From Retrieval to Generation: A Survey of Large-Language-Model-Driven Personalized Content Recommendation and Distribution Technologies

HAN Yuxuan

2026-07-23

ChinaRxiv

Abstract

[Purpose] This paper systematically reviews the paradigm shift in personalized recommendation systems driven by large language models from retrieval-based distribution to generative distribution, traces the trajectory of technological evolution, and proposes a unified analytical framework. [Methods] Based on the GeneRec paradigm and the Gen-RecSys classification framework, this study systematically retrieves and analyzes more than 20 cutting-edge domestic and international studies in the field of generative recommendation, summarizes the four roles of LLM in recommendation, proposes a three-stage technological evolution framework of "augmentation-generation-collaboration" , and conducts a classified analysis and comparison of key technologies. [Results] This paper outlines the technological evolution path from collaborative filtering to LLM-driven recommendation; proposes a three-stage framework covering key technologies such as Prompt representation learning, semantic ID generative retrieval, multimodal sentiment awareness, and multi-agent collaborative filtering; and analyzes industrial applications and governance challenges in light of platform practices such as WeChat Channels, Xiaohongshu, and Douyin. [Limitations] The field of generative recommendation is developing extremely rapidly, and some of the latest industrial developments (2026) may not be fully covered; the three-stage framework is mainly based on the induction of existing technologies, and its predictive capability for future technologies remains to be verified. [Conclusions] Recommendation systems are accelerating their transformation from retrieval-based distribution to generative distribution. The four LLM-driven roles constitute a progressive technical spectrum of capabilities, and the three-stage framework provides a systematic reference perspective for subsequent research.

Full Text

From Retrieval to Generation: A Survey of Large-Language-Model-Driven Personalized Content Recommendation and Distribution Technologies

HAN Yuxuan*

School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu 213159

Corresponding author: HAN Yuxuan*, E-mail: 2300770102@smail.cczu.edu.cn

Abstract:

[Objective] This paper systematically surveys the paradigm shift in personalized recommendation systems driven by large language models, from retrieval-based distribution to generative distribution; it traces the trajectory of techno-

logical evolution and proposes a unified analytical framework.

[**Methods**] Based on the GeneRec paradigm and the Gen-RecSys classification framework, this study systematically retrieves and analyzes more than 20 cutting-edge domestic and international papers in the field of generative recommendation, summarizes the four roles of LLMs in recommendation, proposes a three-stage technological evolution framework of "augmentation—generation—collaboration," and classifies, analyzes, and compares key technologies.

[**Results**] The technological evolution path from collaborative filtering to LLM-driven recommendation is clarified; a three-stage framework is proposed, covering key technologies such as Prompt representation learning, semantic-ID-based generative retrieval, multimodal affective perception, and multi-agent collaborative filtering; and, in combination with platform practices such as WeChat Channels, Xiaohongshu, and Douyin, industrial applications and governance challenges are analyzed.

[**Limitations**] The field of generative recommendation is developing extremely rapidly, and some of the latest industry developments (2026) may not be fully covered; the three-stage framework is mainly an induction based on existing technologies, and its predictive capability for future technologies remains to be verified.

[**Conclusion**] Recommendation systems are accelerating their transformation from retrieval-based distribution to generative distribution. The four LLM-driven roles constitute a capability-progressive technological spectrum, and the three-stage framework provides a systematic reference perspective for subsequent research.

Keywords: generative recommendation; large language model; content distribution; AIGC; multimodal recommendation; personalized recommendation; privacy protection

Classification number: TP311

Survey on Generative AI-Driven Personalized Content

Recommendation and Distribution Technologies

HAN Yuxuan*

School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213159, China

Corresponding author*: HAN Yuxuan, E-mail: 2300770102@smail.cczu.edu.cn

Abstract:

[Objective] To systematically survey the paradigm shift from retrieval-based to generative distribution in personalized content recommendation driven by large language models, and to propose a unified analytical framework for this rapidly evolving field.

[Methods] Grounded in the GeneRec paradigm and Gen-RecSys taxonomy, we systematically review over 20 recent publications in generative recommendation. We identify four roles of LLMs in recommender systems, propose a three-stage evolutionary framework— "Enhancement, Generation, and Collaboration" —and provide a comparative analysis of key techniques including prompt-based representation learning, semantic ID-based generative retrieval, multimodal emotion-aware recommendation, and multi-agent collaborative filtering. Industrial practices from WeChat Channels, Xiaohongshu, and TikTok are integrated with governance dimensions.

[Results] The technological evolution from collaborative filtering to LLM-driven recommendation is traced. The three-stage framework covers prompt-based representation learning, semantic ID-based generative retrieval, multimodal emotion perception, and multi-agent collaborative filtering. The analysis reveals four progressive roles of LLMs that collectively drive the paradigm transition.

[Limitations] The field evolves rapidly; some recent industrial developments may not be fully captured. The framework is derived from existing technology categorization; its predictive validity for future developments awaits verification.

[Conclusions] Recommendation systems are accelerating their shift from retrieval-based to generative distribution. The four LLM roles constitute a capability-progressive spectrum, and the three-stage framework provides a systematic reference for subsequent research.

Keywords: Generative recommendation; Large language model; Content distribution; AIGC; Multimodal recommendation; Personalized recommendation; Privacy protection

1 Introduction

As digital technologies deeply permeate social life, the speed of content production and dissemination is growing exponentially, and the problem of information overload is becoming increasingly severe. As the core bridge connecting users and content, recommender systems have become an indispensable infrastructure for Internet platforms. Traditional recommender systems mainly follow the technical routes of collaborative filtering and content-based filtering; their core logic is to retrieve, from an existing content repository, the content that best matches user preferences and distribute it. However, as users' requirements for personalized experience continue to rise, the limitations of retrieval-based recommendation have become increasingly evident: content supply is constrained by the scale of the repository; users can express preferences only through passive feedback; recommendation results are presented in a "black-box" form and

lack interpretability. Since the end of 2022, the explosive rise of large language models represented by ChatGPT has brought entirely new possibilities to recommender systems—by virtue of their capabilities in semantic understanding, world knowledge, and text generation, large language models can play roles in recommender systems that were previously unavailable. Entering 2025–2026, the industrial trend toward deep integration of generative artificial intelligence and content distribution has become more pronounced: the WeChat Open Platform released an AI ecosystem development guide, Xiaohongshu launched the RED Skill function, and the European Commission issued an algorithmic recommendation control order against Meta. These developments all point to the same trend: recommender systems are accelerating their transformation from a retrieval-based distribution paradigm toward a generative distribution paradigm centered on large language models.

Although research in this field is advancing rapidly, it still presents fragmented characteristics. Different studies enter from separate perspectives such as feature encoding, score prediction, content generation, and conversational interaction, lacking a unified technical classification framework; the boundary between traditional recommendation and generative recommendation remains insufficiently clear; the latest industrial practices such as WeChat Channels and Xiaohongshu RED Skill have rarely been addressed in academic reviews; and governance dimensions such as privacy protection, algorithmic fairness, and user controllability also lack an integrated analysis linked to technological evolution.

In response to these issues, this paper provides a systematic review of generative-AI-driven content recommendation and personalized distribution technologies. The main contributions of this paper include three aspects. First, it systematically traces the technological evolution of recommender systems from collaborative filtering and deep-learning recommendation to LLM-enhanced recommendation and then to generative recommendation. Taking the GeneRec paradigm and the Gen-RecSys classification framework as the theoretical basis, it clarifies the four role positions of large language models in recommender systems and specifies the transition path from "LLM-enhanced recommendation" to "generative recommendation." Second, it proposes a three-stage technological evolution framework of "enhancement—generation—collaboration," and conducts categorical analysis and comparative analysis of key technologies such as Prompt representation learning, semantic-ID generative retrieval, multimodal affect-aware recommendation, and multi-agent collaborative filtering. Third, it incorporates into the review the industrial practices of mainstream platforms such as WeChat Channels, Xiaohongshu RED Skill, and Douyin, and at the same time discusses governance challenges from the dimensions of privacy protection, algorithmic fairness, and user controllability, thereby achieving a unified analysis across the three levels of technological evolution, industrial application, and governance challenges.

