A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys)

Yashar Deldjoo
Polytechnic University of Bari
Bari, Italy
deldjooy@acm.org

Zhankui He
University of California
La Jolla, USA
zhh004@ucsd.edu

Julian McAuley
University of California
La Jolla, USA
jmcauley@ucsd.edu

Anton Korikov
University of Toronto
Toronto, Canada
anton.korikov@mie.utoronto.ca

Scott Sanner
University of Toronto
Toronto, Canada
ssanner@mie.utoronto.ca

Arnau Ramisa
Amazon[*]
Palo Alto, USA
aramisay@amazon.com

René Vidal
Amazon[*]
Palo Alto, USA
vidalr@seas.upenn.edu

Maheswaran Sathiamoorthy
Bespoke Labs
Santa Clara, USA
mahesh@bespokelabs.ai

Atoosa Kasirzadeh
University of Edinburgh
Edinburgh, UK
atoosa.kasirzadeh@gmail.com

Silvia Milano
University of Exeter and LMU Munich
Munich, Germany
milano.silvia@gmail.com

ABSTRACT

Traditional recommender systems typically use user-item rating histories as their main data source. However, deep generative models now have the capability to model and sample from complex data distributions, including user-item interactions, text, images, and videos, enabling novel recommendation tasks. This comprehensive, multidisciplinary survey connects key advancements in RS using Generative Models (Gen-RecSys), covering: interaction-driven generative models; the use of large language models (LLM) and textual data for natural language recommendation; and the integration of multimodal models for generating and processing images/videos in RS. Our work highlights necessary paradigms for evaluating the impact and harm of Gen-RecSys and identifies open challenges. This survey accompanies a **tutorial** presented at ACM KDD'24, with supporting materials provided at: https://encr.pw/vDhLq.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Generative Models, Recommender Systems, GANs, VAEs, LLMs, Multimodal, vLLMs, Ethical and Societal Considerations

[*] This work does not relate to the author's position at Amazon.

KDD '24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671474

ACM Reference Format:
Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671474

1 INTRODUCTION

Advancements in generative models have significantly impacted the evolution of recommender systems (RS). Traditional RS, which relied on capturing user preferences and item features within a specific domain — often referred to as "*narrow experts*" – are now being complemented and, in some instances, surpassed by generative models. These models have introduced innovative ways of conceptualizing and implementing recommendations. Specifically, modern generative models learn to represent and sample from complex data distributions, including not only user-item interaction histories but also text and image content, unlocking these data modalities for novel and interactive recommendation tasks.

Moreover, advances in natural language processing (NLP) through the introduction of large language models (LLMs) such as ChatGPT [121] and Gemini [148] have showcased remarkable *emergent* capabilities [165], including reasoning, in-context few-shot learning, and access to extensive open-world information within their pre-trained parameters. Because of their broad generalist abilities, these pretrained generative models have opened up an exciting new research space for a wide variety of recommendation applications (see Table 1), e.g., enhanced personalization, improved conversational interfaces, and richer explanation generation, among others.

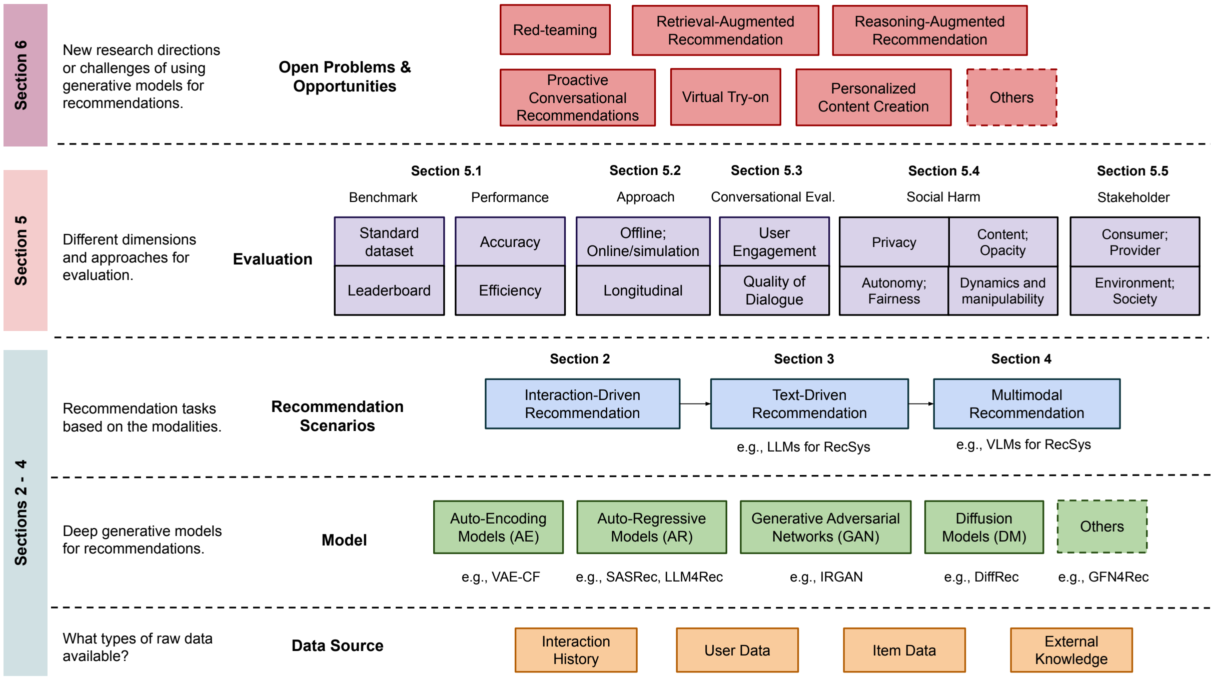
Section 6
New research directions or challenges of using generative models for recommendations.
Open Problems & Opportunities
Red-teaming
Retrieval-Augmented Recommendation
Reasoning-Augmented Recommendation
Proactive Conversational Recommendations
Virtual Try-on
Personalized Content Creation
Others
Section 5
Different dimensions and approaches for evaluation.
Evaluation
Section 5.1
Benchmark
Performance
Standard dataset
Leaderboard
Accuracy
Efficiency
Section 5.2
Approach
Offline; Online/simulation
Longitudinal
Section 5.3
Conversational Eval.
User Engagement
Quality of Dialogue
Section 5.4
Social Harm
Privacy
Content; Opacity
Autonomy; Fairness
Dynamics and manipulability
Section 5.5
Stakeholder
Consumer; Provider
Environment; Society
Sections 2 - 4
Recommendation tasks based on the modalities.
Recommendation Scenarios
Section 2
Interaction-Driven Recommendation
Section 3
Text-Driven Recommendation
e.g., LLMs for RecSys
Section 4
Multimodal Recommendation
e.g., VLMs for RecSys
Deep generative models for recommendations.
Model
Auto-Encoding Models (AE)
e.g., VAE-CF
Auto-Regressive Models (AR)
e.g., SASRec, LLM4Rec
Generative Adversarial Networks (GAN)
e.g., IRGAN
Diffusion Models (DM)
e.g., DiffRec
Others
e.g., GFN4Rec
What types of raw data available?
Data Source
Interaction History
User Data
Item Data
External Knowledge

Figure 1: Overview of the areas of interest in generative models in recommendation.

The core of generative models lies in their ability to model and sample from their training *data distribution* for various inferential purposes, which enables two primary modes of application for RS:

(1) **Directly trained models.** This approach trains generative models, such as VAE-CF (Variational AutoEncoders for Collaborative Filtering) [97] (cf. Section 2.1) directly on user-item interaction data to predict user preferences, without using large, diverse pretraining datasets. These models learn the probability distribution of items a user might like based on their previous interactions.
(2) **Pretrained models.** This strategy uses models pretrained on diverse data (text, images, videos) to understand complex patterns, relationships, and contexts that often exhibit (emergent) generalization abilities to a range of novel tasks [165]. Among a variety of applications, this survey covers the use of pretrained Gen-RecSys models in the following settings:
    - *Zero- and Few-shot Learning* (cf. Section 3.2.1), using in-context learning (ICL) for broad understanding without extra training.
    - *Fine-Tuning* (cf. Section 3.3), adjusting model parameters using specific datasets for tailored recommendations.
    - *Retrieval-Augmented Generation (RAG)* (cf. Section 3.3), integrating information retrieval with generative modeling for contextually relevant outputs.
    - *Feature Extraction for Downstream Recommendation* (cf. Section 3.4), e.g., generating embeddings or token sequences for complex content representation.
    - *Multimodal Approaches* (cf. Section 4), jointly using multiple data types such as text, image, and video to enhance and improve the recommendation experience.

