Journal of Informatics and Web Engineering
Vol. 4 No. 3 (October 2025) eISSN: 2821-370X

Exploring Generative AI Recommender Systems in E-Commerce: Model, Evaluation Metric, and Comparative Review

Wan-Er Kong[1*], Tong-Ern Tai[2], Palanichamy Naveen[3], Kok-Why Ng[4], Lucia Dwi Krisnawati[5]
[1,2,3,4]Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, Cyberjaya, Malaysia
[5]Faculty Information Technology, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia
**corresponding author: (kong.wan.er@student.mmu.edu.my; ORCiD: 0009-0003-4221-1850)*

Abstract - Generative Artificial Intelligence (GAI) is changing what can be done with Recommender Systems (RS) in e-commerce by allowing much more interactive, situationally aware, and highly tailored experiences for users. The purpose of this paper is to provide overall insight into how GAI, including Large Language Models (LLMs), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and other emerging methods, is affecting the building and running of modern e-commerce RS. This paper classifies generative models into groups based on the type of models used, data modality, and specific domain of application. Their involvement in tasks such as personalized product ranking, content generation, and cold-start problem avoidance is discussed comprehensively as well. In addition, we also analyse innovation in design trends, practical challenges, such as explainability, real-time adaptability, computational scalability, and possible trade-offs, as well as pathways ahead through the lens of current literature and empirical systems. By contrasting GAI-RS with traditional RS, we highlight their advantages in handling several problems, such as data sparsity, generating diverse recommendations, and enabling dynamic user interaction. This paper should serve to broaden awareness among scholars and practitioners about the ever-changing convergence of GAI and intelligent recommendation structures within e-commerce, emphasizing both their transformative potential and operational complexities in practice.

Keywords—Machine Learning, E-Commerce System, Recommender System, Generative Artificial Intelligence, Large Language Models.

Received: 9 March 2025; Accepted: 19 June 2025; Published: 16 October 2025

1. INTRODUCTION

In the present digital era, e-commerce businesses bring much focus on RS since these technologies help achieve customer engagement, satisfaction, as well as income growth [1]. These systems cater to online shoppers by assisting the customers with tailored suggestions based on a thorough analysis of user behaviour and interests, which enhances the overall shopping experience [2]. Many e-commerce stores still use traditional recommendation methods, but these often have limitations like data sparsity, cold-start issues, and insufficient diversity within recommendation [3-4].

However, these limitations can often cause the systems to underperform, which can lead to poor recommendation relevancy, which, in turn, alienates the users.

Recent developments in artificial intelligence (AI) have prompted the integration of generative models into RS as a solution to counter these drawbacks. By providing more dynamic, tailored, and context-aware recommendations, Generative Artificial Intelligence (GAI) holds the potential to revolutionize traditional recommendation techniques. In contrast to traditional collaborative filtering or content-based filtering methods, generative models can overcome the data sparsity issue by utilizing probabilistic techniques to produce more significant user-item interactions [5]. GAI-based recommenders can discover hidden user preferences, create synthetic interactions, and generate context-sensitive as well as diverse recommendations even in cold-start situations [6-7]. The models thus developed are applied to introduce probabilistic reasoning and language comprehension in the recommendation task, allowing the systems to mimic a richer user-item interaction dynamic and produce content (like reviews or product descriptions) that enhances decision making.

This review attempts to discuss the changing scenario of GAI-based RS within the e-commerce landscape. Starting from an overview of basic principles, we focus on foundational aspects in GAI models and their implications for the recommendation task. Section 2 outlines the paradigm shift of technologies utilized in e-commerce platforms. Section 3 presents how diverse generative architectures are employed in varied e-commerce contexts; this is followed by a taxonomy of GAI-based recommender system upon architecture, task, data modality, and domain. In addition to summarizing some of the critical insights from recent literature, this paper also illustrates trends in industrial adoption and discusses ongoing challenges such as scalability, explainability, and bias mitigation. Some potential directions for research at the interface between GAI and intelligent recommendation systems are discussed at the end.

2. A PARADIGM SHIFT: FROM TRADITIONAL TO GAI

RS have been instrumental in shaping the user experience on e-commerce platforms through the personalized content recommendations they offer. Historically, these systems have utilized collaborative filtering, content-based filtering, and hybrid models. Although these strategies have seen tremendous success, particularly on enormous platforms like Amazon and Netflix, they also face challenges that primarily revolve in dealing with data sparsity, cold-start issues, and limited contextual understanding [4].

2.1 Traditional Recommender System

One of the earliest and most widely used techniques is collaborative filtering. The approach is based on a principal that users who agreed in the past are likely to disagree, sometime in the future. This can be implemented through user-based or item-based similarity algorithms using rating matrices or implicit feedback signals such as clicks and purchases [8]. Although collaborative filtering is very effective in capturing user preferences without relying on metadata, it suffers from data sparsity and the cold-start problem wherein new users or new items lack sufficient interactions for the system to generate reliable recommendations [9-10].

Content-based filtering relies on the characteristics of items and user profiles [11]. It recommends items similar to what a user has liked previously. Although this reduces the cold-start problem for the user, it is often monotonous and leads to filter bubbles; thus, the user gets exposed only to a limited range of similar content [12].

Hybrid approaches were designed to overcome the weaknesses of pure collaborative and content-based systems. Matrix factorization, deep learning-based neural recommenders, and graph-based RS models have also proven to be interesting extensions of the traditional RS frameworks. On the other hand, these extensions are still subject to the constraint of requiring huge amounts of historical data and often exhibit poor generalization capabilities in dynamic or sparse data contexts.

2.2 GAI-based Recommender System

A GAI-enhanced recommender system is a significant evolution of the traditional system, as it allows the system to generate new data or content to inform its recommendations instead of merely retrieving what has already been [7].

The two broad categories of systems fall under the general heading of generative model-based RS and content generation-based RS (to be discussed in Section 3).

Using models like GANs, Variational Autoencoders (VAEs), and diffusion models, GAI-enhanced RS can simulate missing interactions or augment training data to tackle the sparsity and cold-start problems much better than classical models [13]. For example, a VAE-based RS can generate realistic user-item interactions for a new item, making it possible to recommend the item even before any significant engagement data has been gathered.

Moreover, LLM such as Generative Pre-trained Transformer 3 (GPT-3), Chat Generative Pre-Trained Transformer (ChatGPT), Bidirectional Encoder Representations from Transformers (BERT), and DeepSeek significantly enhance the ability to understand users' intents, actions, and preferences through natural language interactions [6], [14-16]. These models can dynamically generate personalized descriptions, summarize reviews, or even engage in conversational recommendations thus providing a level of context-awareness and personalization that was hardly ever achieved with static RS algorithms.

Figure 1 summarizes the key differences between these two groupings in terms of some features such as context awareness, handling cold start issue, data and so on. From Figure 1, it is obvious that a GAI-based RS not only supports traditional approaches but also often replaces or redefines the recommendation process within next-generation systems. By enabling models to create missing or supplementary data, Generative AI enriches training contexts, boosts personalization, and grants systems higher flexibility.

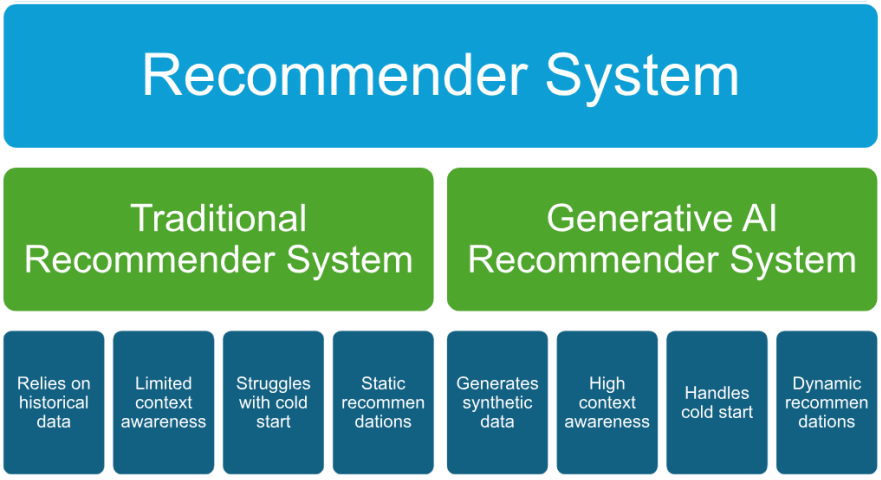

Figure 1. Key Differences between Traditional RS and GAI RS

However, such systems have gained their drawbacks Generative AI models are computationally intensive, may propagate hallucinated outputs or biased content and need strict curation to ensure trustworthy and ethical recommendations [17]. Despite these concerns, the trend is clearly toward the increasing convergence of generative intelligence with recommending systems as an integral feature of hyper-personalized E-commerce platforms that can adapt in real time.