This paper is organized into eight sections. Section 2 reviews the technological evolution and three core dilemmas of traditional recommender systems. Section

3, based on the GeneRec paradigm and the Gen-RecSys classification framework, analyzes the paradigm shift in generative-AI-driven recommender systems. Section 4 proposes the three-stage technological evolution framework of "enhancement—generation—collaboration" and classifies and analyzes key technical methods. Section 5 combines cases from mainstream platforms to explore industrial application scenarios. Section 6 analyzes governance challenges from the three dimensions of privacy protection, algorithmic fairness, and user controllability. Section 7 looks ahead to future development trends. Section 8 concludes the paper.

2 Review of Traditional Recommender Systems and Content Distribution Technologies

The core task of traditional recommender systems can be formalized as estimating the association function of user-item interactions, $f : U \times I \to \mathbb{R}$. Their technological evolution has gone through four stages, from heuristic similarity computation to deep representation learning. In this process, user-item interaction data are usually represented as rating matrices, item sets, interaction sequences, or bipartite graphs, and different forms of representation have given rise to different modeling methods. From memory-based similarity computation to model-based matrix factorization, and then to deep neural networks, the traditional recommendation paradigm has gradually improved prediction accuracy and modeling capability, but has never fully broken away from

the essence of the discriminative architecture.

2.1 Technical Evolution of Classical Recommendation Paradigms

The core task of a traditional recommender system can be formalized as estimating the association function of user-item interactions, $f : U \times I \to \mathbb{R}$. Its technical evolution has gone through four stages, from heuristic similarity computation to deep representation learning.

Figure content:
1990s–2000s: User-based CF; SVD / MF; TF-IDF.
Collaborative filtering / content filtering → deep-learning recommendation → LLM-enhanced recommendation → generative recommendation.
2010s.
2022–2024: LLM as Encoder; Prompt-based; RAG / CRS.
2024–present: GeneRec; Gen-RecSys; Multimodal Gen.

Figure 1 Timeline of the technical evolution of recommender systems

(1) Collaborative Filtering

Collaborative Filtering (CF) is the most classical paradigm in recommender systems, and can be divided into memory-based and model-based approaches.

Memory-based methods directly use the user–item interaction matrix to compute similarity, such as Jaccard similarity, cosine similarity, and Pearson correlation coefficient; model-based methods approximate the high-dimensional interaction matrix as a low-rank product through matrix factorization (MF), such as SVD and its variants. The latent factor model proposed by Koren et al. further introduced user bias and item bias terms to improve prediction accuracy. However, CF methods are highly dependent on the density of historical interaction data and have limited performance on sparse matrices and cold-start users.

(2) Content-Based Filtering and Hybrid Recommendation

Content-Based Filtering (CBF) performs matching by extracting item attribute features, such as TF-IDF, tags, and categories, thereby alleviating CF's inability to handle new items. However, CBF is prone to over-specialization, making it difficult to discover users' potential new interests. To combine the advantages of both, hybrid recommender systems integrate collaborative signals and content signals through weighted fusion, switching strategies, or feature combinations, such as FM and FFM. Although hybrid strategies improve recommendation quality to a certain extent, their fusion rules mostly rely on manual design, and in essence they still do not depart from the discriminative scoring framework, making it difficult to handle high-dimensional, unstructured multimodal content.

(3) Deep Neural Network Recommendation

With the development of deep learning, recommender systems entered the DNN era. Models such as Wide & Deep, DeepFM, and DIN replace inner-product operations with multilayer perceptrons to learn more complex nonlinear interactions between users and items. Wide & Deep jointly trains a shallow memorization component (the wide part, based on logical regression over crossed features) and a deep generalization component (the deep part, an embedding-based DNN), balancing recommendation relevance and diversity; DeepFM conducts end-to-end joint learning of FM' s low-order interactions and DNN' s high-order interactions, avoiding the overhead of feature engineering in Wide & Deep; DIN introduces an attention mechanism to dynamically capture the relevance between users' historical behaviors and candidate items, and computes a weighted representation of user interests adaptively through locally activated units. In sequential recommendation, early work such as FPMC (Factorized Personalized Markov Chain) combined Markov chains with matrix factorization, assuming that the next item depends only on the previous item, and modeled short-range dependencies in user sequences through a factorized transition matrix; Hidasi et al. introduced RNNs into session modeling, capturing the temporal evolution of user behavior through gated recurrent units (GRU) and encoding session sequences into hidden-state vectors; NARM further combined an attention mechanism to focus on recent interactions and distinguish users' primary intentions from incidental behaviors. Subsequent studies used self-attention mechanisms (SASRec) and bidirectional Transformers (BERT4Rec) to capture long-range dependencies, among which SASRec implements parallelization through self-

attention layers.

training, thereby abandoning the sequence-dependence limitations of RNNs; BERT4Rec uses a bidirectional encoder to model bidirectional context in users' histories and conducts pretraining through a masked-item prediction task. In addition, item2vec transfers word2vec from natural language processing to item sequences, learning distributed item representations by maximizing item co-occurrence probability; "user-free" models such as FISM avoid storing massive user vectors by aggregating representations of items in a user' s history, dynamically computing the user representation at inference time, which significantly reduces storage overhead and improves the scalability of industrial deployment. However, these DNN models still estimate the conditional probability $P(Y|X)$ in a discriminative manner, use fixed low-dimensional vectors to represent user–item interactions, fail to fully model the generative mechanism of the data distribution itself, and also find it difficult to directly exploit the high-dimensional semantic information in multimodal content such as text and images.

2.2 Three Major Dilemmas of the Traditional Paradigm

Looking across the above evolutionary path, traditional recommender systems are essentially discriminative, ID-centric architectures: they take user IDs and item IDs as anchors and implement recommendation by optimizing a scoring function or ranking loss, with the core objective of estimating the conditional probability $P(Y|X)$ of user–item interaction. This paradigm has made substantial progress in accuracy and scalability, but its inherent limitations have become increasingly evident and can be summarized as three core dilemmas: the content bottleneck, the interaction bottleneck, and the interpretability bottleneck.

(1) Content bottleneck

Traditional systems mainly rely on discrete IDs and shallow features, and have weak semantic-understanding ability for multimodal content such as textual descriptions, images, and videos. MF and DNN models use fixed low-dimensional vectors to represent interactions, making it difficult to capture complex semantic associations between user preferences and item attributes. For example, traditional models cannot directly understand natural-language intentions such as "a sci-fi film similar to *Inception* but with a more open ending," resulting in homogenized recommendation results and a lack of cross-modal reasoning ability. Li et al. [11] point out that this "content semantic gap" is precisely the core problem that generative models attempt to bridge through probabilistic distribution modeling. Traditional discriminative architectures compress items into static ID embeddings, severing the deep semantic connection between item content and user preferences and limiting the generalization capability of recommender systems in open-domain knowledge environments. In addition, traditional models find it difficult to utilize natural-language information such as item text descriptions and user reviews, and cannot establish cross-modal semantic alignment; as a result, recommendations lack an understanding of the deep semantics of

items and struggle to satisfy users' increasingly complex personalization needs.

(2) Interaction bottleneck

The user–item interaction matrix is extremely sparse, and CF-type methods suffer a sharp performance decline in long-tail scenarios and under cold-start conditions. Although DNN sequential models (such as SASRec) model behavioral co-occurrence through self-attention mechanisms, they are still essentially memorizing and fitting historical interaction patterns and lack deep reasoning about user intent. When facing new users or new items, the system lacks an effective mechanism to transform external knowledge (such as item text descriptions, user profiles, and domain knowledge) into generalizable recommendation signals. This strong dependence on dense interaction data limits the applicability of traditional systems in data-scarce scenarios. In addition, the traditional paradigm struggles to effectively use the history of natural-language interactions between users and the system, and cannot dynamically clarify and refine user preferences through dialogue, further exacerbating the interaction bottleneck. In industrial practice, the cold-start problem directly affects new-user retention and new-item exposure, becoming a key bottleneck that constrains the commercial value of recommender systems. Traditional systems usually rely on popular recommendations or random exploration as compromise solutions for cold start, but this often comes at the cost of sacrificing personalization and cannot fundamentally solve the reasoning difficulties caused by interaction sparsity.