1.1 Recent Surveys and Our Contributions

Recent Relevant Surveys. Recent surveys have marked significant advancements in the field. We highlight our contributions and distinguish our survey by its comprehensive and unique approach.

- Deldjoo et al. [33] explore GAN-based RS across four different recommendation scenarios (graph-based, collaborative, hybrid, context-aware).
- Li et al. [95] explore training strategies and learning objectives of LLMs for RS.
- Wu et al. [171] discuss both the use of LLMs to generate RS input tokens or embeddings as well as the use of LLMs *as* an RS;
- Lin et al. [99] focus on adapting LLMs in RS, detailing various tasks and applications. Fan et al. [38] overview LLMs in RS, emphasizing pre-training, fine-tuning, and prompting, while Vats et al. [150] review LLM-based RS, introducing a heuristic taxonomy for categorization.
- Huang et al. [67], explore using foundation models (FMs) in RS.
- Wang et al. [158] introduce GeneRec, a next-gen RS that personalizes content through AI generators and interprets user instructions to gather user preferences.

While the mentioned surveys offer crucial insights, their scope is often limited to LLMs [38, 95, 99, 150, 171] or, more broadly, FMs [67] and/or specific models such as GANs [33], without considering the wider spectrum of generative models and data modalities. The work by [158] provides a more relevant survey on Gen-RecSys although their work is mostly on personalized content generation.

Core Contributions. Figure 1 illustrates the structure of our Gen-RecSys survey. It categorizes data sources, recommendation models,

Table 1: Example applications of Gen-RecSys methods.

Description	Relevant Section
Utilize text in RS, including item descriptions, user preferences, reviews, queries, and conversation histories. Examples include generative and conversational recommendations and explanations.	Sections 3.2, 3.5
Using and generating images for recommendation, reasoning, and content creation.	Section 4, Sections 4.4
Applying the emergent reasoning abilities of pre-trained models to recommendation tasks, including in-context learning and tool-augmented reasoning.	Sections 3.3, 3.5
Integrating RS with external knowledge sources through retrieval augmented generation.	Section 3.3
Selecting or generating informative user-item interactions to improve RS model training.	Sections 2.3, 2.4
Generating recommendation results with complex structures such as list-wise or page-wise outputs.	Sections 2.1, 2.2, 2.3, 2.5
Facilitating conversational recommendation through full NL dialogue.	Section 3.5

and scenarios, extending to system evaluation and challenges. We present a **systematic** approach to deconstructing the Gen-RecSys recommendation process into distinct components and methodologies. Our contributions are summarized as follows.

(1) Our survey is broader in scope than the surveys mentioned above, encompassing not just LLMs but a wide array of generative models in RS.
(2) We have chosen to classify these models based on the type of data and modality they are used for, such as user-item data (cf. Section 2), text-driven (cf. Section 3), and multimodal (cf. Section 4) models, as shown in the *Rec. Scenario* layer.
(3) Within each modality discussion, we provide an in-depth exploration of deep generative model paradigms as shown in the *Model* layer, yet with a broader scope that spans multiple contexts and use cases, offering a critical analysis of their roles and effectiveness in respective sections.
(4) We study the evaluation of Gen-RecSys with finer details, shedding light on multiple aspects such as benchmarks, evaluation for impact and harm relative to multiple stakeholders, and conversational evaluation. This evaluation framework is particularly notable as it helps to understand the complex challenges intrinsic to Gen-RecSys.
(5) We discuss several open research challenges and issues. Our survey benefits from the expertise of scholars/industry practitioners from diverse institutions and disciplines.

2 GENERATIVE MODELS FOR INTERACTION-DRIVEN RECOMMENDATION

Interaction-driven recommendation is a setup where only the user-item interactions (e.g., "user A clicks item B") are available, which is the most general setup studied in RS. In this setup, we concentrate on the inputs of user-item interactions and outputs of item-recommended lists or grids rather than richer inputs or outputs from other modalities such as textual reviews. Even though no textual or visual information is involved, generative models [47, 64, 84, 144, 149] still show their unique usefulness. In this section, we examine the paradigms of generative models for recommendation tasks with user-item interactions, including auto-encoding models [84], auto-regressive models [64, 149], generative adversarial networks [47], diffusion models [144] and more.

2.1 Auto-Encoding Models

Auto-encoding models learn to reconstruct their inputs. This capability allows them to be used for various purposes, including denoising, representation learning, and generation tasks.

2.1.1 *Preliminaries: Denoising Auto-Encoding Models.*

Denoising Auto-Encoding models are a group of models that learn to recover the original inputs from a corrupted version of the inputs. Traditionally, denoising auto-encoding models refer to a group of Denoising Autoencoders [140, 151] with hidden layers as a "bottleneck". For example, AutoRec [140] tries to reconstruct the input vector, which is partially observed. More broadly, BERT-like models [35, 146, 172] are also treated as denoising auto-encoding models. Such models recover corrupted (i.e., masked) inputs through stacked self-attention blocks [59, 146]. For example, BERT4Rec [146] is trained to predict masked items in given user historical interaction sequences. Therefore, BERT-like [35] models can be used for next-item prediction in the inference phase [59, 146].

2.1.2 *Variational Auto-Encoding Models.*

Variational Autoencoders (VAEs) are models that learn stochastic mappings from an input x from a often complicated probability distribution p to a probability distribution q. This distribution, q, is typically simple (e.g., a normal distribution), enabling the use of a decoder to generate outputs $\hat{x}$ by sampling from q [84]. VAEs find wide applications in traditional RS, particularly for collaborative filtering [97], sequential recommendation [137] and slate generation [29, 74, 106]. Compared to Denoising Autoencoders, VAEs often demonstrate superior performance in collaborative filtering due to stronger modeling assumptions, such as VAE-CF [97]. Additionally, Conditional VAE (CVAE) [145] models learn distributions of preferred recommendation lists for a given user. This makes them useful for generating those lists beyond a greedy ranking schema. Examples like ListCVAE [74] and PivotCVAE [106] use VAEs to generate entire recommendation lists rather than solely ranking individual items.

2.2 Auto-Regressive Models

Given an input sequence $\mathbf{x}$, at step i, auto-regressive models [12] learn the conditional probability distribution $p(x_i|\mathbf{x}_{<i})$, where $\mathbf{x}_{<i}$ represents the subsequence before step i. Auto-regressive models are primarily used for sequence modeling [12, 36, 149]. In RS,

they find wide applications in session-based or sequential recommendations [63, 80], model attacking [181], and bundle recommendations [7, 66], with recurrent neural networks [7, 63, 66], self-attentive models [80], and more.

2.2.1 Recurrent Auto-Regressive Models. Recurrent neural networks (RNNs) [25, 64] have been use to predict the next item in session-based and sequential recommendations, such as GRU4Rec [63] and its variants [62, 182] (e.g., predicting the next set of items in basket or bundle recommendations, such as set2set [66] and BGN [7]). Moreover, using the auto-regressive generative nature of recurrent networks, researchers extract model-generated user behavior sequences, which are used in the research of model attacking [181].

2.2.2 Self-Attentive Auto-Regressive Models. Self-attentive models replace the recurrent unit with self-attention and related modules, inspired by transformers [149]. This group of models can be used in session-based recommendation and sequential recommendation [80, 100, 124, 170], next-basket or bundle prediction [179], and model attacking [181]. Meanwhile, the benefits of self-attentive models are that they handle long-term dependencies better than RNNs and enable parallel training [149]. Additionally, self-attentive models are the *de-facto* option for pre-trained models [35] and large language models [17, 18, 165], which is gaining traction in RS. More details about using such language models for recommendations will be discussed in Section 3.