3. TYPES OF GAI RECOMMENDER SYSTEM

The rise of GAI brings a dramatic evolution to the architecture and functionalities of contemporary RSs. As the traditional AI techniques mainly concentrate on identifying patterns in existing data, GAI greatly broadens the scope by producing new data instances that enhance the calibre and variety of recommendations, such as realistic user-item interactions, natural language text, or Machine Learning (ML) reviews. Due to the changes in paradigm, RS can now function more reliably and adaptably in most dynamic settings such as e-commerce.

Recent advances in GAI research have given rise to two fundamental streams of application within RSs, that are (i) Generative Model-Based, and (ii) Content Generation-Based.

Generative Model-Based RSs use generative models to improve the modelling of user preferences and item characteristics [18]. In particular, GANs, VAEs, and more recently diffusion models have been utilized to obtain latent representations of users and items based on sparse or incomplete interaction data. These models produce synthetic user-item interactions which allow the system to (i) cold-start improvement of performance; (ii) tackle data sparsity by enriching the training dataset with synthetic samples, and (iii) enhance diversity as well as personalization in recommendations. For instance, in GAN-based recommenders, plausible negative as well as positive samples are generated, thus training the system on finer distinctions between relevant and irrelevant items. In contrast, VAEs encode information about users and items into continuous latent variables, which permits smooth interpolation and probabilistic reasoning over unseen preferences [19]. The generative nature of such models enables them to discover hidden patterns and imitate user behaviour, leading to more interesting and tailored recommendations, especially useful in contexts where user intent changes rapidly or interaction history is limited [20].

Table 1 summarizes some of the models under this grouping.

Table 1. Research in Generative Model-Based RSs

Technique	Description	Reference
VAE	VAE-based collaborative filtering enhances recommendations by learning latent factors	[25-28]
GAN	GANs generate user preference distributions to improve recommendations.	[29-31]
Normalizing Flows	Normalizing flows transform latent spaces for more expressive modelling in recommendations.	[32-35]
Transformer based	Utilizes self-attention mechanisms to learn contextual representations in RSs.	[33], [36-38]

The second stream is Content Generation-Based RSs. This stream has a focus on using GAI for the creation of auxiliary content which enhances or supports the recommendation process [21]. Such content includes: (i) dynamic product descriptions tailored to user context, (ii) synthetic reviews or summaries that play a pivotal role in influencing purchasing decisions, and (iii) multimedia content (e.g., product thumbnails or preview videos). The core of this category consists of Transformer-based architectures, especially LLM, which include GPT, and BERT. These models are pre-trained on enormous corpora and can be fine-tuned or prompted to produce extremely relevant, tailored textual content for a user.

In an e-commerce RS pipeline, LLMs can produce personalized product descriptions based on a user's browsing history [22]. Moreover, it can distil lengthy product reviews into snippets highlighting features that might be relevant to a particular user [23]. Furthermore, it suggests packages or related products with the help of natural language explanations [24]. By connecting unstructured information (like customer feedback) (structured recommendation algorithms), content generation-based systems can offer a much richer and more interactive experience. This in turn fosters not only greater user trust and satisfaction but also helps systems understand and exploit contextual cues better, which may include sentiment, intent, or product usage situations. Table 2 summarizes some of the current research on this model.

As a summary, while the two models are used differently, in practice, they are often complementary. A powerful GAI-based RS could use generative modelling to simulate user preferences and at the same time employ language models to populate dynamic, user-relevant content in interface areas.

This comprehensive research focuses on how GAI techniques are used in e-commerce RSs, giving particular attention to how effectively they boost recommendation diversity, reduce cold-start issues, and enhance personalization. By conducting a deep study in GAI-based techniques, this study aims to provide a comprehensive understanding of how these models contribute to the development of RSs and their potential to revolutionize the future of personalized online shopping experiences.

Table 2. Research in Content Generation-based RS

Technique	Description	Reference
GPT	Transformer-based language models are frequently employed for natural language tasks like text creation and summarization because they predict the next word or sequence to produce writing that is human-like.	[22], [39]
BERT	Transformer-based models that have been bidirectionally trained to comprehend the context of words in sentences do very well in tasks such as question answering and sentence categorization.	[40-41]
Co-optimization of Content Generation and Consumption	Emphasize enhancing user interaction and content production at the same time to increase content relevancy and consumption.	[33], [42]

Figure 2 depicts the word cloud of GAI and related terminologies, which captures the crucial themes within the intersection of generative AI and e-commerce, reflecting the growing influence of LLM in dynamically reshaping online retail experiences. Models pivotal to this transition include GPT-3, ChatGPT, and BERT to produce human-like language along with contextually relevant content. Furthermore, in the context of e-commerce, these models have been increasingly applied for product recommendations, virtual shopping assistants, and interactive real-time communication with users. The terms 'personalization', 'session-based recommendations', and 'multi-turn dialogue' capture the additional potential of generative models to respond dynamically to changes in user behaviour - much more than traditional RS can offer. In addition, the word 'cloud' also serves to capture the vital foundational concepts that form the backbone of the operational pipeline for GAI systems.

Figure 2. Word Cloud Terminologies of GAI Influence in E-Commerce

4. THE ROLE OF GAI IN VARIOUS SECTORS

GAI has transformed RSs into a variety of sectors by improving decision-making, efficiency, and customization. Unlike traditional methods which mostly rely on collaborative or content-based filtering, GAI can generate new recommendations by detecting the complex patterns in data. Recommendation models driven by GAI have improved user experience and speed up the overall service delivery process in a variety of sectors, such as urban planning, healthcare, education, and entertainment. There are many recent studies that explored the role of GAI in several domains and further demonstrated its effectiveness in enhancing recommendations, which in precisely, personalizing the content as well as supporting decision-making.

In cosmetic and beauty sectors, GAI-based RSs show improved product discovery by generating personalized products and customizing product suggestions. For example, there are existing papers that employ GAI models such as VAEs and GANs that are implemented to evaluate user preferences based on skin type, colour, and beauty habits to give personalized beauty solutions. To illustrate the makeup effect of the selected makeup face [43], for example, suggested a model that can transfer the makeup style from a reference makeup face picture to the user's face. Besides that, [44] proposed a model that employs GAI for individualized cosmetic recommendations, utilizing user skin and face data to improve the accuracy of recommendations. GAI makes it easier to create virtual experiences, which enhances online shopping by allowing users to see how different items look when it is applied to their skin tone or face. In addition, research by [45] also investigated the effects of GAI-powered virtual try-on experiences on consumer decision-making and product recommendations.

In the electronics sector, GAI-based RSs is changing the way people view and choose electronic products and accessories. The adoption of GAI systems is utilized to investigate three components, which are product specifications, user preferences, and behavioural patterns, to offer personalized product suggestions. With these functions, GAI models can offer proactive recommendations with a broader variety of items by forecasting future demands. This leads to an enhancement in decision-making and reduces reluctance to purchase, which especially benefits high-end devices users. In addition to enhancing augmented reality apps for immersive experiences and user interactions, research by [40] demonstrated how GAI may enhance voice assistants and virtual avatars by enhancing images, understanding natural language, and personalizing interactions. A study by [46] explored the breakthroughs and challenges may face in the future in the electronics industry.

By analysing complex user data and generating personalized suggestions, GAI has been employed to boost recommendation accuracy in the book retail industry. To increase user satisfaction and engagement, GAI algorithms can be utilized in assessing a reader's reading preferences and habits to recommend books that align with their interests. [47] suggested a hybrid recommendation system for e-commerce that combines content-based filtering and collaborative filtering methods to compute the similarities of product description and user profile.

Similarly, GAI applications in e-commerce have improved the home decor industry. GAI technology driven system can assess user preferences and recommend home decor items that fit individual tastes. This would lead to enhancement in shopping experience and assisting clients in creating cohesive interior designs. The potential and difficulties faced by the home renovation industry's recommendation system are covered in [48].

It is evident from these areas that RSs could significantly boost income across a range of industries. These insights obtained by GAI-generated are believed to have the potential to uncover the possibility of optimizing recommendation performance. The incorporation of AI technology into RSs will enhance user experiences and optimize a range of commercial applications as they develop further [49].