(3) Interpretability bottleneck

Traditional recommender systems have long taken prediction accuracy as their primary optimization objective, while transparency and interpretability are relegated to a secondary position. The black-box nature of deep models (such as DeepFM and DIN) makes recommendation decisions difficult for users to understand and audit. Existing explanation methods are mostly limited to visualization based on association rules or attention heat maps and cannot generate personalized recommendation rationales at the natural-language level. Users cannot know "why this item was recommended," nor can they correct system behavior in real time through natural-language feedback, resulting in insufficient trust and controllability. Wu Guodong et al. [5] discussed large language models and personalization

systematically sorted out the pathways for integrating personalized recommendation technologies, providing a classificatory foundation for subsequent research. Li et al. [10] comprehensively reviewed the principles and limitations of the various technical routes of the above deep-learning recommender systems. This lack of interpretability not only reduces users' acceptance of recommender systems, but also hinders systems from rapidly iterating and correcting errors through users' explicit feedback. In application scenarios that require high reliability—such as medical and financial recommendation—the interpretability bottleneck is particularly prominent, severely limiting the application scope of

traditional recommender systems.

3 Generative AI-Driven Recommender Systems: A Paradigm Shift

3.1 The Four Roles Through Which Large Language Models Reshape Recommender Systems

The emergence of large language models has brought recommender systems entirely new capabilities that transcend traditional technical routes. Unlike collaborative filtering, which relies only on user-behavior matrices, or deep-learning models, which extract features only from interaction data, large language models, by virtue of the world knowledge, semantic-understanding ability, and text-generation ability embedded in their massive pretraining corpora, can play multiple roles in recommender systems. According to existing research, the roles of large language models in recommender systems can be summarized into the following four categories:

First, feature encoders. Large language models can transform raw information such as user profiles, item descriptions, and user-behavior sequences into dense vector representations rich in semantic information. Compared with the feature representations of traditional recommendation models based on ID embeddings, the semantic vectors generated by large language models can capture deep semantic associations between texts, thereby still producing meaningful recommendations in cold-start scenarios—when new users or new items lack historical interaction data. For example, the P4R framework uses large language models to generate personalized item-description text through a Prompt strategy, then uses a pretrained BERT model to convert it into semantic embeddings, and finally performs collaborative filtering through a graph convolutional network, achieving performance superior to that of pure ID-embedding methods on multiple recommendation benchmark datasets.

Second, scoring functions. Large language models can use their pretrained knowledge and contextual-understanding ability to score the degree of match between users and items without specialized training. This zero-shot scoring capability enables recommender systems to break through the limitation that "historical interaction data must be available before recommendations can be made." Researchers typically organize user-preference descriptions and candidate-item lists in natural-language form as Prompts, input them into a large language model, and have the model directly output scores or ranking results.

Third, conversational interaction interfaces. Traditional recommender systems operate in a "black-box" manner: users can only passively accept recommendation results and cannot interact and communicate with the system. Large language models endow recommender systems with natural-language dialogue capabilities—users can actively express their needs ("I want to watch a lighthearted and funny short video"), the system can explain recommendation rea-

sons in natural language (“This video is recommended because you like similar pet content”), and users can revise their preferences (“I don’ t really want to watch pet videos recently”). This conversational recommendation fundamentally changes the interaction mode between users and recommender systems. For the complete technical development and classification system of conversational recommender systems, see the review by Jannach et al. [11].

Fourth, content generators. This is the most subversive role that large language models bring to recommender systems. Traditional recommender systems can only retrieve and rank existing content from an existing content library for distribution; recommender systems equipped with generative artificial intelligence can directly generate entirely new personalized content in response to user needs —a piece of text, an image, a short-video script, or even a piece of music that matches the user’ s current mood. The recommender system thereby transforms from a “porter of content” into a “creator of content.”

From “LLM-enhanced recommendation” to “generative recommendation,” the four roles above constitute a progressively advancing technological spectrum: the first three roles are essentially enhancements of traditional recommender systems—large language models optimize existing recommendation processes at the levels of feature extraction, score prediction, and interactive experience; whereas the fourth role—content generator—initiates a fundamental transformation of the recommendation paradigm. Ka Zuming et al. [4] systematically classified and sorted out the multiple roles of large language models in recommender systems from a Chinese-language perspective.

Visible text in Figure 2:

- User description; item ID; behavior sequence → semantic vectors
  Captures deep semantic associations in text and supports cold-start scenarios.
- Dialogue interaction interface
  Capability level: interaction
- Feature encoder
  Capability level: basic
  Zero-shot scoring, contextual understanding, directly predicts the degree of user matching.
- Large language model (LLM)
- Scoring function
  Capability level: basic + reasoning
  Context: directly predicts the degree of user matching.
- Content generator
  Capability level: creation
  Generates entirely new personalized content (text, images, video, music), achieving fundamental transformation.
- Enhancing traditional recommendation → Transforming the recommendation paradigm

Figure 2. Four roles of large language models in recommender systems

3.2 Generative Recommendation Paradigm

(1) Paradigm Definition and Core Architecture

Wang et al. [2] first systematically proposed the conceptual framework of the generative recommendation paradigm in 2023 and named it GeneRec. This paradigm points out that the retrieval-based recommendation model—retrieving, from an existing item catalog, content that matches user preferences—has two fundamental limitations. First, content pre-produced by humans is always finite and can hardly satisfy the information needs of all users across various scenarios. Second, users can usually influence recommendation results only through passive feedback such as clicks and ratings; this indirect way of expressing preferences is inefficient and imprecise. The rise of large language models and AIGC technologies provides key means for overcoming these limitations: generative artificial intelligence can create personalized content to meet users' diverse information needs, while the natural-language interaction capabilities of large language models enable users to express their needs accurately and efficiently. On this basis, GeneRec defines generative recommendation as a recommendation paradigm that "uses artificial-intelligence generators to generate personalized content and uses user instructions to guide content generation."

The core architecture of GeneRec consists of two key modules: the guider module is responsible for integrating and processing users' natural-language instructions and traditional behavioral feedback (clicks, dwell time, likes, etc.) into a unified generative guidance signal; the artificial-intelligence generator module then completes content retrieval, recombination, and creation driven by this guidance signal. Unlike the single output path of traditional recommender systems, GeneRec supports three progressive content-service modes: content retrieval—selecting the most matching content from an existing repository; content re-editing—modifying and reorganizing existing content according to users' personalized needs; and content creation—generating new personalized content from scratch to meet user needs.

Visible text in Figure 3:

- **User input**
  - User natural-language instructions (User Prompts): accurately express needs, express preferences, and provide personalized descriptions.
  - Traditional behavioral feedback (Behavioral Feedback): clicks, dwell time, likes, favorites, etc.
- **Guider module**
  - Instruction parsing: parses the semantics of user instructions.
  - Preference integration: integrates instructions and behavioral feedback.

  - Signal encoding: converts them into a unified generative guidance signal.
- **AI intelligent generator module**
  - AI editor: performs personalized reorganization and rewriting of existing content; shortened/extended videos, reorganized copywriting.
  - AI creator: generates entirely new personalized content from scratch; customized melodies, [[unclear: text describing generated summary/content]].
  - Fidelity Control: multi-level verification to ensure content consistency and authenticity; prevents factual errors and aligns content with the original materials/requirements.
- Existing content repository
- **Content service modes**
  - Content retrieval mode: content retrieval—selects the most matching content from the repository.
  - Content re-editing mode: content re-editing—modifies and reorganizes existing content.
  - Content creation mode: content creation—generates entirely new personalized content.
  - Retrieval result; re-editing result; generation result → final content delivery.

Figure 3. Schematic diagram of the core architecture of the generative recommendation paradigm (GeneRec)

(2) AI Editor and AI Creator

The artificial-intelligence generator in the GeneRec framework can be further divided into two functionally complementary subcomponents: the AI editor and the AI creator.

The function of the AI editor is to perform personalized reorganization and rewriting of existing content. For example, in short-video recommendation scenarios, the AI editor can automatically edit a relatively long video clip, according to the user' s preference for viewing duration, into a refined

a shortened or expanded version; in e-commerce recommendation, the AI editor can reorganize product-description copy according to users' stated preferences, highlighting the core informational dimensions that users care about. The core value of the AI editor lies in the fact that, while fully leveraging existing content resources, it improves the degree of personalized matching through intelligent content modification, while at the same time preserving the informational fidelity and auditability of the original content.

The function of the AI creator is to generate entirely new personalized content from scratch, unconstrained by an existing content repository. For example, in music recommendation scenarios, the AI creator can directly generate a customized melody based on the user's current emotional state and listening history;

in news recommendation, the AI creator can synthesize in real time a topical summary or explanatory text according to the user' s information needs. The AI creator fundamentally breaks through the upper limit of content supply, enabling recommender systems to truly possess the ability to "create on demand."

