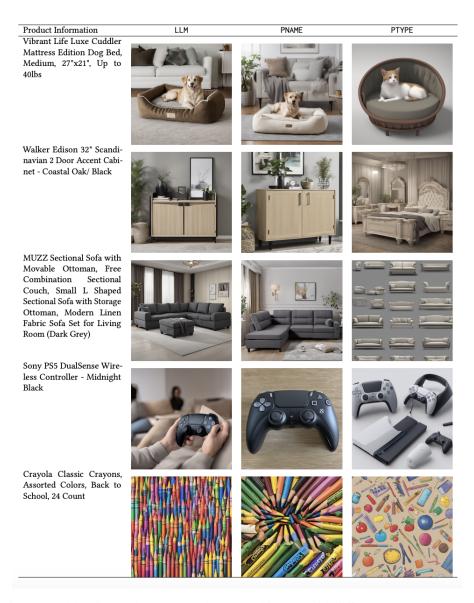
Figure 8: Table with product name and generated images with different approaches [72].

Question	Siamese	BERT	XLNet
Is it single core or multi core?	processor name core i3	internal mic single digital microphone	**number of cores 2**
	processor variant 7100u	processor name core i3	processor name core i3
	os architecture 64 bit	**number of cores 2**	processor brand intel
Does 16 inch laptop fit in to it?	depth 13 inch	**compatible laptop size 15.4 inch**	**compatible laptop size 15.4 inch**
	width 9 inch	laptop sleeve no	depth 13 inch
	height 19 inch	depth 13 inch	height 19 inch

Figure 9: Trained model answering product questions. Top three specifications returned by different models for two questions. Correct specification is highlighted in bold. BERT and XLNet are able to retrieve the correct specifications. [73].

words are denoted by $\mathcal{X}$ and $\mathcal{Y}$. Then the WEAT test statistics are defined as follows:

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in X} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B}),$$

where $s(w, \mathcal{A}, \mathcal{B})$ represents the difference between the average of the cosine similarity of word w with all words in $\mathcal{A}$ and the average of the cosine similarity of word w to all words in $\mathcal{B}$, and it is defined as follows:

$$s(w, \mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \cos(w, a) - \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \cos(w, b),$$

where $w \in \mathcal{X}$ or $\mathcal{Y}$, and $\cos(\cdot, \cdot)$ represents the cosine similarity. The normalized effect size is as follows:

$$d = \frac{\mu\left(\{s(x, \mathcal{A}, \mathcal{B})\}_{x \in \mathcal{X}}\right) - \mu\left(\{s(y, \mathcal{A}, \mathcal{B})\}_{y \in \mathcal{Y}}\right)}{\sigma\left(\{s(t, \mathcal{X}, \mathcal{Y})\}_{t \in \mathcal{F} \cup \mathcal{B}}\right)}$$

4.7.2 Extrinsic Evaluation Metrics

Extrinsic bias evaluation metrics are employed to assess extrinsic bias by measuring the performance gap in the output of downstream tasks. These metrics are often accompanied by benchmark datasets that specifically measure bias in a particular task. Dhamala et al. [82] introduces the Bias in Open-Ended Language Generation Dataset (BOLD), a comprehensive fairness benchmark dataset with a large scale. BOLD focuses on evaluating bias in five domains: gender, race, religion, profession, and political ideology, using natural prompts. By providing prompts that describe specific target populations, BOLD assesses the completions generated by language models over sentiment, toxicity, regard, emotion lexicons and gender polarity. Counterfactual Sentiment Bias (CSB) [83] considers the fairness of the generated text under counterfactual evaluation, which inputs the conditions containing sensitive attributes to GPT-2, and then calculate the sentiment score of the generation. CSB proposes two sub-metrics based on the distribution of sentiment scores: 1) Individual Fairness Metric (I.F.) is the average of the Wasserstein-1 [84] distance of the sentiment score distribution between each counterfactual sentence pair; 2) Group Fairness Metric (G.F.) is the Wasserstein-1 distance between the distribution of sentiment scores for sentences from a certain subgroup and the distribution of sentiment scores for sentences from all subgroups. They are formalized as follows:

$$\text{I.F.} = \frac{2}{M|A|(|A|-1)} \sum_{m=1}^{M} \sum_{a, \hat{a} \in A} W_1\left(P_S\left(x^m\right), P_S\left(\hat{x}^m\right)\right), \ \text{G.F.} = \frac{1}{|A|} \sum_{a \in A} W_1\left(P_S^a, P_S^*\right),$$

where M is the number of templates, A is the set of all subgroups, x and $\hat{x}$ are a pair of counterfactual sentences, a and $\hat{a}$ are their sensitive attributes, $P_S\left(x^m\right)$ and $P_S\left(\hat{x}^m\right)$ are their sentiment score distributions, as well as P_S^a and P_S^* are the sentiment scores distributions over all generated sentences in subgroup a and all subgroups, respectively.

5 Future direction

The future direction of addressing fairness and bias in LLMs and e-commerce platforms requires ongoing research, innovation, and collaboration among researchers, industry practitioners, and policymakers. One key direction is the development of more advanced and nuanced fairness metrics that capture the multifaceted nature of fairness and account for the complex dynamics of e-commerce ecosystems. This involves moving beyond simplistic notions of demographic parity and towards more context-specific and domain-aware fairness criteria. Another important direction is the integration of fairness considerations into the entire AI development pipeline, from data collection and preprocessing to model training, evaluation, and deployment. This requires the establishment of standardized fairness assessment frameworks and the incorporation of fairness checks at every stage of the development process. Researchers should also focus on developing explanatory models that provide insights into the decision-making processes of LLMs and e-commerce algorithms, enabling stakeholders to identify and mitigate sources of bias. Future work should explore the potential of using domain adaptation techniques to transfer fairness-aware models across different e-commerce platforms and contexts, promoting the widespread adoption of fair AI practices. Additionally, there is a need for interdisciplinary collaboration, bringing together experts from computer science, social sciences, ethics, and law to address the societal implications of biased AI systems and develop holistic solutions. By prioritizing fairness and transparency in the development and deployment of LLMs and e-commerce algorithms, we can work towards building a more equitable and trustworthy digital marketplace that benefits all participants.

6 Conclusion

This review provides a comprehensive overview of the principles, applications, and fairness challenges of LLMs in e-commerce, intended to promote further research and exploration in this interdisciplinary field. With the rapid development, LLMs could significantly improve future e-commerce practices and innovations for the benefit of businesses and consumers. However, a critical aspect that demands attention is addressing the fairness challenges that may arise from the integration of LLMs into e-commerce platforms. Ensuring fairness and mitigating potential biases in these models is important for creating equitable and inclusive online shopping experiences for all users. This review highlights the need for sustained interdisciplinary collaboration between e-commerce practitioners, domain experts, and AI researchers. [85]

References

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[2] OpenAI. Gpt-3: Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2021.

[3] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[4] Anthropic. Claude's constitution, May 2023.

[5] Anthropic. Model card and evaluations for claude models, July 2023.

[6] Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce, 2023.

[7] D. Zhang, Z. Yuan, Y. Liu, F. Zhuang, C. H., and H. Xiong. E-bert: Adapting bert to e-commerce with adaptive hybrid masking and neighbor product reconstruction. `https://arxiv.org/pdf/2009.02835`, 2021.

[8] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities, 2020.

[9] Michal Měchura. A taxonomy of bias-causing ambiguities in machine translation. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington, July 2022. Association for Computational Linguistics.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[12] Mike Lewis et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461, 2020.

[13] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020.

[14] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, volume 32, pages 5754–5764, 2019.