5. EVALUATION METRICS

5.1 Normalized Discounted Cumulative Gain (NDCG)

NDCG is a ranking-based evaluation metric commonly used in RSs to measure how well the recommended items are ranked in relation to their relevance. As this technique considers both an item's relevance and its ranking, a higher-ranked relevant item adds more to the final score [50]. Lower-ranked relevant things are penalized by the Discounted Cumulative Gain (DCG), which uses a logarithmic discount factor to calculate the relevance of suggested items. In order to ensure a normalized score between 0 and 1, where 1 denotes a perfect ranking, the NDCG is computed as the ratio of DCG to IDCG, while the Ideal DCG (IDCG) reflects the best possible ranking of pertinent elements. This statistic plays an essential role in a model performance as the user experience can be greatly impacted by the placement of pertinent suggestions in top-N recommendation jobs. In e-commerce, NDCG assists in evaluating how well a RS ranks highly relevant goods for consumers in order to make sure that the most helpful suggestions show up first.

5.2 Precision

Precision measures the proportion of relevant recommendations among all recommended items, assessing how the accuracy of the recommendations [51]. It is defined in Equation (1).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

True Positives (TP) = appropriately suggested relevant ones

False Positives (FP) = incorrect suggested irrelevant things

When it comes to movies or product suggestions, where customers value quality above quantity, precision is essential. In these situations, offering highly relevant recommendations is more important than covering every conceivable relevant thing. However, recall must be used in addition to accuracy since precision does not account for whether all pertinent items are retrieved. In e-commerce, precision can reduce irrelevant recommendations and increase user happiness by guaranteeing that customers receive product recommendations that align with their preferences.

5.3 Recall

Recall measures the proportion of relevant items that were successfully recommended out of all relevant items available [52]. It is defined in Equation (2).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negative} \quad (2)$$

False Negatives (FN) = relevant items that were not recommended

In situations like fraud detection systems or medical diagnosis suggestions, when it is expensive to overlook vital information, recall is especially crucial. High recall is frequently used in conjunction with accuracy since it indicates that the system is able to recover the majority of the pertinent things correctly. However, it does not reveal if the retrieved items contain suggestions that are not relevant. In e-commerce, recall is used to ensure that users are exposed to a wide range of relevant products, and this will increase their chances to find desirable or more interesting items.

5.4 F1-score

F1-score is the harmonic mean of precision and recall, providing a balanced evaluation metric when both are equally important [53]. It is calculated as in Equation (3).

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

When there exists imbalance in the relevance of recommendations, the F1-score can assist balance the accuracy and recall trade-off. A high F1-score indicates that the model shows high accuracy in finding relevant items, which is important in applications like search engines and fraud detection where accuracy and recall are critical. In e-commerce, the F1-score guarantees that RSs are able to find a balance between offering highly relevant product recommendations and accommodating a wide variety of customer preferences.

5.5 Area Under the Curve (AUC)

The AUC curve tends to evaluate the ability of a model to distinguish between relevant and irrelevant recommendations by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) using various classification thresholds [54]. This metric can show the likelihood of a randomly chosen relevant item ranked lower than a relevant one. If the AUC = 0.5, then it denotes random guessing, whereas AUC = 1.0, it denotes a flawless model. This metric can show the likelihood that a randomly chosen relevant item would rank lower than a relevant one. If the AUC = 0.5, then it denotes random guessing, whereas AUC = 1.0, it denotes a flawless model. Because AUC assesses ranking performance rather than absolute classification accuracy, it is very helpful when working with unbalanced datasets. It is frequently employed in ranking issues and binary classification jobs, including RSs, spam detection, and medical diagnostics. In e-commerce, AUC is capable of optimizing product ranking algorithms, and this led to users more likely to engage with recommended products.

These assessment measures are essential to make sure that e-commerce RSs offer highly relevant, varied, and well-ranked product recommendations [55]. In precise, top-quality ranks are guaranteed by NDCG, Precision maximizes accuracy, product discovery is maximized by Recall, both of these metrics are balanced by F1-score, and ranking performance is enhanced by AUC. When combined, they improve customer satisfaction, boost engagement, and boost revenue.

6. LITERATURE REVIEW

A study by [56] presented an Ensemble Variational Autoencoder (EnsVAE) framework for recommendations which integrates side information for domain-specific recommendations to address the shortcomings of previous VAE versions. EnsVAE transforms the projected utility matrices of sub-recommenders into interest probabilities so that the VAE can represent changes in their aggregate. This model combines a GRU-based matrix factorization recommender with a GloVe content-based filtering recommender to assess the framework. The model alternates between sub-recommenders according to context, improving accuracy and speeding up utility matrix reconstruction. Empirical studies on real-world datasets demonstrate that EnsVAE works better than state-of-the-art techniques, solving the cold-start issue by precisely predicting interest probability for new users and resources while preserving information during matrix reconstruction.

Another study by [57] proposed a self-supervised VAE model named SSVAE to improve the ability of VAE model in generalizing datasets with sparse interactions. In particular, after first generating many views for every user through data augmentation, they create a pretext task to match the representations learned from different user viewpoints. The goal of this strategy is to improve the joint purpose of the pretext and suggest tasks. This is essential to aid one another during the learning process. The result demonstrates that the proposed model outperforms the traditional VAE recommendation methods. This paper concludes with a highlighting on how GAI is able to significantly improve the range and precision of recommendations while addressing common problems with traditional systems and opening the door for further advancements in AI-driven personalization.

A study by [58] suggested a hybrid model named Hy-VAE that addresses the problem of posterior collapse in VAE-based collaborative filtering recommendation systems. This technique is formed by combining a deterministic autoencoder with a traditional VAE. By using this technique, the proposed solution successfully reduces the mismatch that causes negative impact to be learning by a more balanced encoder and decoder. From the result, this technique brings a better mapping from the data manifold to the parameterized graph. In short, the main contribution of this study is to introduce a new way to optimize latent variable models for recommendation tasks. This new technique is capable of enhancing model stability by combining the advantages of deterministic autoencoders and VAE. Hence, the proposed model effectively reduces posterior collapse and improves recommendation performance, demonstrating its usefulness in enhancing collaborative filtering systems.

[59] presented a unified foundation model for industrial RSs which aims to handle the complexity of many domains and activities, and this includes retrieval, ranking, explanation generation, and AI-assisted content development. To facilitate sample-efficient adaptation for a variety of recommendation tasks, this study leverages M6 (a large-scale pretrained language model) and expresses user behaviour data as plain text rather than creating distinct algorithms for each task and domain. This study successfully build an enhanced version of prompt tuning that outperforms fine-tuning and drastically decreases task-specific parameters along with several methods such as late interaction, early departing, parameter sharing, and pruning to increase computing efficiency, The study shows how the foundation model may be used for a variety of tasks, allowing for conversational suggestions, tailored content generation, and zero-shot recommendation. The results highlight the ability of the model in reducing data demands, lower carbon footprints, and achieve deployment on both cloud servers and mobile devices, marking a significant step toward efficient and scalable industrial recommendation systems.

[60] introduce a unique method named Item Concept Causal Variational Auto-Encoder (ICCVAE) for top-n recommendation tasks to solve the drawbacks of traditional VAE-based recommenders. The proposed technique presume independence between latent components. ICCVAE is design to increase recommendation efficiency and interpretability by giving item ideas a causal framework. According to the proposed framework, this study incorporates causal reasoning into the VAE framework. This can address the sub-optimality brought on by the independence assumption and offers a deeper understanding of user-item interactions at the same time. Based on the evaluation

result, ICCVAE gain a greater result compared to the baseline models in terms of recall and NDCG metrics, underscoring its superiority in recommendation accuracy and interpretability.

[39] proposed a novel generative framework named GPT4Rec that is inspired by search engines. In simple terms, GPT4Rec is a NLP-based RS that can create fictitious search questions based on the user item history and then searches these queries to return suggestions, in contrast to standard methods that regard items as nothing more than IDs. This approach uses the benefit of NLP language modelling to better interpretably capture user interests. The focus of this work is on how beam search combined with multi-query generation improves recommendation variety and more precisely represents user preferences at different granularities. This system further increases versatility by using query-based retrieval to recommend cold-start objects. According to the qualitative case studies, the produced queries offer interpretable representations of user preferences, making the system efficient and flexible enough to manage expanding item inventories.

In a different study, [61] suggested a new GAN-based method named Variational Collaborative Generative Adversarial Network (VCGAN) which was designed to address sparse and noisy data in tailored RS. VCGAN increases the correlation between generated samples and actual data by including a GAN-based framework. In this model, VAE is utilised as the generator for adversarial training while an Auto-encoder (AE) is used to generate latent vectors in order to handle high-dimensional side input. This structure allows the model to have richer latent representations and further increases recommendation accuracy. The main contribution of this study is the efficient use of side information to reduce rating sparsity and improve user-item interaction accuracy. Comprehensive tests on four real-world datasets also show that VCGAN outperforms baseline techniques and is efficient in producing suggestions of the highest calibre.