To ensure the reliability of generated content, the GeneRec framework places special emphasis on fidelity-control mechanisms: whether an AI editor reorganizes and rewrites existing content or an AI creator generates entirely new content, multiple layers of verification are required to confirm consistency between the content and the original materials or user requirements, thereby preventing the model from introducing factual errors or misleading information during the generation process.

(3) Comparative analysis with traditional retrieval-based recommendation

The core differences between generative recommendation and traditional retrieval-based recommendation are reflected in four dimensions: content source, user interaction mode, personalization mechanism, and content-supply boundary. A systematic comparison of the two paradigms across these four key dimensions is presented below.

Table 1 Comparison between traditional retrieval-based recommendation and generative recommendation

Comparison dimension	Traditional retrieval-based recommendation	Generative recommendation
Content source	Existing content repository	Dynamically generated or intelligent reorganization of existing content
User interaction	Mainly passive feedback (clicks, ratings)	Active instructions + natural-language dialogue
Personalization mechanism	Matches users to existing content	Creates exclusive content for users
Content supply	Limited by the scale of the content repository	Theoretically unlimited supply
Interpretability	Relatively weak; usually requires additional explanation modules	Relatively strong; naturally supports natural-language explanations
Cold-start handling	New users/new items lack signals	Uses world knowledge to provide initial recommendations

Comparison dimension	Traditional retrieval-based recommendation	Generative recommendation
Core challenge	Sparsity and coverage	Generation quality and fidelity control

As shown in Table 1, generative recommendation has significant advantages in terms of personalization upper bounds, interaction naturalness, and supply flexibility, but it also introduces the new challenge of generated-content quality control, which traditional recommender systems do not need to face. The two paradigms are not in a complete substitution relationship—in mass recommendation scenarios where the content repository is sufficiently abundant and user preferences are stable, traditional retrieval-based recommendation still has the advantage of computational efficiency; whereas in scenarios involving long-tail needs, high personalization requirements, and the need for interactive guidance, generative recommendation demonstrates irreplaceable value.

3.3 Gen-RecSys Comprehensive Classification Framework

In the survey published by Li et al. [1] in 2024, the application landscape of large language models in generative recommendation was systematically organized, and a unified classification framework covering multiple technical routes was proposed.

(1) ID-driven generative models

The first class of technical routes consists of ID-driven generative models, which mainly rely on two classic deep generative models: generative adversarial networks and variational autoencoders. Typical applications of such methods in recommender systems include: using generative adversarial networks to learn the interaction distribution between users and items, thereby generating more realistic user-behavior simulation data to alleviate data-sparsity problems; and using variational autoencoders to learn latent-space representations of user preferences, generating diversified recommendation lists by sampling in the latent space. The advantage of ID-driven methods lies in their good compatibility with existing recommender-system technology stacks: they can operate directly on the embedding spaces of user IDs and item IDs. However, their limitations are also relatively obvious—their generative capability is constrained by the amount of information encoded in the ID embedding space, and they cannot introduce external semantic knowledge.

(2) LLM-driven generative recommendation

The second technical route is the core focus of this paper—generative recommendation driven by large language models. Compared with ID-driven methods, LLM-driven methods have three prominent advantages. First, the world knowledge accumulated during large language model pretraining enables them

to understand semantic associations among items, rather than relying only on statistical correlations based on collaborative signals. Second, the natural-language capabilities of large language models make the recommendation process explainable, conversational, and negotiable. Third, large language models can simultaneously complete multiple subtasks in the recommendation pipeline, such as user intent parsing, candidate-content filtering, and recommendation-explanation generation, offering the prospect of unifying multiple dispersed model components in traditional recommender systems. Compared with ID-driven methods, LLM-driven methods have three prominent advantages. Xie Guangming et al. [6] provide a comprehensive survey of recommendation systems driven by large language models, covering the two major technical routes of discriminative and generative approaches. The Alibaba M6-Rec mentioned in the survey by Li et al. [1] is a representative work along this route. It converts user behavioral data into natural-language text, unifies multiple tasks in recommender systems—including retrieval, ranking, explanation generation, and even personalized content creation—within a single large-language-model-based architecture, and uses improved prompt-tuning techniques to achieve cross-task adaptability with very few task-specific parameters.

(3) Multimodal generative recommendation

The third technical route is multimodal generative recommendation. Real-world recommendation scenarios often involve multiple information modalities such as text, images, video, and audio, and user preferences and content characteristics are also embedded in multimodal signals. Multimodal generative recommendation aims to use large language models or multimodal large models to simultaneously understand and generate multimodal content during the recommendation process. A typical example is the DualFashion model, which adopts a dual-diffusion Transformer architecture to process both image and text modalities at the same time. In fashion outfit recommendation, it can generate both outfit-effect images and semantic outfit descriptions. Multimodal generative recommendation effectively overcomes the loss of information in visual, auditory, and other dimensions caused by purely text-based recommendation, and represents an important direction in the development of generative recommendation toward a "full-modality personalized experience." Zhang Rui and Bian Zhipeng [7] systematically survey the classification system and application scenarios of multimodal generation techniques in recommender systems.

4 Core Technologies of Generative Recommendation

4.1 Overview of the Three-Stage Evolution

Recommender-system modeling algorithms have gone through three major technical paradigms: machine-learning recommendation (MLR), deep-learning recommendation (DLR), and generative recommendation (GR). MLR centers on collaborative filtering and matrix decomposition, relying on the user–item interaction matrix to perform statistical modeling. DLR centers on MLPs and

embedding representations, using deep networks to mine nonlinear feature interactions; however, multilayer cascade architectures have long suffered from shortcomings such as semantic fragmentation and feature pulverization.

Generative AI is driving recommender systems from the traditional retrieval-and-ranking mode toward a native generative paradigm. Traditional recommendation can only filter existing content from a fixed item repository and cannot autonomously create content. The GeneRec paradigm proposes using a generator to complete personalized content generation and verifies its feasibility in short-video scenarios. In 2025, Kuaishou released OneRec, the industry' s first deployed industrial-grade end-to-end generative recommendation solution.

Figure 4 Framework for the three-stage technological evolution of generative recommendation

- **Stage 1: LLM-augmented recommendation**
  - Prompt-based representation learning
  - Causality-enhanced behavioral sequence modeling
  - LLM-generated recommendation explanations
  - P4R / CoT / Feng et al.
  - LLM-assisted traditional recommendation process

→ **Deepening of user-intent understanding**

- **Stage 2: Core technologies of generative recommendation**
  - Semantic-ID generative retrieval
  - Multimodal content generation
  - Unified value-alignment generation
  - GeneRec / GPT4Rec / UniVA
  - LLM as a recommendation generator

→ **Expansion of personalization dimensions**

- **Stage 3: Emotion awareness and multi-agent collaboration**
  - Multimodal emotion-aware recommendation
  - Multi-agent collaborative controllable filtering
  - MMEI / Zhang et al.
  - Multi-agent collaboration and emotion awareness

From the perspective of technological iteration logic, generative recommendation can be divided into three development stages: Stage 1 is LLM-augmented recommendation, in which large models are used to optimize components such as user profiling, while core decision-making is still undertaken by traditional models; Stage 2 reconstructs the recommendation task as "next-token generation," realizing discrete tokenization of items through semantic IDs; Stage 3 integrates emotion awareness and multi-agent collaboration to achieve personalized recommendation that is empathetic, controllable, and adjustable. The core differences between traditional recommendation and generative recommendation are shown in the following table.

Table 2 Core differences between traditional recommendation and generative recommendation

Dimension	Traditional recommendation	Generative recommendation
Core task	Retrieve existing items and rank them	Generate personalized content
Modeling approach	Discriminative scoring function	Autoregressive sequence generation
User interaction	Implicit feedback	Natural-language instructions + implicit feedback
Cold-start capability	Relies on interaction data	Zero-/few-shot learning capability

4.2 Stage 1: LLM-Augmented Recommendation

LLM-augmented recommendation is the initial form through which generative AI is implemented in the recommendation domain. Recommendation systems empowered by large models can be divided into two major categories: discriminative and generative. The generative route mainly conducts research in four directions: direct recommendation, representation learning, generative learning, and prompt learning. The gains that large models bring to recommendation systems are mainly reflected in the following three aspects.