2.3 Generative Adversarial Networks

Generative adversarial networks (GANs) [47, 115] are composed of two primary components: a generator network and a discriminator network. These networks engage in adversarial training to enhance the performance of both the generator and the discriminator. GANs are used in RS for multiple purposes [19, 23, 153]. In the interaction-driven setup, GANs are proposed for selecting informative training samples [19, 153], for example, in IRGAN [153, 156], the generative retrieval model is leveraged to sample negative items. Meanwhile, GANs synthesize user preferences or interactions to augment training data [21, 157]. Additionally, GANs have shown effectiveness in generating recommendation lists or pages, such as [23] in whole-page recommendation settings.

2.4 Diffusion Models

Diffusion models [144] generate outputs through a two-step process: (1) corrupting inputs into noise via a forward process, and (2) learning to recover the original inputs from the noise iteratively in a reverse process. Their impressive generative capabilities have attracted growing interest from the RS community.

First, a group of works [152, 159] learns users' future interaction probabilities through diffusion models. For example, DiffRec [159] predicts users' future interactions using corrupted noises from the users' historical interactions. Second, another group of works [104, 173] focuses on diffusion models for training sequence augmentation, showing promising results in alleviating the data sparsity and long-tail user problems in sequential recommendation.

2.5 Other Generative Models

In addition to the previously mentioned generative models, RS also draw upon other types of generative models. For instance, VASER [191] leverages normalizing flows [132] (and VAEs [84]) for session-based recommendation. GFN4Rec [105], on the other hand, adapts generative flow networks [11, 122] for listwise recommendation. Furthermore, IDNP [37] utilizes generative neural processes [43, 44] for sequential recommendation. In summary, various generative models are explored in RS, even in settings without textual or visual modalities.

3 LARGE LANGUAGE MODELS IN RECOMMENDATION

While language has been leveraged by content-based RS for over three decades [107], the advent of pretrained LLMs and their emergent abilities for generalized, multi-task natural language (NL) reasoning [17, 18, 165] has ushered in a new stage of language-based recommendation. Critically, NL constitutes a unified, expressive, and interpretable medium that can represent not only item features or user preferences, but also user-system interactions, recommendation task descriptions, and external knowledge [45]. For instance, items are often associated with rich text including titles, descriptions, semi-structured textual metadata, and reviews. Similarly, user preferences can be articulated in NL in many forms, such as reviews, search queries, liked item descriptions, and dialogue utterances.

Pretrained LLMs provide new ways to exploit this textual data: recent research (e.g., [40, 45, 58, 138, 143]) has shown that in many domains, LLMs have learned useful reasoning abilities for making and explaining item recommendations based on user preferences as well as facilitating conversational recommendation dialogues. As discussed below, these pretrained abilities can be further augmented through prompting (e.g., [103, 138, 143]), fine-tuning (e.g., [45, 54, 78, 189]), retrieval (e.g., [27, 40, 65, 83, 154]), and other external tools (e.g., [40, 160, 183].

We next proceed to survey the developments in LLM-based RS's, first discussing encoder-only LLMs for dense retrieval and cross-encoding (Section 3.1) followed by generative NL recommendation and explanation with sequence-to-sequence (seq2seq) LLMs (Section 3.2). We then review the complementary use of RS and LLMs covering RAG (Section 3.3) and LLM-based feature extraction (Section 3.4), before concluding with a review of conversational recommendation methods (Section 3.5).

3.1 Encoder-only LLM Recommendation

3.1.1 Recommendation as Dense Retrieval. A common task is to retrieve the most relevant items given a NL preference statement using item texts, for which dense retrieval has become a key tool. Dense retrievers [39] produce a ranked list of documents given a query by evaluating the similarity (e.g., dot product or cosine similarity) between encoder-only LLM document embeddings and the query embedding. They are highly scalable tools (especially when used with approximate search libraries like FAISS[1]) because documents and queries are encoded separately, allowing for dense

[1] https://github.com/facebookresearch/faiss

vector indexing of documents before querying. To use dense retrieval for recommendation [123], first, a component of each item's text content, such as its title, description, reviews, etc., is treated as a document and a dense item index is constructed. Then, a query is formed by some NL user preference description, for instance: an actual search query, the user's recently liked item titles, or a user utterance in a dialogue.

Several recent works explore recommendation *as* standard dense retrieval with retrievers that are off-the-shelf [54, 123, 185] and fine-tuned [65, 91, 116]. More complex dense retrieval methods include review-based retrieval with contrastive BERT fine-tuning [2] and multi-aspect query decomposition [86], and the use of a second-level encoder to fuse the embedding of a user's recently liked items into a user embedding before scoring [92, 168].

3.1.2 Recommendation via LLM Item-Preference Fusion. Several works approach rating prediction by *jointly* embedding NL item and preference descriptions in LLM cross-encoder architectures with an MLP rating prediction head [126, 169, 176, 188, 190]. Such fusion-in-encoder methods often exhibit strong performance because they allow interaction between user and item representations, but are much more computationally expensive than dense retrieval and thus may be best used for small item sets or as rerankers [116].

3.2 LLM-based Generative Recommendation

In LLM-based generative recommendation, tasks are expressed as token sequences – called *prompts* – which form an input to a seq2seq LLM. The LLM then generates another token sequence to address the task – with example outputs including: a recommended list of item titles/ids [54, 111, 138, 143], a rating [9, 78], or an explanation [45, 50, 93, 94, 118]. These methods rely on the pretraining of LLMs on large text corpora to provide knowledge about a wide range of entities, human preferences, and commonsense reasoning that can be used directly for recommendation or leveraged to improve generalization and reduce domain-specific data requirements for fine-tuning or prompting [18, 165].

3.2.1 Zero- and Few- Shot Generative Recommendation. Several recent publications [78, 103, 138, 143] have evaluated with off-the-shelf LLM generative recommendation, focusing on domains that are prevalent in the LLM pre-training corpus such as movie and book recommendation. Specifically, these methods construct a prompt with a NL description of user preference (often using a sequence of recently liked item titles) and an instruction to recommend the next k item titles [103, 138, 143] or predict a rating [78, 103]. While, overall, untuned LLMs underperform supervised CF methods trained on sufficient data [78, 143], they are competitive in near cold-start settings [138, 143]. Few-shot prompting (or in-context learning), in which a prompt contains examples of input-output pairs, typically outperforms zero-shot prompting [138].

3.2.2 Tuning LLMs for Generative Recommendation. To improve an LLM's generative recommendation performance and add knowledge to its internal parameters, multiple works focus on fine-tuning [9, 45, 54, 78, 111] and prompt-tuning [26, 94, 189] strategies. Recent works fine-tune LLMs on NL input/output examples constructed from user-system interaction history and task descriptions for rating prediction [9, 78] and sequential recommendation [54, 111], or in the case of P5 [45], both preceding tasks plus top-k recommendation, explanation generation, and review summarization. Other recommendation works study prompt tuning approaches [26, 94, 189], which adjust LLM behaviour by tuning a set of continuous (or soft) prompt vectors as an alternative to tuning internal LLM weights.

Generative Explanation. A line of recent work focuses on explanation generation where training explanations are extracted from reviews, since reviews often express reasons why a user decided to interact with an item. Techniques include fine-tuning [45, 94, 161], prompt-tuning [93, 94], chain-of-thought prompting [129], and controllable decoding [50, 118, 119, 174] – where additional predicted parameters such as ratings steer LLM decoding.

3.3 Retrieval Augmented Recommendation

Adding knowledge to an LLM internal memory through tuning can improve performance, but it requires many parameters and re-tuning for every system update. An alternative is retrieval-augmented generation (RAG) [15, 70, 87], which conditions output on information from an external source such as a dense retriever (Section 3.1). RAG methods facilitate online updates, reduce hallucinations, and generally require fewer LLM parameters since knowledge is externalized [15, 70, 112].

RAG has recently begun to be explored for recommendation, with the most common approach being to first use a retriever or RS to construct a candidate item set based on a user query or interaction history, and then prompt an encoder-decoder LLM to rerank the candidate set [27, 65, 154, 166, 175]. For RAG-based explanation generation, Xie et al. [174] generate queries based on interaction history to retrieve item reviews which are used as context to generate an explanation of the recommendation. RAG is also emerging as a key paradigm in conversational recommendation (c.f. Sec 3.5): for example, RAG is used in [40] to retrieve relevant user preference descriptions from a user "memory" module to guide dialogue, and by Kemper et al. [83] to retrieve information from an item's reviews to answer user questions.