[62] proposed an LLM-driven knowlEdge Adaptive RecommeNdation (LEARN) framework to improve recommendation systems. This is done by integrating collaborative filtering and open-world knowledge. To preserve knowledge and avoid catastrophic forgetting, LEARN uses pretrained LLMs as item encoders while freezing their parameters in contrast to conventional ID embedding-based techniques which ignore important semantic information in textual descriptions. As its twin-tower structure connects the collaborative and open-world realms, the framework is able to adapt to a variety of industrial settings. The study emphasizes its contribution to increasing recommendation accuracy by addressing computing efficiency as well as adding textual understanding. The efficacy of LEARN is further evaluated by employing industrial datasets and online A/B tests. The results indicate that the proposed method achieves state-of-the-art performance on six Amazon Review datasets, and this proves its supremacy in practical recommendation tasks.

Different research by [40] suggested a special web Application Programming Interface (API) recommendation mechanism for creating mashups. Based on the model's structure, this approach combines quality-of-service (QoS) and content-based approaches with BERT Variants and Graph GAN. It shows significant improvement in the recommendation accuracy compared to the existing methods that depend on keyword matching and historical data. While Graph GAN learns from mashup-service invocation data to produce more pertinent suggestions, the BERT Variants enhance contextual comprehension, allowing the system to collect intricate functional descriptions. This main contribution of this study is making the RS system flexible enough to respond to a variety of natural language inquiries by utilising its strong semantic enrichment component. This is achieved by using paraphrase mining to increase vocabulary and improve similarity metrics. Based on the result, this technique has gained a great improvement in suggestion accuracy, and this especially benefits the developers with little domain expertise or when dealing with confusing service descriptions.

[63] proposed a new model named VCFGAN, an improved collaborative filtering model based on CFGAN to address the issues raised by high-dimensional side information (SI) and sparse user rating matrices (URM). In order to reduce the sparsity of the data before feeding it into the GAN for training, the suggested method uses VAEs to extract latent features from both SI and URM. By using this innovative method, the GAN model can have a better understanding of user preferences and increase the recommendation accuracy. This research highlights the use of VAEs for feature extraction, which improves the training data while preserving computational effectiveness. Experimental results demonstrate that VCFGAN outperforms baseline models such as CFGAN, IRGAN, and CDAE in terms of accuracy and recall.

Table 3 concludes the reviewed research papers in this study.

Table 3. Reviewed Works

Reference	Findings	Evaluation
[56]	Proposed a method to convert the projected utility matrix of sub-recommenders into interest probabilities so that the VAE might capture the variance in their aggregate.	Mean Average Precision, NDCG
[57]	Suggest a self-supervised VAE model to enhance the VAE model's capacity to generalize on datasets with sparse interactions.	Recall, NDCG
[58]	Proposed a hybrid model that reduces the mismatch that impedes learning by better balancing the encoder and decoder, enabling a better mapping from the data manifold to the parameterized graph	Masked Mean Absolute Error, Masked Root Mean Square Error
[59]	Leverages a large-scale pretrained language model, and expresses user behaviour data as plain text rather than creating distinct algorithms for each task and domain	Precision, Recall, F1-score
[60]	Developed ICCVAE which solves the issue of conventional VAE-based recommenders, which presume independence between latent components.	Recall, NDCG
[39]	Proposed an NLP-based RS that generates hypothetical search queries based on the user item history and retrieves recommendations by searching these queries	Recall
[61]	Proposed VCGAN framework that is designed to address data sparsity and noise in personalized recommendation systems	Precision
[62]	Employed pre-trained LLMs as item encoders and freezing their parameters to minimize computing complexity and avoid catastrophic forgetting using collaborative and open-world knowledge.	AUC
[40]	proposed a web API recommendation system for mashup composition that leverages natural text descriptions, semantic enrichment, and deep learning techniques	Mean Reciprocal Rank, NDCG, Mean Average Precision
[63]	Proposed a two-stage collaborative filtering framework, VCFGAN, which uses two VAEs to extract features from URM and side information, followed by training a GAN network with the extracted latent vectors.	Precision, Recall

The results of this study confirm that using VAEs to preprocess data prior to GAN-based recommendation greatly improves performance, and it also provides a reliable remedy for sparsity problems of collaborative filtering systems.

7. COMPARATIVE INSIGHTS AND DISCUSSIONS

The traditional RS to GAI-RS transition marks a shift influenced by significant breakthroughs recently [64-65]. This section offers comparative study along several dimensions, including recommendation accuracy, data handling capabilities, system adaptability, diversity, explainability, and scalability.

7.1 Representation Learning

In traditional RS models of the sort collaborative filtering or matrix factorization, users and items are represented as fixed points in some low-dimensional space vector [66]. These embeddings are derived from historical interactions and, moreover, are usually linear. They offer speed and interpretability but cannot capture complex or subtle user behaviour.

In contrast, GAI-RS utilize neural-based encoders as in VAE or Transformer models which encode user preferences into a non-linear and more expressive latent space [67]. Such models can capture much richer patterns and structures in the interaction data, leading to better generalization and personalization even in sparse or incomplete interaction data [68].

7.2 Handling Data Sparsity and Cold-Start Issues

In traditional RS approaches, performance is heavily reliant on user-item interaction histories, which are designed principally under collaborative filtering or matrix factorization models. However, in realistic e-commerce settings, these interaction data are often sparse users tend to interact with a few items only, and new users or items keep entering the system without much history [69-70]. Consequently, such systems cannot provide meaningful and accurate recommendations due to the lack of sufficient co-occurrence patterns, leading to a significant reduction in personalization and engagement for the user.

Traditional approaches attempt to reduce this problem effect using item metadata or content-based filtering. In this case, characteristics of the products being filtered include categories, brands, and keywords. Although these strategies provide some relief by deriving preferences from known attributes of items, they are significantly inadequate in capturing subtle or dynamic user preferences. Additionally, these methods require heavy feature engineering and are also susceptible to the quality and availability of metadata. For instance, in cold-start situations where a user or item has no interaction data at all - is a severe challenge to these models leading to too generic or irrelevant recommendations which may soon lose the user's trust.

GAI-RS introduce a radically new perspective on sparsity and cold-start challenges. These systems rely on the generative models' ability, instead of being restricted to the only observed data, to discover the true distributions of data and generate new, realistic user-item interactions [71]. For instance, in training models like VAEs and GANs, one trains these models to capture latent representations of users and items; thereby they can simulate missing interactions by extrapolating from known patterns. This allows for better generalization even when historical data is limited, effectively addressing the sparsity issue.

In cold-start situations, GAI-RS can create enriched profiles of users or items by using auxiliary information from different sources. For instance, user preferences can be derived from very small activity signals or demographic information; recommendations for new products can be generated from product descriptions, visual content, or embeddings of similar items. Even in the total absence of structured interaction data, generative models can rely on pretrained language models or multimodal embeddings that have learned generalizable representations across domains.

One particularly exciting recent development in this area is the application of LLMs such as GPT and BERT for zero-shot and few-shot recommendation tasks. Given that these models are pretrained on enormous corpora, they should be able to generalize from very little contextual input and make useful predictions without relying on specific historical data. Hence, they can perform effectively even for unseen users or items, providing highly personalized outputs based on user queries, descriptions, or conversation context.

Besides inference and interaction simulation, GAI-RS also has the capability of real-time adaptation [72]. In interactive contexts, like session-based shopping or conversational RSs, these models can continuously adjust their interpretation of user intent and accordingly produce dynamic recommendations. This makes them particularly well suited for situations in which user preferences change rapidly or where explicit feedback is scarce.

Hybrid architecture also emerges as a great solution capturing the advantages of traditional models and generative ones. In particular, a GAI model could be applied to produce synthetic user-item interactions such that the generated interactions are integrated into a collaborative filtering framework aimed at bettering matrix completion. Additionally,

content-based features that are obtained through generative models can be combined with graph-based RS for increased robustness as well as diversity.

GAI allows a dynamic and imaginative take on recommendations, were indeed, the lack of data cannot be seen as a fundamental obstacle anymore. By creating what is not yet there, tuning dynamically, and generalizing even over sparse contexts, GAIRS can effectively address one of the most historical problems within the field. Thus, they are in e-commerce environments increasingly complex and often data limited as key enablers for delivering personalized experiences.