First is Prompt-based representation learning. With the aid of Prompt templates, LLMs can integrate heterogeneous information and generate vector representations rich in semantic information; when combined with instance-level prompts and multi-agent reinforcement learning, they can customize dedicated discrete prompts for different users, effectively alleviating the cold-start problem. The Prompt4NR framework converts user click logs and candidate news into natural-language text and designs dedicated prompt templates.

Second is behavioral sequence modeling that integrates causal reasoning. Traditional sequential models can only capture temporal associations and have difficulty distinguishing correlation from causation; by contrast, relying on chain-of-thought reasoning capabilities, LLMs can mine underlying real needs from users' behavioral sequences. A path-free prompt constructor can transform behavioral graph-structure information into textual prompts, and this approach has already been implemented and validated in an online recruitment recommendation business.

Third is the generation of recommendation explanations by large models. Compared with traditional recommendation schemes, LLMs, in terms of semantic understanding and contextual percep-

knowledge, external knowledge, and interpretability, but this technology still suffers from factual hallucinations and inherent biases, and faces six major challenges, including recommendation bias and prompt fragility.

4.3 Phase Two: Core Technologies for Generative Recommendation

Phase two is the key stage in which generative recommendation shifts from "enhancement" toward "reconstruction." Recommendation systems are redefined as sequence-generation tasks, capable of actively acquiring knowledge from the external world to improve cold-start and generalization capabilities. This transformation involves four core technical dimensions.

Table 3 Technical Dimensions of Phase Two

Dimension	Core Positioning	Key Output
Semantic ID generative retrieval	Representation revolution	Next-token prediction paradigm
Multimodal content generation	Expansion of capability boundaries	New content generation
Unified value alignment	Industrial implementation	Integration of commercial value
Inference acceleration	Engineering lifeline	Real-time services

Semantic ID generative retrieval is the most central innovation. Semantic ID encodes each item as a discrete token sequence, transforming the recommendation task into an autoregressive generation problem of "given the SID sequence of a user' s historical interactions, predict the next SID token," fully aligning with the next-token prediction paradigm of LLMs.

Multimodal content generation introduces generative models such as diffusion models, enabling the system to create new content. The GPT4Rec framework first generates hypothetical queries and then performs retrieval-based recommendation; in short-video scenarios, DimeRec uses diffusion models to achieve sequential recommendation enhancement and has already been deployed on platforms with hundreds of millions of users. Multimodal semantic IDs incorporate modality fusion into the SID learning process.

Industrial-grade unified value alignment is the key to real-world deployment. The UniVA framework consists of three layers: the Commercial SID tokenizer injects value attributes into the SID construction process; the Generation-as-Ranking decoder jointly optimizes generation and ranking through supervised learning and reinforcement learning; and value-guided personalized beam search reuses logits as online guidance. On the WeChat Channels advertising platform, UniVA improved offline Hit Rate@100 by 37.04% and increased online GMV by 1.5%.

Inference acceleration is the engineering bottleneck for real-time services. The RcLLM series designs mechanisms such as reusable block decomposition, reducing first-token latency by 1.31× to 9.51×, with negligible impact on accuracy.

4.4 Phase Three: Emotional Perception and Multi-Agent Collaboration

Phase three focuses on emotional perception and multi-agent collaboration, including two technical routes: multimodal emotional perception and multi-agent collaborative controllable filtering. After the first two phases separately address semantic enhancement and generative reconstruction, phase three extends the technical perspective from "items" to "users," exploring the value of emotional signals and user controllability in generative recommendation.

(1) Empathetic Conversational Recommendation

Zhang et al. [3] proposed a technical framework for empathetic conversational recommendation systems at RecSys 2024. This study introduces empathy into recommendation systems, enabling the system to capture users' emotional-state signals during conversational interaction and dynamically adjust personalized recommendation strategies accordingly. Unlike traditional recommendation systems, which take click-through rate and conversion rate as their core optimization objectives, empathetic recommendation emphasizes the system's perception of and response to users' emotional needs—users are not only passive recipients of information, but participants who engage in bidirectional empathetic interaction with the recommendation system. This study, through large-scale [[unclear: text continues after "实"]].

verified the positive impact of empathy signals on recommendation relevance and user satisfaction, and was published at a top international conference in the field of recommender systems. The deployment of empathetic recommender systems faces technical challenges such as the complexity of modeling conversational context, the subjectivity of annotating empathy signals, and latency requirements for real-time interaction. Yet these challenges also indicate an important direction for recommender systems to move from "precise distribution" toward "services with warmth."

(2) Multi-Agent Collaborative Controllable Filtering Framework

The multi-agent collaborative filtering framework realizes transparent and controllable recommendation filtering through multi-agent orchestration with edge-cloud collaboration. This framework targets two technical bottlenecks: purely textual LLMs cannot identify inappropriate content at the visual level; and users' negative feedback may be incorrectly generalized by LLMs to similar topics, producing an "over-association" hallucination. The framework contains two core designs: first, a fact-grounded adjudication pipeline introduces fact-checking agents, shifting filtering decisions from model inference to support by factual evidence. It includes roles such as data retrievers, quality supervisors,

hallucination detectors, and candidate evaluators, with outputs aggregated by a coordinator. Second, a dynamic two-layer preference graph divides user preferences into a "long-term preference layer" and an "incremental adjustment layer"; the latter allows users to make explicit and reversible preference corrections through Delta adjustments, while the two layers are updated orthogonally to avoid catastrophic forgetting. Experiments show that this framework reduces the false-positive rate by 74.3%, and that its F1 score is significantly better than that of the pure-text baseline. Field studies verify improvements in algorithmic transparency and users' sense of control. This framework indicates that generative recommender systems can move from algorithmic autonomy toward human–machine co-governance through multi-agent collaboration, providing a technical pathway for compliance in platform content distribution.

Visible labels in the diagram: explicit Delta adjustment correction; long-term preference layer; incremental adjustment layer; more details; long-term & incremental features; recommendation generation and distribution; user-personalized and compliant recommendation stream.

Figure 5 Dynamic Two-Layer Preference Graph

4.5 Comparative Summary of Key Technical Methods

Table 2 systematically compares the key technical methods in the three stages of generative recommendation across seven dimensions: core objectives, technical paradigms, representative methods, user interaction modes, core limitations, and evolutionary directions.

Table 4 Comparative Summary of Key Technical Methods in the Three Stages of Generative Recommendation

Comparison Dimension	Stage 1: LLM-Enhanced Recommendation	Stage 2: Core Technologies of Generative Recommendation	Stage 3: Emotion Perception and Multi-Agent Collaboration
Core objective	Use the semantic capabilities of LLMs to improve recommendation accuracy	Reconstruct the recommendation process with generative models	Perceive user emotions and endow the system with controllability
Technical paradigm	LLMs are embedded into traditional processes as auxiliary modules	LLMs serve as the core engine for end-to-end generation	Multi-agent collaborative decision-making, driven by emotional signals

Representative methods	Prompt representation learning; causal behavioral sequence modeling; LLM-based recommendation explanations	Semantic ID retrieval; multimodal generation; UniVA alignment; RcLLM acceleration	Empathic conversational recommendation [3]; multi-agent collaborative filtering
User interaction mode	Implicit inference	Implicit generation	Explicit + implicit fusion
Core limitations	High inference cost; limited real-time performance	Unstable content quality; weak interpretability	Emotion recognition depends on data quality and privacy compliance
Evolution direction	From feature enhancement toward capability substitution	From candidate generation toward decision generation	From algorithmic autonomy toward human–machine co-governance

Stage I uses LLMs to enhance representation learning, sequential modeling, and explanation generation in traditional recommendation pipelines; the LLM remains an auxiliary module, while the decision-making subject is still a discriminative model. Stage II reconstructs recommendation as a "next-token generation"problem, elevating the LLM to the core engine and shifting recommendation from "retrieving existing items" to "generating personalized content." Stage III introduces affective signals and user-control mechanisms, enabling recommender systems to possess empathic capacity and interveneability. In addition to the vertical technological iteration across the three stages, knowledge-enhancement techniques can serve as a general optimization path to compensate for common defects in generative recommendation across stages, such as factual insufficiency and unreliable reasoning.