3.4 LLM-based Feature Extraction

Conversely to how RS or retrievers are used in RAG to obtain inputs for LLMs (Section 3.3), LLMs can also be used to generate inputs for RS [54, 60, 91, 116, 130, 180]. For instance: LLM2-BERT4Rec [54] initializes BERT4Rec (Section 2.1.1) item embeddings of item texts; Query-SeqRec [60] includes LLM query embeddings as inputs to a transformer-based recommender; and TIGER [130] first uses an LLM to embed item text, then quantizes this embedding into a semantic ID, and finally trains a T5-based RS to generate new IDs given a user's item ID history. Similarly, MINT [116] and GPT4Rec [91] produce inputs for a dense retriever by prompting an LLM to generate a query given a user's interaction history.

3.5 Conversational Recommendation

The recent advances in LLMs have made fully NL system-user dialogues a feasible and novel recommendation interface, bringing in a new stage of conversational recommendation (ConvRec) research. This direction studies the application of LLMs in multi-turn, multi-task, and mixed-initiative NL recommendation conversations

[40, 72], introducing dialogue history as a rich new form of interaction data. Specifically, ConvRec includes the study and integration of diverse conversational elements such as dialogue management, recommendation, explanation, QA, critiquing, and preference elicitation [72, 110]. While some research [58] approaches ConvRec with a monolithic LLM such as GPT4, other works rely on an LLM to facilitate NL dialogue *and* integrate calls to a recommender module which generates item recommendations based on dialogue or interaction history [5, 22, 52, 77, 96, 160, 175]. Further research advances ConvRec system architectures with multiple tool-augmented LLM modules, incorporating components for dialogue management, explanation generation, and retrieval [40, 42, 75, 83, 162, 183].

4 GENERATIVE MULTIMODAL RECOMMENDATION SYSTEMS

In recent years, users have come to expect richer interactions than simple text or image queries. For instance, they might provide a picture of a desired product along with a natural language modification (e.g., a dress like the one in the picture but in red). Additionally, users want to visualize recommendations to see how a product fits their use case, such as how a garment might look on them or how a piece of furniture might look in their room. These interactions require new RS that can discover unique attributes in each modality. In this section, we discuss RS that utilize multiple data modalities. In Sections 4.1-4.2 we discuss motivations and challenges to the design of multimodal RS. In Sections 4.3-4.4 we review contrastive and generative approaches to multimodal RS, respectively.

4.1 Why Multimodal Recommendation?

Retailers often have multimodal information about their customers and products, including product descriptions, images and videos, customer reviews and purchase history. However, existing RS typically process each source independently and then combine the results by fusing unimodal relevance scores.

In practice, there are many use cases in which such a "late fusion" approach may be insufficient to satisfy the customer needs. One such use case is the *cold start problem*: when user behavioral data cannot be used to recommend existing products to new customers, or new products to existing customers, it is useful to gather diverse information about the items so that preference information can be transferred from existing products or customers to new ones.

Another use case occurs when different modalities are needed to understand the user request. For example, to answer the request "best metal and glass black coffee table under $300 for my living room", the system would need to reason about the appearance and shape of the item in context with the appearance and shape of other objects in the customer room, which cannot be achieved by searching with either text or image independently. Other examples of multimodal requests include an image or audio of the desired item together with text modification instructions (e.g., a song like the sound clip provided but in acoustic), or a complementary related product (e.g., a kickstand for the bicycle in the picture).

A third use case for multimodal understanding is in RS with complex outputs, such as virtual try-on features or intelligent multimodal conversational shopping assistants.

4.2 Challenges to Multimodal Recommendation

The development of multimodal RS faces several challenges. First, collecting data to train multimodal systems (e.g., image-text-image triplets) is significantly harder than for unimodal systems. As a result, annotations for some modalities may be incomplete [128].

Second, combining different data modalities to improve recommendation results is not simple. For instance, existing contrastive learning approaches [73, 89, 90, 127] map each data modality to a common latent space in which all modalities are approximately aligned. However, such approaches often capture information that is shared across modalities (e.g., text describing visual attributes), but they overlook complementary aspects that could benefit recommendations (e.g., text describing non visual attributes) [49]. In general we would like the modalities to compensate for one another and result in a more complete joint representation. While fusion-based approaches [89, 90] do learn a joint multimodal representation, ensuring the alignment of information that is shared and leaving some flexibility to capture complementary information across modalities remains a challenge. Third, learning multimodal models requires orders of magnitude more data than learning models for individual data modalities.

Despite these challenges, we believe multimodal generative models will become the standard approach. Indeed, recent literature shows significant advances on the necessary components to achieve effective multimodal generative models for RS, including (1) the use of LLMs and diffusion models to generate synthetic data for labeling purposes [16, 117, 135], (2) high quality unimodal encoders and decoders [56, 85], (3) better techniques for aligning the latent spaces from multiple modalities into a shared one [46, 89, 127], (4) efficient re-parametrizations and training algorithms [71], and (5) techniques to inject structure to the learned latent space to make the problem tractable [144].

4.3 Contrastive Multimodal Recommendation

As discussed before 4.2, learning multimodal generative models is very difficult because we need to not only learn a latent representation for each modality but also ensure that they are aligned. One way to address this challenge is to first learn an alignment between multiple modalities and then learn a generative model on "well-aligned" representations. In this subsection, we discuss two representative contrastive learning approaches: CLIP and ALBEF.

Contrastive Language-Image Pre-training (CLIP) [127] is a popular approach, in which the task is to project images and associated text into the same point of the embedding space with parallel image and text encoders. This is achieved with a symmetric cross-entropy loss over the rows and columns of the cosine similarity matrix between all possible pairs of images and text in a training minibatch.

Align Before you Fuse (ALBEF) [90] augments CLIP with a multimodal encoder that fuses the text and image embeddings, and proposes three objectives to pre-train the model: Image-text contrastive learning (ITC), masked language modeling (MLM), and image-text matching (ITM). The authors also introduce momentum distillation to provide pseudo-labels in order to compensate for the potentially incomplete or wrong text descriptions in the noisy web training data. Using their proposed architecture and training objectives, ALBEF obtains better results than CLIP in several zero-shot and fine-tuned multimodal benchmarks, despite using orders of magnitude less images for pre-training.

Contrastive-based alignment has shown impressive zero-shot classification and retrieval results [8, 61, 120], and has been successfully fine-tuned to a multitude of tasks, such as object detection [48], segmentation [192] or action recognition [69]. The same alignment objective has also been used between other modalities [24, 51, 68], and with multiple modalities at the same time [46].

4.4 Generative Multimodal Recommendation

Despite their advantages, the performance of purely contrastive RS often suffers from data sparsity and uncertainty [163]. Generative models address these issues by imposing suitable structures on their latent spaces. Moreover, generative models allow for more complex recommendations, e.g., those requiring to synthesize an image. In what follows, we discuss thee representative generative approaches: VAEs, diffusion models, and multimodal LLMs.

Multimodal VAEs: While VAEs (see Section 2.1.2) could be applied directly to multimodal data, a better approach that leverages modality specific encoders and decoders trained on large corpus of data is to partition both the input and latent spaces per modality, say image and text. However, this approach reduces the multimodal VAE to two independent VAEs, one per modality. In ContrastVAE [163], both modalities are aligned by adding a contrastive loss between the unimodal latent representations to the ELBO objective. Experiments show that ContrastVAE improves upon purely contrastive models by adequately modeling data uncertainty and sparsity, and being robust to perturbations in the latent space.

Diffusion models, explained in Section 2.4, are state-of-the-art models for image generation. While they can also be used for text generation, e.g., by using a discrete latent space with categorical transition probabilities [6], text encoders based on transformers or other sequence-to-sequence models are preferred in practice. As a consequence, multimodal models for both text and images, such as text-to-image generation models, combine text encoders with diffusion models for images. For instance, DALL-E [131] uses the CLIP embedding space as a starting point to generate novel images, and Stable Diffusion [134] uses a UNet autoencoder separately pretrained on a perceptual loss and a patch-based adversarial objective. Several works have built on and expanded diffusion models by increasing controllability of the generated results [187], consistency on the generated subjects identity [136], or for virtual try on [193].