7.3 Handling Diversity

Traditional RSs, which include collaborative filtering and matrix factorization-based models, are typically optimized according to some accuracy, RMSE, or precision@k. Although these predict which items a user is likely to enjoy very effectively, they can recommend only popular or very similar items to what the user has already consumed; thus, creating a filter around the user. This phenomenon limits users' exploration of content and can lead to recommendation fatigue: When users are exposed to repeated variations of the same content, their long-term engagement and satisfaction decrease.

Due to their inherent probabilistic and generative nature, GAI-RS support diversity as well as novelty in recommendations. Such models are capable of sampling from a huge latent space of possible items or contents; hence, they can generate or retrieve recommendations that are not only relevant but also diverse and surprising.

From a technical standpoint, GAI-RS achieve this in several ways:

- Stochastic Sampling: In techniques such as top-k sampling, nucleus sampling (top-p), or temperature scaling, the generated items are made less frequent or more non-obvious by allowing items that are contextually relevant but go beyond what is normally expected [73].
- Conditional Generation: By conditioning on user history, preferences, or context embeddings, a GAI-RS can produce items more relevant to the user's intent but still explore lesser-known areas of the content space [74].
- Diversity-Conscious Objectives: Training objectives can be adapted to introduce a penalty for redundancy and promote the selection of diverse items, such as by incorporating maximal marginal relevance (MMR), submodular optimization, or contrastive loss functions that are required to be far from each other for similar recommendations [75].
- Multi-Objective Optimization: GAI-RS can be configured to simultaneously pursue several objectives such as improving accuracy, diversity, serendipity, and novelty through the application of reinforcement learning (e.g., reward shaping), Pareto optimization, or hybrid scoring systems [76].
- Newness in the Content Generated: Unlike the traditional RS, generative models have the capability of producing new content (text, images, products) that is tailored to the tastes of a user [77]. Such content is never part of the catalogue; hence novelty and interest are perceived to be higher.

Thus, GAI-RS not only enlarges the recommendation space but also provides a more flexible adaptation to changing user interests and long-tail content. Thus, it offers a richer and more engaging experience to the user, which finally fosters discovery, reduces churn, and increases the overall robustness of the system.

7.4 Explainability and User Trust

Traditional RSs, including collaborative filtering and content-based models, provide a simple and often rule-based explanation for the recommendations offered. Such explanations in the form of "users who bought this item also bought that" or "recommended because you watched X" are straight out related to the core heuristics or similarity measures defining the model. Although these justifications are very shallow and non-personalized, they grant not only transparency but also a high degree of interpretability. The user easily gains knowledge about the recommendation's rationale, which fosters trust, especially when the system's decisions are consistent and predictable.

In contrast, GAI-RS incorporating LLMs offers much greater flexibility and customization in the generation of explanations [78]. Such systems can dynamically craft elaborate, natural language justifications by considering a much

larger context: user history, inferred preferences, temporal trends, and even subtle signals gleaned from conversations or implicit feedback. For instance, an LLM-enabled RS might produce an explanation like: “Considering your recent interest in Scandinavian crime novels and the high ratings you have given to psychological thrillers, we think you would enjoy this new release from a Norwegian author.”

However, despite their fluency and contextual richness, explanations produced by LLMs are uniquely susceptible to a fundamental problem: a disconnection between the generation process and the underlying recommendation logic. Most LLMs are not at all grounded in the model's internal reasoning or feature space; therefore, their outputs may sound plausible yet embody completely wrong reasons—what can be called explanation hallucination. Such unfaithfulness cannot but mislead users, thereby shaking confidence whenever inconsistency arise and hiding any possible biases embedded in the recommendation engine.

Current research focuses on ensuring these natural language explanations are aligned with the true decision-making process of the recommendation model. Post-hoc alignment using attention weights, training explanation models jointly with recommendation models, and integration of symbolic reasoning components are all attempts to improve on this. However, the simultaneous fluency in interpretability and truthfulness of the fact in the explanation has been a problematic challenge, especially in high-stakes domains like healthcare, finance, or education where transparency and accountability are paramount.

As GAI-RS continues to evolve, explainability will be a key factor in building user trust, especially in cases where the reasoning was opaque or non-deterministic. Thus, strong mechanisms for faithful, user-aligned, and context-sensitive explanations will be crucial towards the wider adoption of GAI in RSs.

7.5 Real-Time Adaptability and Interaction Modelling

Conventional RSs have been designed to use batch learning or incremental updates at fixed time intervals [79]. These systems are a step later in processing user feedback like clicks, purchases, or ratings by adding it into the user profile or collaborative matrix during periodic retraining cycles. This makes them lag in personalization and thus ineffective in quickly changing user scenarios like dynamic preferences during a browsing session or transient interests such as searching for a gift.

Moreover, traditional frameworks portray user engagement as rigid and uniform, hence they do not capture the dynamic and temporal characteristics of a user's choice process. Although there have been some significant advancements in session-based recommendations or reinforcement learning tools, flexibility is still limited by computational overheads and model rigidity.

GAI-RS consumes streaming user input in real time, including text, clicks, and contextual metadata, thereby updating their internal state representations accordingly. This allows for fine-grained, on-the-fly personalization since the model adapts its recommendations dynamically based on the latest signals from the user.

Moreover, GAI-RS support multi-turn interaction modelling, thus enabling them to conduct long-term dialogues with users. For instance, in the case of conversational RSs, they can simply maintain and update a belief state or preference model during an interaction session through methods based on attention-oriented context tracking, memory networks, or reinforcement learning. These systems can ask clarifying questions to disambiguate user intents and iteratively refine suggestions based on natural language exchanges. A virtual shopping assistant might say something like "I'm looking for something more affordable" or "Show me options with eco-friendly materials," and it would immediately adjust the recommendations accordingly.

Such an interactive feedback loop makes the recommendation process move from a static prediction task to a dynamic, user-in-the-loop system. The model adapts not only to explicit input but also can infer latent intent from subtle behavioural cues. The capabilities discussed here are enabled by architectures such as sequence-to-sequence models, dialogue policy networks, and transformer encoders trained on rich interaction histories; pretraining on large-scale conversational data aids generalization to unseen interaction patterns.

However, real-time adaptability comes with technical challenges. These include:

- Latency constraints for inference in live systems
- Efficient memory management to retain session-level context without performance degradation

- Robustness to noise or ambiguous input, especially in natural language
- Continual learning frameworks that update models incrementally without catastrophic forgetting

Despite these challenges, the combination of GAI with interactive modelling in RSs has been encouraging for applications that need in-session personalization, like digital shopping assistants, live content discovery, and personalized tutoring systems. With the ongoing development of transformer-based models, especially their new instances such as retrieval-augmented generation (RAG) and parameter-efficient fine-tuning (PEFT), the static versus truly adaptive recommendation system dichotomy is fading fast.

7.6 Computational Efficiency and Scalability

Traditional RS often rely heavily on sparse representations, linear models, and precomputed similarity scores make these systems computationally inexpensive while highly effective in large-scale environments [80]. Thus, they can provide recommendations to millions of users across millions of items with relatively modest hardware requirements. Moreover, such models are generally easy to parallelize, cache; deploy with standard serving frameworks, which makes them a preferred choice in latency-sensitive production environments like e-commerce sites or streaming platforms.

GAI-based RSs have a very different computational profile compared to traditional RS. GAI-RS typically employ deep neural architectures, such as transformers, graph neural networks (GNNs), VAEs) and GANs [64]. The number of parameters in models like LLMs or sequence-to-sequence transformers is exorbitant, reaching billions of weights. Consequently, the resources consumed during training and inference are magnified orders of magnitude.

GAI-RS requires large-scale data and powerful hardware accelerators such as GPUs or TPUs during training [81]. Training procedures involve intricate data preprocessing, fine-tuning, and often distributed training pipelines to manage model size and memory requirements. In addition, the non-linearity and depth of these architectures make them less interpretable and harder to optimize in terms of compute-to-performance ratio.

In inference, GAI-RS are also much slower and memory-intensive. Real-time generation of tailored recommendations, especially in the case of multi-turn interactions or dialogue-driven settings, may require multiple forward passes through the network, computations of attention, and contextual encodings. Therefore, these needs impose on infrastructure—from large-scale model hosting to memory-efficient model compression and intelligent caching strategies to meet the requirements for real-time performance.

To mitigate these challenges, the following techniques can be adopted.

- Model distillation (to create smaller, faster student models)
- Quantization along with pruning for the reduction of model size and enhancement of inference speed.
- Parameter-efficient fine-tuning (e.g., LoRA, adapters)
- Retrieval-augmented generation (RAG), a technique intended to transfer some of the knowledge from the model to an external memory, is under active research for balancing capability with practical deployment.