4.6 Knowledge-Enhanced Generative Recommendation

Stages I to III respectively address semantic enhancement, paradigm reconstruction, and affective controllability, but all three share an underlying limitation: LLMs' understanding of recommended items comes from textual co-occurrence rather than factual knowledge, and knowledge-intensive scenarios are prone to producing results that are "semantically fluent but insufficiently grounded." Through external knowledge graphs, retrieval-augmented generation, and struc-

tured knowledge injection, knowledge enhancement builds a bridge for generative recommendation from "semantic association" to "factual reasoning."

(1) Recommendation reasoning enhanced by knowledge graphs

Knowledge-graph enhancement mainly has three integration pathways:

Entity alignment and preference propagation: Items are linked to graph entities for graph-structured preference propagation. Unlike the implicit vector similarity of semantic IDs, this pathway provides explicit reasoning chains.

Joint GNN-LLM encoding: A dual-encoder architecture is adopted: the GNN aggregates multi-hop entity information to generate knowledge embeddings, while the LLM encodes user queries; cross-modal attention is then used for fusion, alleviating the problem of data sparsity.

Knowledge-aware explanation generation: Explanations are generated by retrieving the shortest paths between preferred entities and recommended items in the graph. Each step is supported by factual triples, providing a supplementary path for addressing the "weak interpretability" of Stage III.

(2) Recommendation applications of retrieval-augmented generation

RAG provides dynamic knowledge support through "retrieval–generation." In profile construction, background knowledge related to user behavior is retrieved, expanding the profile from "what the user clicked" to "what the user cares about." In content generation, factual information is retrieved and injected into the LLM as decoding constraints, ensuring consistency between the content and the source material and forming a progression of capability together with recommendation-explanation generation. In cold start, external knowledge retrieval is performed on the metadata of new items to construct initial representations, supplementing the limitation that semantic IDs depend on interaction data.

[Flowchart labels: User query; Current profile; Profile update (clicks → depth of attention); Dynamic retriever; Retrieval strategy; Retrieved factual knowledge; External knowledge base / fact source; Evidence injection; LLM generator; Decoding-constraint injection; LLM model; Recommended content.]

Figure 6 Workflow of RAG-enhanced recommendation

(3) Structured Knowledge Injection and Preference Alignment

Business-rule encoding converts constraints such as price sensitivity and school-stage matching into rule templates and incorporates them into decoding, complementing UniVA value alignment. Explicit symbolic modeling of preferences represents preferences as an editable rule set, complementing the "incremental adjustment layer" of the dynamic two-layer preference graph. The three types of methods each have their own emphasis: knowledge graphs are suitable for vertical domains, RAG is suitable for time-sensitive scenarios, and structured

injection is suitable for scenarios with strong rules; in industrial practice, they are often used in combination.

The evolution of the three stages is progressive and mutually complementary. Stage 1 provides a preliminary validation of LLM capabilities for Stage 2; Stage 2 lays the technical foundation of generative models for Stage 3; and Stage 3 introduces emotional signals and mechanisms for explicit user intervention, extending the optimization objective from "accuracy" to "empathy and controllability." Knowledge enhancement continuously injects factual knowledge at all stages, so that the capability improvement of recommender systems goes beyond being "more accurate" and "more understanding," and moves toward being "more reliable." At present, industrial generative recommendation is in the period of large-scale implementation for Stage 2; systems such as WeChat Channels UniVA have already verified technical feasibility in real-world scenarios. The affective perception and multi-agent collaboration of Stage 3, together with the guarantee of factual reliability provided by knowledge enhancement, jointly represent the next evolutionary direction from "technically feasible" toward "trusted by users."

5 Typical Application Scenarios and Industrial Practice

Recommender systems driven by generative AI and LLMs have moved from academic research to large-scale industrial implementation. In 2025–2026, mainstream platforms at home and abroad have successively integrated generative AI into the recommendation pipeline, from feature encoding and user understanding to content production and distribution, forming distinctive technical paths. This section reviews typical application scenarios from three dimensions—social-media short video, e-commerce advertising, and digital content platforms—and conducts a comparative analysis of industrial practices in China and abroad.

5.1 Social Media and Short-Video Platforms

(1) The recommendation system of Douyin/TikTok

Douyin and its international version TikTok are benchmarks for engineering practice in industrial-grade recommender systems. Their traditional recommender systems adopt a multi-stage funnel architecture of "recall → coarse ranking → fine ranking → reranking," with the core logic of traffic distribution built on deep learning. In the recall stage, deep-learning-based model recall strategies are used; in the fine-ranking stage, the architecture has evolved from MLP to more complex structures such as DeepFM, DCN, MMOE, and PLE, and LHUC has been introduced to achieve neuron-level personalized perception.

After the introduction of generative AI, Douyin' s recommender system was further upgraded. At the user-understanding level, users' historical behaviors are input into an LLM to generate fine-grained interest tags—for example, inferring deeper preferences such as "snowboarding" from "liked skiing videos." At the content-understanding level, multimodal large models are used to automatically

annotate and summarize videos, generating multidimensional representations such as scenes, people, actions, and emotions. In terms of generative recommendation, Douyin has tested generative retrieval technology, treating candidate-set generation as a sequence-generation problem and directly generating sequences of video IDs that match user interests, yielding positive benefits in long-tail distribution.

(2) Industrial practice of WeChat Channels

Relying on the social relationship chain of the WeChat ecosystem, WeChat Channels has formed a differentiated recommendation path. In June 2026, WeChat opened AI ecosystem interfaces, allowing third-party AI services to access the recommendation and creation pipeline and fully embrace generative AI.

The core feature of the Channels recommender system lies in integrating social relationships with personal interests. LLMs are used to understand "good

the two layers of semantics— "why friends like it" and "whether the user will also like it" —thereby more accurately balancing social signals and individual preferences. On the content-production side, Video Accounts provides AI writing assistants, AI video-editing tools, and AI title-generation tools; on the distribution side, generative AI dynamically generates personalized cover images, teaser clips, and copywriting, achieving full-chain coverage from content production to distribution. In addition, Video Accounts is exploring LLM-based end-to-end generative recommendation, generating candidate sets conditioned on users' histories and social contexts so as to reduce information loss in multi-stage funnels.

(3) Xiaohongshu RED Skill

RED Skill, launched by Xiaohongshu in 2025–2026, is a core component of its AI ecosystem, embedding AI capabilities into all stages of content recommendation and creation.

The application of RED Skill in recommendation is mainly reflected in four aspects. First, in intent understanding and query rewriting, it uses LLMs to perform deep analysis of users' search terms and recommends personalized plans by integrating information such as weather, transportation, and budget. Second, in multimodal content tag generation, it automatically generates structured tags for images, text, and videos, enhancing the granularity of content understanding. Third, in personalized content summarization, it generates different summary copy or cover images for the same note according to user preferences, achieving personalization at the presentation layer. Fourth, on the creator side, it provides functions such as AI-assisted writing, image optimization, and title suggestions, thereby indirectly enriching the recommendation candidate pool.

5.2 E-commerce and Advertising Recommendation

(1) Alibaba M6-Rec Unified Recommendation Foundation

M6-Rec, derived from Alibaba' s M6 large-model series, is a representative industrial-grade unified recommendation foundation. In terms of technical architecture, M6-Rec is based on a unified Transformer architecture, integrating recall, ranking, and re-ranking into a single model. It takes user behavior sequences, product attributes, and user profiles as inputs, and learns interaction patterns through self-attention. M6-Rec uses LLMs to semantically encode product titles, descriptions, and reviews, enabling cold-start products to obtain reasonable recall and ranking based on content features, thereby compensating for the shortcomings of traditional systems that rely on IDs and statistical features. In generative recommendation practice, M6-Rec attempts to directly generate product titles or combinations of attributes that users may be interested in, showing potential in satisfying open-ended needs.

(2) Generative Advertising Recommendation

Generative AI is reshaping the advertising recommendation pipeline. In terms of ad creative generation, advertisers can use AI to quickly generate multiple versions of creative materials and conduct A/B testing, while the system automatically selects the optimal creative combination based on user profiles. In terms of optimizing user–advertisement matching, LLMs are used to understand the semantic match between advertising intent and user needs; for example, after a user searches for "fitness,"the system can infer that the user is at a "beginner stage" and push introductory tutorials or lightweight equipment. In terms of interactive advertising, some platforms are experimenting with AI-driven conversational advertisements: users can interact with ads through natural language, and AI generates personalized recommendations and answers in real time, improving conversion rates compared with traditional advertising.