[15] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

[16] Meta AI. Galactica: A large language model for science. *ArXiv*, abs/2211.12522, 2022.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[18] Pengcheng He, Xiaodong Liu, Weizhu Chen, and Jianfeng Gao. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2021.

[19] Meta AI. Llama: Open and efficient foundation models. *ArXiv*, abs/2302.13971, 2023.

[20] Tom B. Brown et al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[21] Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

[22] Stanford CRFM. Alpaca: A strong, replicable instruction-following model. *Stanford Alpaca Project*, 2023. Available at `https://stanford-alpaca.gitbook.io/everything-you-need-to-know`.

[23] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[24] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[25] Qinjin Jia, Yang Liu, Shaoyuan Xu, Huidong Liu, Daoping Wu, Jinmiao Fu, Roland Vollgraf, and Bryan Wang. Kg-flip: Knowledge-guided fashion-domain language-image pre-training for e-commerce. In *ACL 2023*, 2023.

[26] Kaidi Chen, Ben Chen, Dehong Gao, Huangyu Dai, Wen Jiang, Wei Ning, Shanqing Yu, Libin Yang, and Xiaoyan Cai. General2specialized llms translation for e-commerce, 2024.

[27] 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, 2024.

[28] Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. Gpt4rec: A generative framework for personalized recommendation and user interests interpretation, 2023.

[29] Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu, and Dongfang Liu. MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada, July 2023. Association for Computational Linguistics.

[30] Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1256–1265, New York, NY, USA, 2022. Association for Computing Machinery.

[31] Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, Florence, Italy, July 2019. Association for Computational Linguistics.

[32] Yibo Wang, Congying Xia, Guan Wang, and Philip Yu. Continuous prompt tuning based textual entailment model for e-commerce entity typing, 2022.

[33] Pan Li, Yuyan Wang, Ed H. Chi, and Minmin Chen. Prompt tuning large language models on personalized aspect extraction for recommendations, 2023.

[34] Elliot Dang, Zheyuan Hu, and Tong Li. Enhancing collaborative filtering recommender with prompt-based sentiment analysis, 2022.

[35] A. Vaswani and et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[36] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[37] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics.

[38] Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, and Xipeng Qiu. Full parameter fine-tuning for large language models with limited resources, 2023.

[39] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2023.

[40] Jianghong Zhou, Weizhi Du, Md Omar Faruk Rokon, Zhaodong Wang, Jiaxuan Xu, Isha Shah, Kuang chih Lee, and Musen Wen. Enhanced e-commerce attribute extraction: Innovating with decorative relation correction and llama 2.0-based annotation, 2023.

[41] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 363–370, USA, 2005. Association for Computational Linguistics.

[42] Qian Zhao, Hao Qian, Ziqi Liu, Gong-Duo Zhang, and Lihong Gu. Breaking the barrier: Utilizing large language models for industrial recommendation systems through an inferential knowledge graph, 2024.

[43] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[44] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[45] google-research-datasets/mave. `https://github.com/google-research-datasets/MAVE`.

[46] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23. ACM, November 2023.

[47] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models, 2024.

[48] Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. Unmasking and quantifying racial bias of large language models in medical report generation, 2024.

[49] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. A survey on popularity bias in recommender systems. `https://arxiv.org/pdf/2308.01118.pdf`, 2023. MediaFutures: Research Centre for Responsible Media Technology & Innovation, University of Bergen, Norway; AAU Klagenfurt, Austria.

[50] Zikun Ye, Reza Yousefi Maragheh, Lalitesh Morishetti, Shanu Vashishtha, Jason Cho, Kaushiki Nag, Sushant Kumar, and Kannan Achan. Seller-side outcome fairness in online marketplaces. `https://arxiv.org/html/2312.03253v1`, 2023.

[51] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7–es, 2007.

[52] Himan Abdollahpouri and Olfa Nasraoui. Addressing seller-side fairness in recommender systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):13805–13806, 2020.

[53] Aravind et al. Cikm. Beyond binary ratings: Debiasing recommender systems with exposure adjustment. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2749–2758, 2021.