Moreover, horizontal scalability which was a classic strength is not so easy with GAI-RS because of the need to share and manage a common state (like user embeddings or session context) and also due to the expensive inference that is context-aware. Serving models at any global scale with reasonable latency and uptime requires specialized serving infrastructure, such as model parallelism, dynamic batching, and hardware accelerators, thus increasing operational complexity and cost.

In practical applications, this results in a significant trade-off: while GAI-RS can provide richer, more personalized and context-aware recommendations, their computational demands might be out of line with any tight performance or cost constraints. This is leading to hybrid systems within which lightweight models handle candidate retrieval and GAI components are applied in a selective manner for re-ranking or personalized explanation thus providing a compromise between the quality of the recommendation and operational viability. Table 4 contrasts the models across key dimensions.

Table 4. Contrast between Traditional RS and GAI-RS

Dimension	Traditional RS	GAI-RS
Representation Learning	Use fixed, simple vectors based on past interactions.	Use deep neural networks to learn rich, flexible, and non-linear representations.
Data Sparsity & Cold Start	Struggle with sparsity and new users/items. Use basic metadata to guess preferences.	Handle missing data better by generating new interactions and using other info (e.g.: text, images).
Diversity	Recommend similar/popular items but may feel repetitive.	Can generate surprising and varied recommendations using smart sampling and multi-goal training.
Explainability & User Trust	Provide simple, rule-based reasons.	Can explain in natural language but might sound right without truly reflecting how the choice was made.
Real-Time Adaptability & Interaction Modelling	Update slowly; limited dynamic user modelling.	Update instantly based on user actions or conversations; can adjust in real-time.
Computational Efficiency & Scalability	Fast, light, easy to use at scale; require low computational cost	Resource-heavy; slower, especially for live interactions; more complex to run.

The fundamental change brought by the current developments in ML is demonstrated by contrasting the GAI-RS with traditional RS. Although traditional RS are interpretable and computationally efficient, they still suffer from cold-start issues, limited variety, static user modelling, and sparse data. On the other hand, GAI-RS employs strong neural architectures such as VAEs, Transformers, and LLMs to learn rich user-item representations, handle sparse and cold-start data by generating synthetic interactions. Besides that, GAI-RS can also encourage a variety of unconventional suggestions, and allow for dynamic, real-time personalization through multi-turn interactions. However, due to the discrepancy between model logic and produced outputs, these advantages come at the expense of more computing complexity, slower inference times, and difficulties in offering accurate explanations. Overall, GAI-RS offer superior adaptability and personalization but require significant resources and careful design to ensure transparency and efficiency.

8. CONCLUSION

In a nutshell, this study demonstrates how GAI can improve RSs in the e-commerce sector. This can be done by tackling issues of traditional RSs such data sparsity, cold-start issues, and variety shortage in suggestions. GAI-powered systems can produce suggestions that are richer, more precise, and tailored, greatly increasing user happiness and engagement by utilizing cutting-edge methods such as transformer-based models, GANs, and VAEs. By comprehending intricate user-item interactions, these models not only can improve suggestion quality but are also capable of offering a more flexible and dynamic recommendation process, and this has completely changed the customized shopping experience. The future of AI-driven recommendation systems in online retail is assured by the promise that GAI integration into e-commerce platforms would maximize user pleasure, improve product discovery, and generate more income. Notably, although incorporating GAI into RS offers fascinating possibilities, there are still several real-world and moral issues that need to be taken into consideration. For example, deep generative models are known to be expensive and difficult to implement, hence leading them to be impractical for some e-commerce platforms, especially those with limited resources or scale. Furthermore, many GAI models limit the transparency and make it more difficult to understand and confirm suggestions, and this might cause a negative impact to user confidence. Bias and fairness are also major issues since generative models trained on biased data may marginalize particular user groups or perpetuate prejudices and produce suggestions that might be unjust or discriminatory.

Future research will include focusing more on investigating hybrid models. Normally, a hybrid model is designed to combine the advantage of two or more methods. For example, a combination of multimodal data might bring more

insights into this study. This could lead to a better personalization and understanding of user preferences. Besides that, exploring fairness and transparency in recommendations is also a good idea for future research. This step is essential to ensure diverse and unbiased suggestions for all users. Furthermore, more scalable and effective solutions should be considered so that the smaller companies can also use GAI without spending intense resources. For instance, user recommendations can be updated to reflect changing user preferences by including real-time feedback loops. Last but not least, it is critical to routinely evaluate the ethical implications of GAI-driven recommendations to ensure that they fully support ethical AI practices while improving online purchasing experience. The changes brought by GAI to e-commerce RSs are required to be made in a way that is ethical, inclusive, and sustainable by complying with these suggestions.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Wan-Er Kong: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Tong-Ern Tai: Project Administration, Writing – Review & Editing;
Palanichamy Naveen: Project Administration, Supervision, Writing – Review & Editing;
Kok-Why Ng: Project Administration, Supervision, Writing – Review & Editing;
Lucia Dwi Krisnawati: Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENT

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline, https://publicationethics.org/.

REFERENCES

[1] S.K.R. Sharma, and S. Gaur, "The role of artificial intelligence in personalized e-commerce recommendations," *International Journal for Research Publication and Seminar*, vol. 15, no. 1, pp. 64–71, Mar. 2024, doi: 10.36676/jrps.v15.i1.09.

[2] Q. Zhang and Y. Xiong, "Harnessing AI potential in e-commerce: Improving user engagement and sales through deep learning-based product recommendations," *Current Psychology*, vol. 43, no. 38, pp. 30379–30401, Oct. 2024, doi: 10.1007/s12144-024-06649-3.

[3] H. Yuan, and A. A. Hernandez, "User cold start problem in recommendation systems: A systematic review," *IEEE Access*, vol. 11, pp. 136958–136977, 2023, doi: 10.1109/ACCESS.2023.3338705.

[4] M. Jangid, and R. Kumar, "Deep learning approaches to address cold start and long tail challenges in recommendation systems: A systematic review," *Multimedia Tools and Applications*, vol. 84, no. 5, pp. 2293–2325, Oct. 2024, doi: 10.1007/s11042-024-20262-3.

[5] X. Ma, M. Li, and X. Liu, "Advancements in recommender systems: A comprehensive analysis based on data, algorithms, and evaluation," unpublished, Jul. 2024.

[6] Q. Wang et al., "Towards next-generation LLM-based recommender systems: A survey and beyond," unpublished, Oct. 2024.

[7] M.O. Ayemowa, R. Ibrahim, and M.M. Khan, "Analysis of recommender system using generative artificial intelligence: A systematic literature review," *IEEE Access*, vol. 12, pp. 87742–87766, 2024, doi: 10.1109/ACCESS.2024.3416962.

[8] D.P. Gohel and P. A. Vanjara, "Analyzing user-based and item-based recommender systems: A comparative examination," *Educational Administration: Theory and Practice*, vol. 30, no. 6(S), Jun. 2024, doi: 10.53555/kuey.v30i6(S).5331.

[9] T. Anwar, V. Uma, Md. I. Hussain, and M. Pantula, "Collaborative filtering and KNN based recommendation to overcome cold start and sparsity issues: A comparative analysis," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 35693–35711, Oct. 2022, doi: 10.1007/s11042-021-11883-z.

[10] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A.H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data," *Expert Systems with Applications*, vol. 149, pp. 113248, Jul. 2020, doi: 10.1016/j.eswa.2020.113248.

[11] M. Liao and S.S. Sundar, "When e-commerce personalization systems show and tell: Investigating the relative persuasive appeal of content-based versus collaborative filtering," *Journal of Advertising*, vol. 51, no. 2, pp. 256–267, Mar. 2022, doi: 10.1080/00913367.2021.1887013.

[12] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data*, vol. 9, no. 1, pp. 59, Dec. 2022, doi: 10.1186/s40537-022-00592-5.

[13] Y. Xu, E. Wang, Y. Yang, and H. Xiong, "GS²-RS: A generative approach for alleviating cold start and filter bubbles in recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2023, doi: 10.1109/TKDE.2023.3290140.

[14] H. Lu et al., "DeepSeek-VL: Towards real-world vision-language understanding," unpublished, Mar. 2024.

[15] M.N.-U.-R. Chowdhury, A. Haque, and I. Ahmed, "DeepSeek vs. ChatGPT: A comparative analysis of performance, efficiency, and ethical ai considerations," *TechRxiv Preprint*, Feb. 11, 2025, doi: 10.36227/techrxiv.173929663.35290537/v1.

[16] A. Rasool, M.I. Shahzad, H. Aslam, V. Chan, and M.A. Arshad, "Emotion-aware embedding fusion in Large Language Models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for intelligent response generation," *AI*, vol. 6, no. 3, pp. 56, Mar. 2025, doi: 10.3390/ai6030056.