Multimodal LLMs (MLLM) provide a natural language interface for users to express their queries in multiple modalities, or even see responses in different modalities to help visualize the products. Given the complexity of training large generative models end-to-end, researchers typically assemble systems composed of discriminatively pre-trained encoders and decoders, usually connected by adaptation layers to ensure that unimodal representations are aligned. Another approach that involves little or no training is to allow a "controller" LLM to use external foundation models, or tools, to deal with the multimodal input and output [184]. Then, instruction tuning is an important step to make LLMs useful task solvers. Llava [102] is a multimodal LLM that accepts both text and image inputs, and produces useful textual responses. The authors connect a CLIP encoder with an LLM decoder using a simple linear adaptation layer. In [101] the authors change the connection layer from a linear projection to a two-layer MLP and obtain better results. Although MLLM research is still in its inception, some works already start using them in recommendation applications [81].

5 EVALUATING FOR IMPACT AND HARM

Evaluating RS is a complex and multifaceted task that goes beyond simply measuring a few key metrics of a single model . These systems are composed of one or more recommender models and various other ML and non-ML components, making it highly nontrivial to assess and evaluate the performance of an individual model. Moreover, these systems can have far-reaching impacts on users' experiences, opinions, and actions, which may be difficult to quantify or predict, which adds to the challenge. The introduction of Gen-RecSys further complicates the evaluation process due to the lack of well-established benchmarks and the open-ended nature of their tasks. When evaluating RS, it is crucial to distinguish between two main targets of evaluation: the system's performance and capabilities, and its potential for causing safety issues and societal harm. We review these targets, discuss evaluation metrics, and conclude with open challenges and future research directions.

5.1 Evaluating for Offline Impact

The typical approach to evaluating a model involves understanding its accuracy in an offline setting, followed by live experiments.

5.1.1 *Accuracy Metrics.*

The usual metrics used for discriminative tasks are recall@k, precision@k, NDCG@k, AUC, ROC, RMSE, MAE, etc. Many recent works on generative RS (e.g., [9, 58, 78, 79, 130]) incorporate such metrics for discriminative tasks.

For the generative tasks, we can borrow techniques from NLP. For example, the BLEU score is widely used for machine translation and can be useful for evaluating explanations[45], review generation, and conversational recommendations. The ROUGE score, commonly used for evaluating machine-generated summarization, could be helpful again for explanations or review summarization. Similarly, perplexity is another metric that could be broadly useful, including during the training process to ensure that the model is learning the language modeling component appropriately [108].

5.1.2 *Computational Efficiency.*

Evaluating computational efficiency is crucial for generative recommender models, both for training and inference, owing to their computational burden. This is an upcoming area of research.

5.1.3 *Benchmarks.*

Many existing benchmark datasets popular in discriminative recommender models, such as Movielens [53], Amazon Reviews [57], Yelp Challenge[1], Last.fm [139], and Book-Crossing [194], are still useful in generative recommender models, but only narrowly. Some recent ones, like ReDial [96] and INSPIRED [55], are useful datasets for conversational recommendations. [30, 32, 186] propose benchmarks called cFairLLM and FaiR-LLM, to evaluate consumer fairness in LLMs based on the sensitivity of pretrained LLMs to protected attributes in tailoring recommendations. We note that some benchmarks such as BigBench[10] which are commonly used by the LLM community, have recommendations tasks. It will be specifically useful for the RS community to develop new benchmarks for tasks unlocked by Gen-RecSys models.

5.2 Online and Longitudinal Evaluations

Offline experiments may not capture an accurate picture because of the interdependence of the different models used in the system and other factors. So, A/B experiments help understand the model's performance along several axes in real-world settings. Note that [155] proposes a new paradigm of using simulation using agents to evaluate recommender models. In addition to the short-term impact on engagement/satisfaction, the platform owners will be interested in understanding the *long-term impact*. This can be measured using business metrics such as revenue and engagement (time spent, conversions). Several metrics could be used to capture the impact on users (daily/monthly active users, user sentiment, safety, harm).

5.3 Conversational Evaluation

BLEU and perplexity are useful for conversational evaluation but should be supplemented with task-specific metrics (e.g., recall) or objective-specific metrics (e.g., response diversity [88]). Strong LLMs can act as judges, but human evaluation remains the gold standard. Toolkits like CRSLab [76] simplify building and evaluating conversational models, but lack of labeled data in industrial use cases poses a challenge. Some studies use LLM-powered user simulations to generate data.

5.4 Evaluating for Societal Impact

Previous work has investigated categories of interest for societal impacts of traditional RS [113] and generative models [14, 167] independently. In the context of RS literature, six categories of harms are found to be associated with RS: *content*, *privacy* violations and data misuse, threats to human *autonomy* and well-being, *transparency and accountability*, harmful *social effects* such as filter bubbles, polarisation, manipulability, and *fairness*. In addition, RS based on generative models can present new challenges [14, 167]:

- LLMs use out-of-domain knowledge, introducing different sources of societal bias that are not easily captured by existing evaluation techniques [30, 31, 141].
- The significant computational requirements of LLMs lead to heightened environmental impacts [13, 109].
- The automation of content creation and curation may displace human workers in industries such as journalism [28], creative writing, and content moderation, leading to social and economic disruption [4].
- Recommender systems powered by generative models may be susceptible to manipulation and could have unintended and unexpected consequences for users [20, 82].
- Generative recommendations can expose users to the potential pitfalls of hyper-personalization [41, 133].

5.5 Holistic Evaluations

As mentioned above, thoroughly evaluating RS for offline metrics, online performance, and harm is highly non-trivial. Moreover, different stakeholders (e.g. platform owners and users) [3, 114, 147] may approach evaluation differently. The complexity of Gen-RecSys evaluation presents an opportunity for further research and specialized tools. Drawing inspiration from the HELM benchmark [98], a comprehensive evaluation framework tailored for Gen-RecSys would benefit the community.

6 CONCLUSIONS AND FUTURE DIRECTIONS

While many directions for future work have been highlighted above, the following topics constitute especially important challenges and opportunities for Gen-RecSys:

- **RAG** (cf. Section 3.3), including: data fusion for multiple (potentially subjective) sources such as reviews [177, 178], end-to-end retriever-generator training [15, 70, 87], and systematic studies of generative reranking alternatives [125].
- **Tool-augmented LLMs** for conversational recommendation, focusing on architecture design for LLM-driven control of dialogue, recommender modules, external reasoners, retrievers, and other tools [18, 40, 112, 162], especially methods for *proactive* conversational recommendation.
- **Personalized Content Generation** such as virtual try-on experiences [193], which can allow users to visualize themselves wearing recommended clothing or accessories, improving customer satisfaction and reducing returns.
- **Red-teaming** – in addition to the standard evaluations, real-world generative RS will have to undergo red-teaming (i.e., adversarial attacks) [34, 142, 164] before deployment to stress test the system for prompt injections, robustness, alignment verification, and other factors.

Despite being a short survey, this work has attempted to provide a foundational understanding of the rich landscape of generative models within recommendation systems. It extends the discussion beyond LLMs to a broad spectrum of generative models, exploring their applications across user-item interactions, textual data, and multimodal contexts. It highlights key evaluation challenges, addressing performance, fairness, privacy, and societal impact, thereby establishing a new benchmark for future research in the domain.

REFERENCES

[1] 2014. Yelp Dataset. (2014). https://www.yelp.com/dataset

[2] M. M. Abdollah Pour, P. Farinneya, A. Toroghi, A. Korikov, A. Pesaranghader, T. Sajed, et al. 2023. Self-supervised Contrastive BERT Fine-tuning for Fusion-Based Reviewed-Item Retrieval. In *ECIR*. Springer, 3–17.

[3] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *UMUAI* 30 (2020), 127–158.

[4] A.R. Arguedas and F.M Simon. 2023. Automating democracy: Generative AI, journalism, and the future of democracy. (2023).

[5] D. E. Austin, A. Korikov, A. Toroghi, and S. Sanner. 2024. Bayesian Optimization with LLM-Based Acquisition Functions for Natural Language Preference Elicitation. *arXiv preprint arXiv:2405.00981* (2024).

[6] J. Austin, D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *NeurIPS* 34 (2021), 17981–17993.

[7] J. Bai, C. Zhou, J. Song, X. Qu, W. An, Z. Li, and J. Gao. 2019. Personalized bundle list recommendation. In *WWW*. 60–71.