5.3 AIGC Distribution on Digital Content Platforms

Digital content platforms are using generative AI to optimize content distribution. For news and information platforms, AI generates personalized news summaries and morning briefings based on users' reading histories, and uses LLMs to interpret events from multiple perspectives. For online education platforms, AI recommendation systems generate personalized exercises and review plans using LLMs according to learners' knowledge graphs and weak points, realizing an integrated "recommendation + generation" learning path. For digital reading platforms, personalized book-list recommendation copy and book summaries are generated based on user preferences, and user trust is enhanced by displaying the operating mechanism of the model.

Radar chart title: Multidimensional Comparison Radar Chart of AI Recommendation Practices on Mainstream Content Platforms in China and Abroad
Legend: WeChat Channels; Xiaohongshu; Douyin; TikTok
Dimensions: Depth of personalization; Naturalness of interaction; User controllability; Degree of privacy protection; Richness of the content ecosystem

Figure 7. Multidimensional Comparison of AI Recommendation

5.4 Comparison of Industrial Practices in China and Abroad

Industries in China and abroad differ in their strategic emphases in responding to generative recommendation. Benefiting from ultra-large-scale user data and rich content ecosystems, domestic platforms have made rapid progress in LLM-enhanced recommendation; overseas efforts are more cutting-edge in end-to-end generative recommendation and privacy-preserving computation technologies. Future recommendation systems should place greater emphasis on user data security and privacy protection, and adopt stricter data-protection measures.

6 Challenges and Governance: Privacy, Fairness, and Controllability

While generative recommendation technology significantly improves the capability for personalized recommendation, it also introduces a series of governance challenges that cannot be ignored. The more users need precise personalized services, the more deeply the system must understand users' behavioral patterns, interest preferences, and even emotional states. This inevitably pushes recommendation systems into a fundamental tension: the higher the degree of personalization, the deeper the dependence on user data, and the greater the accompanying privacy risks and fairness concerns. This section conducts a systematic analysis from four dimensions: privacy protection and data security, algorithmic fairness and bias, user controllability and algorithmic transparency, and the quality and safety of generated content.

6.1 User Privacy Protection and Data Security

(1) The Privacy Paradox and Technical Responses

Personalized recommendation systems face a classic privacy paradox: the better the recommendation performance, the deeper the system' s understanding of the user; yet such understanding must necessarily be built upon extensive collection and in-depth analysis of users' personal data. Traditional recommendation systems usually collect multidimensional data such as users' browsing histories, click behaviors, dwell time, social relationships, and geographic locations. On this basis, generative recommendation further involves natural-language instructions actively entered by users and feedback on emotional states; the information density and sensitivity of these data far exceed those of behavioral logs. Taking multimodal affect-aware recommendation as an example, the system needs to collect three modalities of user data—facial expressions, speech intonation, and textual comments. Among them, facial expressions and speech intonation are highly sensitive biometric information; once leaked or abused, they will cause serious privacy harm to users.

In response to this challenge, academia and industry have explored multiple

privacy-preserving technologies. Federated recommendation is one of the most widely discussed directions, and Yin et al. [9] provide a systematic review of it. Its core idea is to complete joint training of recommendation models under the premise that user data never leave local devices: each user device trains the model locally using personal data, and uploads only encrypted model gradients to a central server for aggregation, thereby avoiding centralized storage and transmission of raw personal data. Differential privacy technology, by contrast, injects controlled noise into model parameters during training to ensure that attackers cannot infer any individual user' s personal information from the model outputs. In addition, edge-side inference technology migrates part of the computation of large language models from the cloud to users' local devices, so that sensitive user instructions and feedback data remain in

processing is completed locally, with only the desensitized recommendation request sent to the remote service. These technologies each have their advantages and disadvantages: federated recommendation entails relatively high communication overhead; differential privacy may appropriately reduce recommendation accuracy as the cost of protection; and on-device inference is constrained by the computational capacity of the device. In practical applications, multiple technologies often need to be combined in order to achieve an acceptable balance between the degree of privacy protection and the quality of recommendation services.

(2) Regulatory compliance requirements

Technical privacy-protection measures must be coordinated with legal compliance requirements. Worldwide, data-privacy legislation is advancing rapidly. China' s *Personal Information Protection Law*, formally implemented in 2021, clarified principles such as minimum necessity, informed consent, and purpose limitation in personal-information processing, and requires that automated decision-making processes ensure transparency and fairness and impartiality of results. The EU *General Data Protection Regulation* grants users the "right not to be subject to purely automated decision-making," and, in the *Artificial Intelligence Act* introduced in 2025, brings recommender systems within the regulatory scope of high-risk AI systems, requiring transparency disclosures and risk assessments for recommendation algorithms. In June 2026, the European Commission issued a provisional measures order to Meta, requiring it to provide WhatsApp users with a service option not subject to AI recommendation-content ranking algorithm controls, marking the shift of recommendation-algorithm governance from statutory provisions toward concrete enforcement practice. These regulatory requirements mean that generative recommender systems must incorporate compliance into architectural considerations from the outset of design, rather than relying on remedial measures after the fact.

6.2 Algorithmic Fairness and Bias

(1) The stereotype problem in LLM recommendations

Large language models are pretrained on massive volumes of Internet text, and their training corpora inevitably contain stereotypes and biases that already exist in society. When large language models are applied to recommendation tasks, the biases learned during pretraining may influence recommendation results in hidden ways. Wang et al. [8] systematically reviewed fairness in recommender systems, pointing out that recommender systems may produce differentiated service quality across multiple dimensions—including race, gender, and socioeconomic status—and that users often lack awareness of such algorithmic bias. This bias permeates personalized recommendation in a manner that is difficult for users to detect: users may believe that recommendations are objective results based purely on their own interests and preferences, whereas in fact the recommender system has already introduced systematic unequal treatment along the dimension of user identity. From the perspective of digital publishing and content distribution, this means that users from different social groups may receive content recommendations of significantly different quality without being aware of it, and long-term accumulation will lead to inequality in opportunities to obtain information.

(2) Information cocoons and over-association

The "information cocoon" problem for which traditional recommender systems have been widely criticized—namely, that algorithms continuously push homogeneous content in which users have already shown interest, causing users' horizons to gradually narrow—may intensify in new forms in the era of generative recommendation. When a recommender system not only chooses what content to push but can also generate what kind of content, the boundary between personalization and information siloing becomes even more blurred. Even more hidden is the "over-association" problem revealed by existing studies: when recommendation systems based on large language models interpret users' negative preferences, they tend to incorrectly generalize a user' s aversion to a particular piece of content to semantically similar but unproblematic information categories. For example, if a user expresses aversion to anxiety-inducing marketing content, the model may therefore mistakenly block normal health-science articles. This kind of over-association causes recommender systems to generate large numbers of false positives, invisibly depriving users of the right to obtain beneficial information, while users lack awareness of and channels to correct it.

6.3 User Controllability and Algorithmic Transparency

(1) Human-machine collaborative governance mechanisms

In the face of the increasing complexity and influence of recommendation algorithms, the idea of transforming users from passive recipients into active participants in governance is gaining more and more attention. Existing research has

proposed a transparent and controllable recommendation-filtering framework whose core innovations include two key designs: first, constructing a two-layer dynamic preference graph to explicitly record users' acceptance and rejection preferences across different information dimensions, and allowing users to directly modify nodes and edges in the graph in a 方

···adjust their own preference configurations; second, introducing a multi-agent collaborative fact-checking pipeline, in which multiple AI agents with different professional judgment capabilities cross-validate recommendation filtering decisions, thereby reducing misjudgments caused by inference from a single model. In a seven-day longitudinal user study of this framework, participants not only reported a significant increase in satisfaction with the recommendation results, but also indicated that the improved transparency of the algorithm alleviated their previous concern about missing important information by mistake. This study reveals a deeper trend in the field: the design goals of recommendation systems should not be limited to increasing click-through rates and dwell time, but should also include enhancing users' understanding of, and sense of control over, the algorithm.