[54] Guan Tece, Yiwei Kang, Jingyi Jiang, and Xin Wang. Alleviating popularity bias in recommender systems: A multi-process fusion approach. *Transactions on Emerging Telecommunications Technologies*, page e4443, 2023.

[55] Xiang Cheng, Mitchell Bowden, Bhushan Ramesh Bhange, Priyanka Goyal, Thomas Packer, and Faizan Javed. An end-to-end solution for named entity recognition in ecommerce search. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 1099–1107, 2023.

[56] Fabiano M Bel'em, Rodrigo M Silva, Claudio MV de Andrade, Gabriel Pessoa, Felipe Mingote, Raphael Ballet, Helton Alpontı, Henrique P de Oliveira, Jussara M Almeida, and Marcos A Gonçalves. Fixing the curse of the bad product descriptions: Search-boosted tag recommendation for e-commerce products. *Information Processing & Management*, 60(1):102912, 2023.

[57] Menting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. Addressing marketing bias in product recommendations. `https://nijianmo.github.io/paper/wsdm19.pdf`, 2020.

[58] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P Gummadi. Alexa, in you, i trust! fairness and interpretability issues in e-commerce search through smart speakers. In *Proceedings of the ACM Web Conference 2022*, pages 3695–3705, 2022.

[59] Shaochen Xu, Zihao Wu, Huaqin Zhao, Peng Shu, Zhengliang Liu, Wenxiong Liao, Sheng Li, Andrea Sikora, Tianming Liu, and Xiang Li. Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis, 2024.

[60] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5), 2023.

[61] Aminu Da'u and Naomie Salim. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, 53:2709 – 2748, 2019.

[62] Damien Sileo, Wout Vossen, and Robbe Raymaekers. Zero-shot recommendation as language modeling, 2021.

[63] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation, 2023.

[64] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, February 2023.

[65] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian, Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. Semantic product search, 2019.

[66] Aounon Kumar and Himabindu Lakkaraju. Manipulating large language models to increase product visibility, 2024.

[67] Haixun Wang and Taesik Na. Rethinking e-commerce search, 2023.

[68] Alexander Brinkmann, Roee Shraga, and Christian Bizer. Product attribute value extraction using large language models, 2024.

[69] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn,

Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[70] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[71] S. Liu and J. Liu. A survey of large language models. In *Proceedings of the 2021 International Conference on Data Science and Business Analytics (DSBA 2021)*. ACM, 2021.

[72] Shanu Vashishtha, Abhinav Prakash, Lalitesh Morishetti, Kaushiki Nag, Yokila Arora, Sushant Kumar, and Kannan Achan. Chaining text-to-image and large language model: A novel approach for generating personalized e-commerce banners, 2024.

[73] Kalyani Roy, Smit Shah, Nithish Pai, Jaidam Ramtej, Prajit Prashant Nadkarn, Jyotirmoy Banerjee, Pawan Goyal, and Surender Kumar. Using large pretrained language models for answering user queries from product specifications, 2020.

[74] Yuanxing Liu, Wei-Nan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li, and Wanxiang Che. Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue, 2023.

[75] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models, 2021.

[76] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022.

[77] Tien T. Nguyen, Pik Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web*, WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web, pages 677–686. Association for Computing Machinery, April 2014. 23rd International Conference on World Wide Web, WWW 2014 ; Conference date: 07-04-2014 Through 11-04-2014.

[78] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods and applications, 2023.

[79] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017.

[80] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[81] Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21. ACM, July 2021.

[82] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021.

[83] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online, November 2020. Association for Computational Linguistics.

[84] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification, 2019.

[85] Zhenglin Li, Haibei Zhu, Houze Liu, Jintong Song, and Qishuo Cheng. Comprehensive evaluation of mal-api-2019 dataset by machine learning in malware detection. *International Journal of Computer Science and Information Technology*, 2(1):1–9, March 2024.