[8] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *ICCV*. 15338–15347.

[9] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *ACM RecSys*. 1007–1014.

[10] BIG bench authors. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *TMLR* (2023).

[11] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio. 2021. Flow network based generative models for non-iterative diverse candidate generation. *NeurIPS* 34 (2021), 27381–27394.

[12] Y. Bengio, R. Ducharme, and P. Vincent. 2000. A neural probabilistic language model. *NeurIPS* 13 (2000).

[13] A. Berthelot, E. Caron, M. Jay, and L. Lefèvre. 2024. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. In *CIRP LCE 2024 - 31st Conference on Life Cycle Engineering, Turin, Italy*. 1–10.

[14] C. Bird, E. Ungless, and A. Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 396–410.

[15] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, et al. 2022. Improving language models by retrieving from trillions of tokens. In *ICML*. PMLR, 2206–2240.

[16] T. Brooks, A. Holynski, and A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*. 18392–18402.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al. 2020. Language models are few-shot learners. *NEU* 33 (2020), 1877–1901.

[18] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712

[19] L. Cai and WY. Wang. 2018. KBGAN: Adversarial Learning for Knowledge Graph Embeddings. In *NAACL, Volume 1 (Long Papers)*. 1470–1480.

[20] Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. 2022. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*. PMLR, 2686–2708.

[21] D-K. Chae, J-S. Kang, S-W. Kim, and J-T. Lee. 2018. CFGAN: A generic collaborative filtering framework based on generative adversarial networks. In *CIKM*. 137–146.

[22] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. 2019. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391* (2019).

[23] X. Chen, S. Li, H. Li, S. Jiang, Y. Qi, and L. Song. 2019. Generative adversarial user model for reinforcement learning based recommendation system. In *ICML*. PMLR, 1052–1061.

[24] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang. 2020. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *MM*. 3884–3892.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

[26] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084* (2022).

[27] S. Dai, N. Shao, H. Zhao, W. Yu, Z. Si, C. Xu, et al. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. *arXiv preprint arXiv:2305.02182* (2023).

[28] D. De Cremer, N. Bianzino, and B. Falk. 2023. How generative AI could disrupt creative work. *Harvard Business Review* 13 (2023).

[29] R. Deffayet, T. Thonet, J-M. Renders, and M. de Rijke. 2023. Generative slate recommendation with reinforcement learning. In *WSDM*. 580–588.

[30] Y. Deldjoo. 2024. FairEvalLLM. A Comprehensive Framework for Benchmarking Fairness in Large Language Model Recommender Systems. *arXiv preprint arXiv:2405.02219* (2024).

[31] Yashar Deldjoo. 2024. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. *arXiv preprint arXiv:2401.10545* (2024).

[32] Y. Deldjoo and T. Di Noia. 2024. CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System. *preprint arXiv:2403.05668* (2024).

[33] Y. Deldjoo, T. Di Noia, and F. A. Merra. 2021. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[34] B. Deng, W. Wang, F. Feng, Y. Deng, Q. Wang, and X. He. 2023. Attack prompt generation for red teaming and defending large language models. *arXiv preprint arXiv:2310.12505* (2023).

[35] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*.

[36] S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* 12 (2016).

[37] J. Du, Z. Ye, B. Guo, Z. Yu, and L. Yao. 2023. Idnp: Interest dynamics modeling using generative neural processes for sequential recommendation. In *WSDM*. 481–489.

[38] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, J. Tang, and Q. Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).

[39] Y. Fan, X. Xie, Y. Cai, J. Chen, X. Ma, X. Li, et al. 2022. Pre-training methods in information retrieval. *FnT-IR* 16, 3 (2022), 178–317.

[40] L. Friedman, S. Ahuja, D. Allen, T. Tan, H. Sidahmed, C. Long, et al. 2023. Leveraging Large Language Models in Conversational Recommender Systems. *arXiv preprint arXiv:2305.07961* (2023).

[41] I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, et al. 2024. The Ethics of Advanced AI Assistants. *arXiv preprint arXiv:2404.16244* (2024).

[42] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524* (2023).

[43] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. M. A. Eslami. 2018. Conditional neural processes. In *ICML*. PMLR, 1704–1713.

[44] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. Eslami, and Y. W. Teh. 2018. Neural processes. *arXiv preprint arXiv:1807.01622* (2018).

[45] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang. 2022. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *RecSys*. 299–315.

[46] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, KV. Alwala, A. Joulin, and I. Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*. 15180–15190.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

[48] X. Gu, T. Lin, W. Kuo, and Y. Cui. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*. https://api.semanticscholar.org/CorpusID:238744187

[49] W. Guo, J. Wang, and S. Wang. 2019. Deep multimodal representation learning: A survey. *Ieee Access* 7 (2019), 63373–63394.

[50] D V Hada and S.K Shevade. 2021. Rexplug: Explainable recommendation using plug-and-play language model. In *SIGIR*. 81–91.

[51] P. Hager, M. Menten, and D. Rueckert. 2023. Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data. In *CVPR*. 23924–23935.

[52] K. Handa, Y. Gal, E. Pavlick, N. Goodman, J. Andreas, A. Tamkin, and B. Li. 2024. Bayesian preference elicitation with language models. *arXiv:2403.05534* (2024).

[53] F Maxwell Harper and Joseph A Konstan. 2015. The Movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* (2015).

[54] J. Harte, W. Zorgdrager, P. Louridas, A. Katsifodimos, D. Jannach, and M. Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *ACM RecSys*. 1096–1102.

[55] S.A. Hayati, D. Kang, Q. Zhu, W. Shi, and Z. Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *EMNLP*. ACL, Online, 8142–8152.

[56] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[57] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *International World Wide Web Conference*.

[58] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, and J. McAuley. 2023. Large language models as zero-shot conversational recommenders. *arXiv preprint arXiv:2308.10053* (2023).

[59] Z. He, H. Zhao, Z. Lin, Z. Wang, A. Kale, and J. McAuley. 2021. Locker: Locally constrained self-attentive sequential recommendation. In *CIKM*. 3088–3092.

[60] Z. He, H. Zhao, Z. Wang, Z. Lin, A. Kale, and J. Mcauley. 2022. Query-Aware Sequential Recommendation. In *CIKM*. 4019–4023.

[61] M. Hendriksen, M. Bleeker, S. Vakulenko, N. van Noord, E. Kuiper, and M. de Rijke. 2022. Extending CLIP for Category-to-image Retrieval in E-commerce. In *ECIR*. Springer, 289–303.

[62] B. Hidasi and A. Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *CIKM*. 843–852.

[63] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. 2016. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*.

[64] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* (1997).

[65] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, et al. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845* (2023).

[66] H. Hu and X. He. 2019. Sets2sets: Learning from sequential sets with neural networks. In *ACM SIGKDD*. 1491–1499.

[67] C. Huang, T. Yu, K. Xie, S. Zhang, L. Yao, and J. McAuley. 2024. Foundation Models for Recommender Systems: A Survey and New Perspectives. *arXiv preprint arXiv:2402.11143* (2024).

[68] W. Huang. 2023. Multimodal Contrastive Learning and Tabular Attention for Automated Alzheimer's Disease Prediction. In *ICCV*. 2473–2482.

[69] X. Huang, H. Zhou, K. Yao, and K. Han. 2024. FROSTER: Frozen CLIP is A Strong Teacher for Open-Vocabulary Action Recognition. In *ICLR*.

[70] G. Izacard and E. Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *preprint arXiv:2007.01282* (2020).

[71] E. Jang, S. Gu, and B. Poole. 2016. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*.

[72] D. Jannach, A. Manzoor, W. Cai, and L. Chen. 2020. A Survey on Conversational Recommender Systems. *arXiv preprint arXiv:2004.00646* (2020).

[73] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. Le, Y. Sung, Z. Li, and T. Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*. PMLR, 4904–4916.

[74] R. Jiang, S. Gowal, Y. Qian, T. Mann, and D. J. Rezende. 2018. Beyond Greedy Ranking: Slate Optimization via List-CVAE. In *ICLR*.

[75] H. Joko, Sh. Chatterjee, A. Ramsay, A. P de Vries, J. Dalton, and F. Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. (2024).