[17] R. Gupta, S. Tiwari, and P. Chaudhary, "Computational foundation of generative AI models," in *Generative AI: Concepts and Applications*, Springer, 2025, pp. 23–44, doi: 10.1007/978-3-031-82062-5_2.

[18] Y. Deldjoo et al., "Recommendation with generative models," unpublished, Sep. 2024.

[19] D. Liang, R.G. Krishnan, M.D. Hoffman, and T. Jebara, "Variational Autoencoders for collaborative filtering," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW '18)*, New York, NY, USA: ACM Press, pp. 689–698, 2018, doi: 10.1145/3178876.3186150.

[20] R. Nahta, G. S. Chauhan, Y.K. Meena, and D. Gopalani, "Deep learning with the generative models for recommender systems: A survey," *Computer Science Review*, vol. 53, pp. 100646, Aug. 2024, doi: 10.1016/j.cosrev.2024.100646.

[21] J. Liu, C. Shi, C. Yang, Z. Lu, and P.S. Yu, "A survey on heterogeneous information network based recommender systems: Concepts, methods, applications and resources," *AI Open*, vol. 3, pp. 40–57, 2022, doi: 10.1016/j.aiopen.2022.03.002.

[22] K.I. Roumeliotis, N.D. Tselikas, and D.K. Nasiopoulos, "Precision-driven product recommendation software: unsupervised models, evaluated by GPT-4 LLM for enhanced recommender systems," *Software*, vol. 3, no. 1, pp. 62–80, Feb. 2024, doi: 10.3390/software3010004.

[23] A.Y. Tsai et al., "Leveraging LLM reasoning enhances personalized recommender systems," unpublished, Jul. 2024.

[24] J. Lin et al., "How can recommender systems benefit from Large Language Models: A Survey," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–47, Mar. 2025, doi: 10.1145/3678004.

[25] Z. Gao et al., "Mitigating the filter bubble while maintaining relevance," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2022, pp. 2524–2531, doi: 10.1145/3477495.3531890.

[26] D. Antognini and B. Faltings, "Fast multi-step critiquing for VAE-based recommender systems," in *Proceedings of the Fifteenth ACM Conference on Recommender Systems*, New York, NY, USA: ACM, pp. 209–219, Sep. 2021, doi: 10.1145/3460231.3474249.

[27] I. Shenbin, A. Alekseev, E. Tutubalina, V. Malykh, and S.I. Nikolenko, "RecVAE: A new Variational Autoencoder for Top-N recommendations with implicit feedback," *ACM Conference Paper*, Dec. 2019, doi: 10.1145/3336191.3371831.

[28] W. Ma, X. Chen, W. Pan, and Z. Ming, "VAE++: Variational Autoencoder for heterogeneous one-class collaborative filtering," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, pp. 666–674, Feb. 2022, doi: 10.1145/3488560.3498436.

[29] W. Shafqat and Y.-C. Byun, "A hybrid GAN-based approach to solve imbalanced data problem in recommendation systems," *IEEE Access*, vol. 10, pp. 11036–11047, 2022, doi: 10.1109/ACCESS.2022.3141776.

[30] E. Dervishaj, and P. Cremonesi, "GAN-based matrix factorization for recommender systems," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA: ACM, Apr. 2022, pp. 1373–1381, doi: 10.1145/3477314.3507099.

[31] P. Chonwiharnphan, P. Thienprapasith, and E. Chuangsuwanich, "Generating realistic users using Generative Adversarial Network with recommendation-based embedding," *IEEE Access*, vol. 8, pp. 41384–41393, 2020, doi: 10.1109/ACCESS.2020.2976491.

[32] F. Zhou, Y. Mo, G. Trajcevski, K. Zhang, J. Wu, and T. Zhong, "Recommendation via collaborative autoregressive flows," *Neural Networks*, vol. 126, pp. 52–64, Jun. 2020, doi: 10.1016/j.neunet.2020.03.010.

[33] P. Zhang et al., "TransGNN: Harnessing the collaborative power of transformers and Graph Neural Networks for recommender systems," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, pp. 1285–1295, Jul. 2024, doi: 10.1145/3626772.3657721.

[34] Z.M. Ziegler and A.M. Rush, "Latent normalizing flows for discrete sequences," unpublished, Jan. 2019.

[35] F. Wang, W. Liu, C. Chen, M. Zhu, and X. Zheng, "HCFRec: Hash collaborative filtering via normalized flow with structural consensus for efficient recommendation," unpublished, May 2022.

[36] H. Liu, Y. Wei, X. Song, W. Guan, Y.-F. Li, and L. Nie, "MMGRec: Multimodal generative recommendation with transformer model," unpublished, Apr. 2024.

[37] C. Li, L. Xia, X. Ren, Y. Ye, Y. Xu, and C. Huang, "Graph transformer for recommendation," unpublished, Jun. 2023, doi: 10.1145/3539618.3591723.

[38] S. J. Kalaiarasi, and K. Nimala, "Enhancing e-commerce product recommendations using LLMs and transformer-based deep learning architectures," in *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, IEEE, pp. 1–8, Dec. 2024, doi: 10.1109/ICSES63760.2024.10910646.

[39] J. Li, W. Zhang, T. Wang, G. Xiong, A. Lu, and G. Medioni, "GPT4Rec: A generative framework for personalized recommendation and user interests interpretation," unpublished, Apr. 2023.

[40] V. Chamola, S. Sai, R. Sai, A. Hussain, and B. Sikdar, "Generative AI for consumer electronics: Enhancing user experience with cognitive and semantic computing," *IEEE Consumer Electronics Magazine*, vol. 14, no. 2, pp. 10–19, Mar. 2025, doi: 10.1109/MCE.2024.3387049.

[41] I. Karabila, N. Darraz, A. El-Ansari, N. Alami, and M. El Mallahi, "BERT-enhanced sentiment analysis for personalized e-commerce recommendations," *Multimedia Tools and Applications*, vol. 83, pp. 56463–56488, Dec. 2023, doi: 10.1007/s11042-023-17689-5.

[42] Z. Zhang et al., "Co-optimize content generation and consumption in a Large Scale video recommendation system," in *Proceedings of the 18th ACM Conference on Recommender Systems*, New York, NY, USA: ACM, pp. 762–764, Oct. 2024, doi: 10.1145/3640457.3688033.

[43] T. Li et al., "BeautyGAN: Instance-Level facial makeup transfer with deep Generative Adversarial Network," in *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA: ACM, pp. 645–653, Oct. 2018, doi: 10.1145/3240508.3240618.

[44] K. N, R. S. K. V, P. R. Rasal, R.J. Jadhav, M. Saidireddy, and K.G. Kharade, "Artificial intelligence-based smart cosmetics suggestion system based on skin condition," in *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, pp. 797–801, Dec. 2022, doi: 10.1109/ICACRS55517.2022.10029120.

[45] T. Islam, A. Miron, X. Liu, and Y. Li, "Deep learning in virtual try-on: A comprehensive survey," *IEEE Access*, vol. 12, pp. 29475–29502, 2024, doi: 10.1109/ACCESS.2024.3368612.

[46] Y. Archana, "Gen AI-driven electronics: Innovations, challenges and future prospects," unpublished, 2023.

[47] T. Badriyah, E. T. Wijayanto, I. Syarif, and P. Kristalina, "A hybrid recommendation system for e-commerce based on product description and user profile," in *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, IEEE, pp. 95–100, Aug. 2017, doi: 10.1109/INTECH.2017.8102435.

[48] K. Al Jadda, "Recommendation in home improvement industry, challenges and opportunities," in *Proceedings of the 13th ACM Conference on Recommender Systems*, New York, NY, USA: ACM, pp. 528, Sep. 2019, doi: 10.1145/3298689.3346960.

[49] Q. Zhang, J. Lu, and Y. Jin, "Artificial intelligence in recommender systems," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 439–457, Feb. 2021, doi: 10.1007/s40747-020-00212-w.

[50] M. Mendoza and N. Torres, "Evaluating Content Novelty in Recommender Systems," *Journal of Intelligent Information Systems*, vol. 54, no. 2, pp. 297–316, Apr. 2020, doi: 10.1007/s10844-019-00548-x.

[51] Y.-M. Tamm, R. Damdinov, and A. Vasilev, "Quality metrics in recommender systems: Do we calculate metrics consistently?," in *Proceedings of the Fifteenth ACM Conference on Recommender Systems*, New York, NY, USA: ACM, pp. 708–713, Sep. 2021, doi: 10.1145/3460231.3478848.