(2) Design of Explainable Recommendation Explainability is the foundation of users' trust in recommendation systems. The explainability of traditional recommendation systems is usually presented in the form of post hoc rule explanations or visualizations of attention weights, and the threshold for user understanding is relatively high. The powerful natural-language capability of large language models provides an entirely new path for implementing explainable recommendation. Relevant large-scale user experiments show that when a large language model recommends a film while also providing a contextual explanation based on the user' s past viewing preferences—for example, "This film is recommended because you previously showed strong interest in similar suspense-themed works, but the narrative rhythm of this work is gentler than the films you usually like" —such situated explanatory capability can effectively satisfy users' cognitive needs and significantly increase their willingness to watch the recommended film. However, this study also reveals a noteworthy phenomenon: overly detailed or overly frequent explanations make users feel excessively attended to and thus uncomfortable. This subtle boundary suggests that the design of explainability in recommendation systems needs to find a delicate balance between informational sufficiency and user comfort.

6.4 Quality and Safety of Generated Content

Generative recommendation endows recommendation systems with the ability to create content, but this capability also introduces new risks that traditional recommendation systems do not need to face: the generated content may contain factual errors, logical contradictions, inappropriate information, or even malicious content. Traditional recommendation distributes human-produced and reviewed content, so content quality has a basic assurance threshold; by

contrast, the content dynamically created by generative recommendation lacks a prior review stage. If the generative model hallucinates or is attacked by malicious prompts, it may output harmful information to users. Fidelity control in generative recommendation has therefore become a key technical challenge. The multilayer verification mechanism proposed by the GeneRec framework—including fact checking at the content level, format-standardization checks, and consistency verification of requirements at the user-instruction level—represents a technical exploration in this direction. In addition, assigning responsibility for reviewing generated content to three levels— "model self-checking—platform spot checks—user feedback" —and building a human-machine collaborative quality and safety protection system are gradually becoming an industry consensus.

7 Future Prospects and Conclusion

7.1 Technical Trends

At the technical level, generative recommendation will continue to evolve in the following directions. First, the paradigm shift from "recommendation as retrieval" to "recommendation as generation" will deepen further. As the inference efficiency of large language models continues to improve, more and more recommendation scenarios will adopt generative modes to replace or supplement retrieval modes. Second, recommendation architectures featuring multimodality and multi-agent collaboration will become the next technological focus. Through the division of labor among multiple specialized agents—such as intention understanding, content generation, quality verification, and user feedback—a collaborative pipeline from demand perception to content delivery will be built. Third, privacy-preserving recommendation driven by on-device large models will accelerate deployment as chip computing power improves and model-compression technologies mature. Fourth, empathic recommendation systems with real-time affective perception will gradually move from academic exploration toward productized applications.

7.2 Industry Trends

At the industry level, represented by the WeChat Open Platform AI ecosystem and Xiaohongshu RED Skill, China' s leading content platforms are taking the lead in building "AI-native recommendation platforms." The three major links of content production, distribution, and consumption are being connected in series by generative AI into a real-time interactive feedback loop. The business model of recommendation systems will also change accordingly—when platforms can generate exclusive content directly for users, recommendation systems will evolve from pure traffic-distribution tools into value platforms that integrate content creation and distribution.

7.3 Governance Trends

At the governance level, algorithm auditing and recommendation transparency will shift from voluntary initiatives to legally mandated requirements. As users' awareness of data sovereignty continues to grow, legal rights such as "data portability rights"and "the right not to be subject to automated decision-making constraints" will become increasingly enforceable through the transformation of technical infrastructure. Human–machine collaborative governance will move from an academic concept into engineering practice, and users' control over recommendation results will evolve from a coarse-grained binary choice into multidimensional, fine-grained preference configuration.

7.4 Open Issues

Although generative recommendation technology is developing rapidly, the following open issues remain to be explored. First, how to establish a quantitative trade-off model between the depth of personalization and the strength of privacy protection still lacks a unified theoretical framework. Second, the inference latency and computational consumption of large language models remain far higher than those of traditional retrieval-based recommendation, and the economic feasibility of large-scale industrial deployment remains to be verified. Third, establishing a standardized evaluation system for generative recommendation requires consideration of multiple dimensions, including content quality, factual accuracy, diversity, and users' emotional experience. Fourth, how to improve the transferability of generative recommendation across domains and cultures remains at an early stage of research.

7.5 Conclusion

Taking "from retrieval to generation"as its main thread, this paper has systematically reviewed content recommendation and personalized distribution technologies driven by generative AI, and has mainly reached the following conclusions.

First, recommendation systems are undergoing a paradigm shift from retrieval-based distribution to generative distribution. Generative recommendation represented by the GeneRec paradigm realizes intelligent reorganization of existing content and on-demand generation of entirely new personalized content through two core components—AI editors and AI creators. In this process, large language models play a complete capability-spectrum role, from feature encoding to content generation.

Second, the three-stage technological evolution framework of "enhancement—generation—collaboration" proposed in this paper provides a structured perspective for understanding the technological trajectory of this field, corresponding respectively to three progressive stages: LLM-assisted traditional recommendation, LLMs as recommendation generators, and multi-agent collaboration with affective perception.

Third, privacy protection, algorithmic fairness, user controllability, and content safety constitute the four major governance pillars for the sustainable development of generative recommendation. Technological progress and governance development must advance in parallel.

Fourth, leading Chinese content platforms such as WeChat Channels, Xiaohongshu, and Douyin are taking the lead in exploring the deep integration of generative AI and recommendation systems, with the role of recommendation systems expanding from information-filtering tools to value-creation platforms.

The limitations of this paper are mainly as follows: generative recommendation is a rapidly evolving field, so the timeliness of the review is inherently limited; there is relatively little quantitative analysis of commercial feasibility assessments and user experience; and research outcomes in languages other than Chinese and English have not been included. Looking ahead, key issues to be addressed include solutions for balancing personalization depth with privacy protection, trade-offs between generation quality and computational cost, and the establishment of standardized evaluation systems.

Author Contribution Statement

Han Yuxuan: proposed the research questions and review framework; designed the three-stage technological evolution framework of "enhancement–generation –collaboration"; completed the literature search, classification analysis, and full manuscript writing; revised and finalized the full text.

References:

[1] Li L, Zhang Y, Liu D, Chen L. Large Language Models for Generative Recommendation: A Survey and Visionary Discussions[A]. Calzolari N, Kan M-Y, Hoste V, et al. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) [C]. Torino, Italia: ELRA and ICCL, 2024: 10146-10159.

[2] Wang W, Lin X, Feng F, He X, Chua T-S. Generative Recommendation: Towards Personalized Multimodal

Content Generation[A]. Proceedings of the ACM Web Conference 2025 (WWW ' 25) [C]. New York: ACM, 2025: 2421–2425.

[3] Zhang X, Xie R, Lyu Y, Xin X, Ren P, Liang M, Zhang B, Kang Z, de Rijke M, Ren Z. Towards Empathetic Conversational Recommender Systems[A]. Proceedings of the 18th ACM Conference on Recommender Systems (RecSys ' 24) [C]. New York: ACM, 2024: 84–93.

[4] Bian Zuming, Zhao Peng, Zhang Bo, et al. A Survey of Recommender Systems Oriented Toward Large Language Models[J]. Computer Science, 2024, 51(11A): 1–10.

[5] Wu Guodong, Qin Hui, Hu Quanxing, et al. Research on Large Language Models and Personalized Recommendation[J]. Journal of Intelligent Systems, 2024, 19(6): 1351-1365.

[6] Xie Guangming, Bai Yanbing, Wu Zi' ang, et al. A Survey of Recommender Systems Based on Large Language Models[J]. Journal of Intelligent Systems, 2025, 20(6): 1520-1533.

[7] Zhang Rui, Bian Zhipeng. A Survey of Multimodal Generation Research for Recommender Systems[J]. Computer Science and Exploration, 2025, 19(12): 3224-3242.

[8] Wang Y, Ma W, Zhang M, Liu Y, Ma S. A Survey on the Fairness of Recommender Systems[J]. ACM Transactions on Information Systems, 2023, 41(3): 1-43.

[9] Yin H, Sun Y, Li X, Zhang W, Liu G. A Comprehensive Survey on Federated Recommendation Systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(8): 1-20.

[10] Li C, Chen H, Liu Y, et al. Deep Learning for Recommendation: A Comprehensive Survey[J]. ACM Computing Surveys, 2024, 56(9): 1-37.

[11] Jannach D, Manzoor A, Cai W, Chen L. A Survey on Conversational Recommender Systems[J]. ACM Computing Surveys, 2022, 54(5): 1-36.

Note: Figure translations are in progress. See original paper for figures.