[76] Y. Zhou C. Shang Y. Cheng WX. Zhao Y. Li J-R. Wen K. Zhou, X. Wang. 2021. CRSLab: An Open-Source Toolkit for Building Conversational Recommender System. *arXiv preprint arXiv:2101.00939* (2021).

[77] D. Kang, A. Balakrishnan, P. Shah, P. Crook, Y-L. Boureau, and J. Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. In *Empirical Methods in Natural Language Processing*.

[78] W. Kang, J. Ni, N. Mehta, M. Sathiamoorthy, L. Hong, E. Chi, and D. Cheng. 2023. Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction. *arXiv preprint arXiv:2305.06474* (2023).

[79] W-C. Kang, C. Fang, Z. Wang, and J. McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *ICDM*. IEEE.

[80] W.-C. Kang and J. McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*. IEEE, 197–206.

[81] SR. Karra and T. Tulabandhula. 2024. InteraRec: Interactive Recommendations Using Multimodal Large Language Models. *preprint arXiv:2403.00822* (2024).

[82] A. Kasirzadeh and C. Evans. 2023. User tampering in reinforcement learning recommender systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 58–69.

[83] S. Kemper, J. Cui, K. Dicarlantonio, K. Lin, D. Tang, A. Korikov, and S. Sanner. 2024. Retrieval-Augmented Conversational Recommendation with Prompt-based Semi-Structured Natural Language State Tracking. In *SIGIR*. ACM.

[84] D. Kingma and M. Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[85] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W. Lo, et al. 2023. Segment anything. In *CVPR*. 4015–4026.

[86] A. Korikov, G. Saad, E. Baron, M. Khan, M. Shah, and S. Sanner. 2024. Multi-Aspect Reviewed-Item Retrieval via LLM Query Decomposition and Aspect Fusion. In *SIGIR Workshop on Information Retrieval's Role in RAG Systems*.

[87] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS* 33 (2020), 9459–9474.

[88] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv:1510.03055* (2015).

[89] J. Li, D. Li, C. Xiong, and S. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 12888–12900.

[90] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and SCH. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS* 34 (2021), 9694–9705.

[91] J. Li, W. Zhang, T. Wang, G. Xiong, A. Lu, and G. Medioni. 2023. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation. *arXiv preprint arXiv:2304.03879* (2023).

[92] J. Li, J. Zhu, Q. Bi, G. Cai, L. Shang, Z. Dong, et al. 2022. MINER: Multi-interest matching network for news recommendation. In *ACL*. 343–352.

[93] L. Li, Y. Zhang, and L. Chen. 2020. Generate neural template explanations for recommendation. In *CIKM*. 755–764.

[94] L. Li, Y. Zhang, and L. Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM TOIS* 41, 4 (2023), 1–26.

[95] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2023. Large language models for generative recommendation: A survey and visionary discussions. *arXiv preprint arXiv:2309.01157* (2023).

[96] R. Li, S.E. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

[97] D. Liang, R. G Krishnan, M. D Hoffman, and T. Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[98] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[99] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, X. Li, C. Zhu, H. Guo, Y. Yu, R. Tang, et al. 2023. How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817* (2023).

[100] J. Lin, W. Pan, and Z. Ming. 2020. FISSA: Fusing item similarity models with self-attention networks for sequential recommendation. In *RecSys*. 130–139.

[101] H. Liu, C. Li, Y. Li, and YJ Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).

[102] H. Liu, C. Li, Q. Wu, and YJ Lee. 2024. Visual instruction tuning. *NeurIPS* 36 (2024).

[103] J. Liu, C. Liu, R. Lv, K. Zhou, and Y. Zhang. 2023. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149* (2023).

[104] Q. Liu, F. Yan, X. Zhao, Z. Du, H. Guo, R. Tang, and F. Tian. 2023. Diffusion augmentation for sequential recommendation. In *CIKM*. 1576–1586.

[105] S. Liu, Q. Cai, Z. He, B. Sun, J. McAuley, D. Zheng, P. Jiang, and K. Gai. 2023. Generative flow network for listwise recommendation. In *SIGKDD*. 1524–1534.

[106] S. Liu, F. Sun, Y. Ge, C. Pei, and Y. Zhang. 2021. Variation control and evaluation for generative slate recommendations. In *Web Conf.* 436–448.

[107] P. Lops, M. De Gemmis, and G. Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.

[108] Y. Lu, J. Bao, Y. Song, Z. Ma, S. Cui, Y. Wu, and X. He. 2021. RevCore: Review-augmented conversational recommendation. *arXiv preprint arXiv:2106.00957* (2021).

[109] A. S. Luccioni, Y. Jernite, and E. Strubell. 2023. Power hungry processing: Watts driving the cost of ai deployment? *arXiv preprint arXiv:2311.16863* (2023).

[110] S. Lyu, A. Rana, S. Sanner, and M. R. Bouadjenek. 2021. A Workflow Analysis of Context-driven Conversational Recommendation. In *ACM WWW*.

[111] Z. Mao, H. Wang, Y. Du, and K-F. Wong. 2023. UniTRec: A Unified Text-to-Text Transformer and Joint Contrastive Learning Framework for Text-based Recommendation. In *ACL*.

[112] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, et al. 2023. Augmented Language Models: a Survey. arXiv:2302.07842 [cs.CL]

[113] S. Milano, M. Taddeo, and L. Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35 (2020), 957–967.

[114] S. Milano, M. Taddeo, and L. Floridi. 2021. Ethical aspects of multi-stakeholder recommendation systems. *The Information Society* 37 (2021), 35–45.

[115] M. Mirza and S. Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[116] S. Mysore, A. McCallum, and H. Zamani. 2023. Large language model augmented narrative driven recommendations. In *ACM RecSys*. 777–783.

[117] Q. Nguyen, T. Vu, A. Tran, and K. Nguyen. 2024. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *NeurIPS* 36 (2024).

[118] J. Ni, J. Li, and J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *IJCNLP*.

[119] J. Ni and J. McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Annual Meeting of the Association for Computational Linguistics*.

[120] Z. Novack, J. McAuley, ZC. Lipton, and S. Garg. 2023. Chils: Zero-shot image classification with hierarchical label sets. In *ICML*. PMLR, 26342–26362.

[121] TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI* (2022).

[122] L. Pan, N. Malkin, D. Zhang, and Y. Bengio. 2023. Better training of gflownets with local credit and incomplete trajectories. In *ICML*. PMLR, 26878–26890.

[123] G. Penha and C. Hauff. 2020. What does BERT know about books, movies and music? probing bert for conversational recommendation. In *ACM RecSys*. 388–397.

[124] Aleksandr V Petrov and Craig Macdonald. 2023. Generative sequential recommendation with gptrec. *arXiv preprint arXiv:2306.11114* (2023).

[125] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563* (2023).

[126] Z. Qiu, X. Wu, J. Gao, and W. Fan. 2021. U-BERT: Pre-training user representations for improved recommendation. In *AAAI*, Vol. 35. 4320–4327.

[127] A. Radford, JW. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.

[128] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion* 81 (2022), 203–239.

[129] B. Rahdari, H. Ding, Z. Fan, Y. Ma, Z. Chen, and A. others Deoras. 2024. Logic-scaffolding: Personalized aspect-instructed recommendation explanation generation using llms. In *WSDM*. 1078–1081.

[130] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2024. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2024).

[131] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[132] D. Rezende and S. Mohamed. 2015. Variational inference with normalizing flows. In *ICML*. PMLR, 1530–1538.

[133] Matthias C Rillig and Atoosa Kasirzadeh. 2024. AI Personal Assistants and Sustainability: Risks and Opportunities. *Environmental Science & Technology* 58, 17 (2024), 7237–7239.

[134] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.

[135] A. Rosenbaum, S. Soltan, W. Hamza, A. Saffari, M. Damonte, and I. Groves. 2022. CLASP: Few-Shot Cross-Lingual Data Augmentation for Semantic Parsing. *AACL-IJCNLP 2022* (2022), 444.

[136] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*. 22500–22510.

[137] N. Sachdeva, G. Manco, E. Ritacco, and V. Pudi. 2019. Sequential variational autoencoders for collaborative filtering. In *WSDM*. 600–608.

[138] S. Sanner, K. Balog, F. Radlinski, B. Wedin, and L. Dixon. 2023. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *RecSys*. 890–896.