[52] G. Shani and A. Gunawardana, "Evaluating recommendation systems," in *Recommender Systems Handbook*, Boston, MA: Springer US, pp. 257–297, 2011, doi: 10.1007/978-0-387-85820-3_8.

[53] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," unpublished, Feb. 2014.

[54] J. Chen, A. Muscoloni, I. Abdelhamid, Y. Wu, and C. V. Cannistraci, "Generalizing the AUC-ROC for unbalanced data, early retrieval and link prediction evaluation," *Preprints*, Jul. 22, 2024, doi: 10.20944/preprints202209.0277.v2.

[55] C. C. Aggarwal, "Evaluating recommender systems," in *Recommender Systems*, Cham: Springer International Publishing, pp. 225–254, 2016, doi: 10.1007/978-3-319-29659-3_7.

[56] A. Drif, H.E. Zerrad, and H. Cherifi, "EnsVAE: Ensemble Variational Autoencoders for recommendations," *IEEE Access*, vol. 8, pp. 188335–188351, 2020, doi: 10.1109/ACCESS.2020.3030693.

[57] J. Wang, G. Liu, J. Wu, C. Jia, and Z. Zhang, "Self-supervised Variational Autoencoder for recommender systems," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, pp. 831–835, Nov. 2021, doi: 10.1109/ICTAI52525.2021.00132.

[58] J. Liu, Y. Xiao, K. Zhu, W. Zheng, and C.-H. Hsu, "Hybrid Variational Autoencoder for collaborative filtering," in *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, pp. 251–256, May 2022, doi: 10.1109/CSCWD54268.2022.9776247.

[59] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang, "M6-Rec: Generative pretrained language models are open-ended recommender systems," unpublished, May 2022.

[60] J. Feng, Q. Wang, Z. Huang, and L. Yang, "ICCVAE: Item concept causal Variational Auto-encoder for Top-N recommendation," in *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, IEEE, pp. 908–913, Apr. 2023, doi: 10.1109/ICSP58490.2023.10248832.

[61] C.-H. Zhou and Y.-L. Chen, "VCGAN: Variational Collaborative Generative Adversarial Network for recommendation systems," in *ICC 2023 - IEEE International Conference on Communications*, IEEE, pp. 6324–6330, May 2023, doi: 10.1109/ICC45041.2023.10278637.

[62] J. Jia et al., "LEARN: Knowledge adaptation from Large Language Model to recommendation for practical industrial application," unpublished, May 2024.

[63] C.-W. Huang, L. Dinh, and A. Courville, "Augmented normalizing flows: Bridging the gap between generative flows and latent variable models," unpublished, Feb. 2020.

[64] M.O. Ayemowa, R. Ibrahim, and M. M. Khan, "Analysis of recommender system using generative artificial intelligence: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 87742–87766, 2024, doi: 10.1109/ACCESS.2024.3416962.

[65] W. Wang, Y. Zhang, and T.-S. Chua, "Recommendation in the era of generative artificial intelligence," in *Generative Artificial Intelligence: Applications and Implications*, Cham: Springer, pp. 201–221, 2025, doi: 10.1007/978-3-031-73147-1_8.

[66] N. Ohsaka and R. Togashi, "Curse of 'low' dimensionality in recommender systems," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, pp. 537–547, Jul. 2023, doi: 10.1145/3539618.3591659.

[67] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative AI: A comprehensive review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," *IEEE Access*, vol. 12, pp. 69812–69837, 2024, doi: 10.1109/ACCESS.2024.3397775.

[68] I. Pesovski, R. Santos, R. Henriques, and V. Trajkovik, "Generative AI for customizable learning experiences," *Sustainability*, vol. 16, no. 7, pp. 3034, Apr. 2024, doi: 10.3390/su16073034.

[69] Z. Chen, W. Gan, J. Wu, K. Hu, and H. Lin, "Data scarcity in recommendation systems: A survey," *ACM Transactions on Recommender Systems*, vol. 3, no. 3, pp. 1–31, Sep. 2025, doi: 10.1145/3639063.

[70] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data*, vol. 9, no. 1, pp. 59, Dec. 2022, doi: 10.1186/s40537-022-00592-5.

[71] H.H. Rashidi et al., "Statistics of Generative Artificial Intelligence and nongenerative predictive analytics machine learning in medicine," *Modern Pathology*, vol. 38, no. 3, pp. 100663, Mar. 2025, doi: 10.1016/j.modpat.2024.100663.

[72] H. Du et al., "The age of Generative AI and AI-generated everything," *IEEE Network*, vol. 38, no. 6, pp. 501–512, Nov. 2024, doi: 10.1109/MNET.2024.3422241.

[73] H. Cao et al., "A Survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 2814–2830, Jul. 2024, doi: 10.1109/TKDE.2024.3361474.

[74] M. Ibrahim et al., "Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges," *Computers in Biology and Medicine*, vol. 189, pp. 109834, May 2025, doi: 10.1016/j.compbiomed.2025.109834.

[75] J. Chan, and Y. Li, "Enhancing team diversity with Generative AI: A novel project management framework," in *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, pp. 1648–1652, Jul. 2024, doi: 10.1109/COMPSAC61105.2024.00259.

[76] Q. Dang, G. Zhang, L. Wang, S. Yang, and T. Zhan, "A Generative Adversarial Networks model based evolutionary algorithm for multimodal multi-objective optimization," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–10, 2024, doi: 10.1109/TETCI.2024.3397996.

[77] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, Feb. 2024, doi: 10.1007/s12599-023-00834-7.

[78] X. Pan, "Enhancing efficiency and innovation with Generative AI," *Journal of Artificial Intelligence and Autonomous Intelligence*, vol. 1, no. 1, pp. 72–81, 2024, doi: 10.54364/JAIAI.2024.1105.

[79] Y. Xi et al., "Efficient and deployable knowledge infusion for open-world recommendations via Large Language Models," *ACM Transactions on Recommender Systems*, Mar. 2025, doi: 10.1145/3725894.

[80] R. Alatrash, R. Priyadarshini, and H. Ezaldeen, "Collaborative filtering integrated fine-grained sentiment for hybrid recommender system," *The Journal of Supercomputing*, vol. 80, no. 4, pp. 4760–4807, Mar. 2024, doi: 10.1007/s11227-023-05600-w.

[81] Y. Park et al., "Inference optimization of foundation models on AI accelerators," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, pp. 6605–6615, Aug. 2024, doi: 10.1145/3637528.3671465.

BIOGRAPHIES OF AUTHORS

Wan-Er Kong is a master's student at Multimedia University, currently pursuing a Master of Science. Holding a degree in Computer Science with a specialization in Data Science from the same university, Wan Er is passionate about exploring the field of Data Science. Her research focuses on Machine Learning and recommender systems. Driven by a desire to contribute to society, she is dedicated to advancing her knowledge and skills to make meaningful impacts in the world through Data Science. She can be contacted at email: KONG.WAN.ER@student.mmu.edu.my.

Tong-Ern Tai is currently studying for a Master of Science degree at Multimedia University. Previously graduated from the same institution with a degree in computer science (specialization in data science). Her work focuses on helpdesk ticket completion time prediction and machine learning. She can be contacted at email: TAI.TONG.ERN@student.mmu.edu.my.

Palanichamy Naveen joined the Faculty of Computing and Informatics, Multimedia University after receiving Ph.D. from Curtin University, Malaysia. She received her Bachelor of Engineering (CSE) and Master of Engineering (CSE) from Anna University, India. Her research interest includes Smart Grid, Cloud Computing, Machine Learning, Deep Learning and Recommender system. She is involved in multiple research projects funded by Multimedia University. She can be contacted at email: p.naveen@mmu.edu.my.

Kok-Why Ng is an Associate Professor in the Faculty of Computing and Informatics (FCI) in Multimedia University (MMU), Malaysia. His research interests are in Recommender System, Artificial Intelligence, Deep Learning, Image Processing, 3D Modelling and Segmentation. He leads several research grants, collaborates with some universities and companies. He can be contacted at email: kwng@mmu.edu.my.

Lucia Dwi Krisnawati is an associate professor at Informatics Department, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia. She received her Doktor der Philosophie (Dr.phil.) as well as master's degree in Natural Language Processing from Ludwig-Maximilian Universität, Munich, Germany. Based on this, her research interests include Natural Language Processing (Text Similarity, Forensic Text, Dialogue System, recommender system), Optical Character Recognition (OCR), Machine Learning, and Human-Computer Interaction. Recently, she is interested in applying those research areas for educational applications. She can be contacted at email: krisna@staff.ukdw.ac.id.