[139] M. Schedl. 2016. The lfm-1b dataset for music retrieval and recommendation. In *ICMR*. 103–110.

[140] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web*. 111–112.

[141] T. Shen, J. Li, M. Bouadjenek, Z. Mai, and S. Sanner. 2023. Towards Understanding and Mitigating Unintended Biases in Language Model-driven Conversational Recommendation. *Information Processing & Management* 60, 1 (2023), 103139.

[142] D. Shu, M. Jin, S. Zhu, B. Wang, Z. Zhou, C. Zhang, and Y. Zhang. 2024. AttackEval: How to Evaluate the Effectiveness of Jailbreak Attacking on Large Language Models. arXiv:2401.09002 [cs.CL]

[143] D. Sileo, W. Vossen, and R. Raymaekers. 2022. Zero-shot recommendation as language modeling. In *ECIR*. Springer, 223–230.

[144] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. PMLR, 2256–2265.

[145] K. Sohn, H. Lee, and X. Yan. 2015. Learning structured output representation using deep conditional generative models. *NeurIPS* 28 (2015).

[146] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*. 1441–1450.

[147] Ö. Sürer, R. Burke, and E. C Malthouse. 2018. Multistakeholder recommendation with provider constraints. In *ACM RecSys*. 54–62.

[148] Gemini Team, R. Anil, S. Borgeaud, Y. Wu, J-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. Dai, A. Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[149] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[150] A. Vats, V. Jain, R. Raja, and A. Chadha. 2024. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. *arXiv preprint arXiv:2402.18590* (2024).

[151] P. Vincent, H. Larochelle, Y. Bengio, and P-A. Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*. 1096–1103.

[152] J. Walker, T. Zhong, F. Zhang, Q. Gao, and F. Zhou. 2022. Recommendation via collaborative diffusion generative model. In *KSEM*. Springer, 593–605.

[153] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, and D. Zhang. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *SIGIR*. 515–524.

[154] L. Wang and E. Lim. 2023. Zero-Shot Next-Item Recommendation using Large Pretrained Language Models. *arXiv preprint arXiv:2304.03153* (2023).

[155] L. Wang, J. Zhang, X. Chen, Y. Lin, R. Song, W. Zhao, and J. Wen. 2023. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552* (2023).

[156] Q. Wang, H. Yin, Z. Hu, D. Lian, H. Wang, and Z. Huang. 2018. Neural memory streaming recommender networks with adversarial training. In *ACM SIGKDD*. 2467–2475.

[157] Q. Wang, H. Yin, H. Wang, Q. V. H. Nguyen, Z. Huang, and L. Cui. 2019. Enhancing collaborative filtering with generative augmentation. In *SIGKDD*. 548–556.

[158] W. Wang, X. Lin, F. Feng, X. He, and T-S. Chua. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516* (2023).

[159] Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. 2023. Diffusion Recommender Model. *arXiv preprint arXiv:2304.04971* (2023).

[160] X. Wang, K. Zhou, J. Wen, and W. X. Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *ACM SIGKDD*. 1929–1937.

[161] Y. Wang, Z. He, Z. He, H. Xu, and J. McAuley. 2024. Deciphering Compatibility Relationships with Textual Descriptions via Extraction and Explanation. In *AAAI*, Vol. 38. 9133–9141.

[162] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, et al. 2023. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296* (2023).

[163] Y. Wang, H. Zhang, Z. Liu, L. Yang, and P. Yu. 2022. Contrastvae: Contrastive variational autoencoder for sequential recommendation. In *CIKM*. 2056–2066.

[164] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966* (2023).

[165] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[166] W. Wei, X. Ren, J. Tang, Q. Wang, L. Su, S. Cheng, et al. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM WSDM*. 806–815.

[167] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[168] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, and X. Xie. 2019. Neural news recommendation with multi-head self-attention. In *EMNLP*. 6389–6394.

[169] C. Wu, F. Wu, T. Qi, and Y. Huang. 2021. Empowering news recommendation with pre-trained language models. In *SIGIR*. 1652–1656.

[170] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. 2020. SSE-PT: Sequential recommendation via personalized transformer. In *RecSys*. 328–337.

[171] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, et al. 2023. A survey on large language models for recommendation. *preprint arXiv:2305.19860* (2023).

[172] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *WSDM*. 153–162.

[173] Z. Wu, X. Wang, H. Chen, K. Li, Y. Han, L. Sun, and W. Zhu. 2023. Diff4Rec: Sequential Recommendation with Curriculum-scheduled Diffusion Augmentation. In *ACM MM*. 9329–9335.

[174] Z. Xie, S. Singh, J. McAuley, and B. Majumder. 2023. Factual and informative review generation for explainable recommendation. In *AAAI*, Vol. 37. 13816–13824.

[175] B. Yang, C. Han, Y. Li, L. Zuo, and Z. Yu. 2022. Improving Conversational Recommendation Systems' Quality with Context-Aware Item Meta-Information. In *NAACL*. 38–48.

[176] S. Yao, J. Tan, X. Chen, J. Zhang, X. Zeng, and K. Yang. 2022. ReprBERT: distilling BERT to an efficient representation-based relevance model for e-commerce. In *ACM SIGKDD*. 4363–4371.

[177] Q. Ye, I. Beltagy, M. Peters, X. Ren, and H. Hajishirzi. 2023. FiD-ICL: A Fusion-in-Decoder Approach for Efficient In-Context Learning. In *ACL*. 8158–8185.

[178] D. Yu, C. Zhu, Y. Fang, W. Yu, S. Wang, Y. Xu, X. Ren, Y. Yang, and M. Zeng. 2021. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *arXiv preprint arXiv:2110.04330* (2021).

[179] L. Yu, L. Sun, B. Du, C. Liu, H. Xiong, and W. Lv. 2020. Predicting temporal sets with deep neural networks. In *ACM SIGKDD*. 1083–1091.

[180] Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni. 2023. Where to go next for recommender systems? ID-vs. modality-based recommender models revisited. In *SIGIR*. 2639–2649.

[181] Z. Yue, Z. He, H. Zeng, and J. McAuley. 2021. Black-box attacks on sequential recommenders via data-free model extraction. In *RecSys*. 44–54.

[182] Z. Yue, Y. Wang, Z. He, H. Zeng, J. McAuley, and D. Wang. 2024. Linear recurrent units for sequential recommendation. In *WSDM*. 930–938.

[183] Y. Zeng, A. Rajasekharan, P. Padalkar, K. Basu, J. Arias, and G. Gupta. 2024. Automated interactive domain-specific conversational agents that understand human dialogs. In *PADL*. Springer, 204–222.

[184] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *preprint arXiv:2401.13601* (2024).

[185] H. Zhang, A. Korikov, P. Farinneya, M. M. Abdollah Pour, M. Bharadwaj, A. Pesaranghader, et al. 2023. Recipe-MPR: A Test Collection for Evaluating Multi-aspect Preference-based Natural Language Retrieval. In *Proceedings of the 46th ACM SIGIR*. 2744–2753.

[186] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *RecSys*. 993–999.

[187] L. Zhang, A. Rao, and M. Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*. 3836–3847.

[188] Q. Zhang, J. Li, Q. Jia, C. Wang, J. Zhu, Zh. Wang, et al. 2021. UNBERT: User-News Matching BERT for News Recommendation.. In *IJCAI*, Vol. 21. 3356–3362.

[189] Y. Zhang, F. Feng, J. Zhang, K. Bao, Q. Wang, and X. He. 2023. COLLM: Integrating collaborative embeddings into large language models for recommendation. *arXiv preprint arXiv:2310.19488* (2023).

[190] Z. Zhang and B. Wang. 2023. Prompt learning for news recommendation. In *SIGIR*. 227–237.

[191] T. Zhong, Z. Wen, F. Zhou, G. Trajcevski, and K. Zhang. 2020. Session-based recommendation via flow-based deep generative networks and Bayesian inference. *Neurocomputing* 391 (2020), 129–141.

[192] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *CVPR*. 11175–11185.

[193] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, and I. Kemelmacher-Shlizerman. 2023. Tryondiffusion: A tale of two unets. In *CVPR*. 4606–4615.

[194] C. Ziegler, S.M McNee, J.A Konstan, and G. Lausen. 2005. Improving recommendation lists through topic diversification. In *International World Wide Web Conference*.