

Uncertainty-aware Generative Recommendation

Chenxiao Fan
simonfan@mail.ustc.edu.cn
University of Science and Technology
of China
Hefei, China

Chongming Gao*
chongminggao@ustc.edu.cn
University of Science and Technology
of China
Hefei, China

Yaxin Gong
gyx2022@mail.ustc.edu.cn
University of Science and Technology
of China
Hefei, China

Haoyan Liu
liuhaoyan@ustc.edu.cn
University of Science and Technology
of China
Hefei, China

Fuli Feng
fulifeng93@gmail.com
University of Science and Technology
of China
Hefei, China

Xiangnan He*
xiangnanhe@gmail.com
University of Science and Technology
of China
Hefei, China

Abstract

Generative Recommendation has emerged as a transformative paradigm, reformulating recommendation as an end-to-end autoregressive sequence generation task. Despite its promise, existing preference optimization methods typically rely on binary outcome correctness, suffering from a systemic limitation we term *uncertainty blindness*. This issue manifests in the neglect of the model’s intrinsic generation confidence, the variation in sample learning difficulty, and the lack of explicit confidence expression, directly leading to unstable training dynamics and unquantifiable decision risks. In this paper, we propose **Uncertainty-aware Generative Recommendation (UGR)**, a unified framework that leverages uncertainty as a critical signal for adaptive optimization. UGR synergizes three mechanisms: (1) an uncertainty-weighted reward to penalize confident errors; (2) difficulty-aware optimization dynamics to prevent premature convergence; and (3) explicit confidence alignment to empower the model with confidence expression capabilities. Extensive experiments demonstrate that UGR not only yields superior recommendation performance but also fundamentally stabilizes training, preventing the performance degradation often observed in standard methods. Furthermore, the learned confidence enables reliable downstream risk-aware applications. Our code is available at: <https://github.com/cxfann/UGR>.

CCS Concepts

• Information systems → Recommender systems.

Keywords

Generative Recommendation, Large Language Models, Reinforcement Learning, Uncertainty Modeling, Preference Alignment

ACM Reference Format:

Chenxiao Fan, Chongming Gao, Yaxin Gong, Haoyan Liu, Fuli Feng, and Xiangnan He. 2026. Uncertainty-aware Generative Recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and*

*Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License. *KDD 2026, Jeju Island, Republic of Korea.*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
<https://doi.org/10.1145/3770855.3817975>

Data Mining V.2 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.
ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770855.3817975>

Resource Availability:

The source code of this paper has been made publicly available at <https://doi.org/10.5281/zenodo.20287710>.

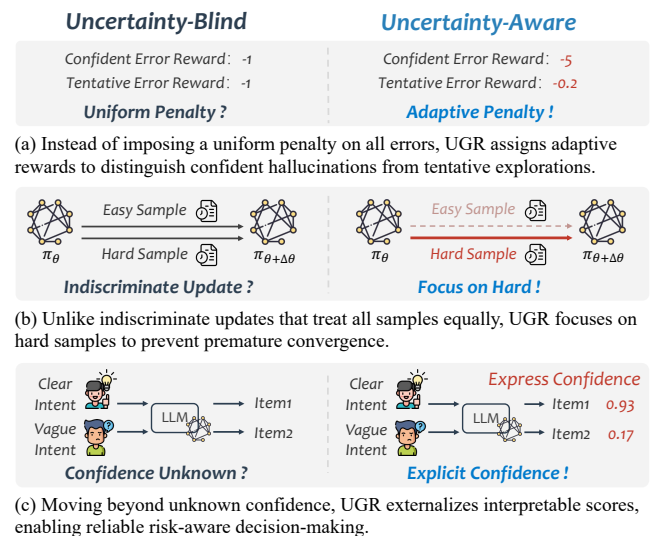


Figure 1: Comparison between existing Uncertainty-Blind methods (Left) and our proposed UGR (Right).

1 Introduction

The field of recommender systems is undergoing a fundamental paradigm shift from discriminative ranking to generative modeling [15, 25, 40]. Diverging from traditional discriminative approaches that rely on multi-stage “retrieval-then-ranking” pipelines to estimate user-item relevance scores [14, 37], Generative Recommendation unifies the task as an end-to-end conditional sequence generation problem [31, 35]. Under this paradigm, items are no longer treated as atomic indices but are serialized into sequences of discrete tokens, typically represented as Semantic IDs (SIDs) [31, 39, 47] or textual identifiers [2, 11, 12]. Consequently,

the underlying logic of recommendation is fundamentally transformed into an autoregressive decoding process, where the model constructs target items by sequentially sampling from a predicted probability distribution over the vocabulary [3, 4, 29].

To align the generative policy with user intent, preference optimization has been widely adopted in generative recommendation [5, 7, 26, 30]. However, existing objectives are typically predicated solely on outcome correctness [22, 35]. This paradigm reduces the generation process to a binary right-or-wrong assessment, overlooking the variations in the model’s internal confidence and sample learning difficulty [13, 18]. We term this systemic oversight *uncertainty blindness* (illustrated in Figure 1), which manifests in three critical limitations: First, by imposing uniform penalties on all errors, the model fails to distinguish the severity of mistakes—specifically, differentiating between confident hallucinations characterized by high probability assignments to incorrect items, and tentative explorations stemming from a flat probability distribution. Second, treating all training instances equally ignores the disparity in learning difficulty, leading to premature convergence on trivial samples while under-fitting hard ones; Finally, the lack of explicit confidence expression prevents the model from quantifying decision risk, forcing it to rely solely on sampling strategies even in high-uncertainty scenarios. Taken together, these issues underscore the necessity of explicitly modeling and leveraging uncertainty within the preference optimization loop.

To this end, rather than overlooking this intrinsic information, we propose to transform uncertainty into a high-value signal guiding model optimization. We propose **Uncertainty-aware Generative Recommendation (UGR)**, a framework designed to explicitly model and leverage generative uncertainty to enhance robustness and trustworthiness. Specifically, the framework incorporates three synergistic mechanisms: First, we introduce an *uncertainty-weighted reward*. Leveraging the model’s intrinsic predictive distribution to distinguish the nature of errors, this mechanism imposes severe penalties on confident hallucinations while preserving the model’s latitude for tentative exploration. Second, we design *difficulty-aware optimization dynamics*. By dynamically sensing the learning difficulty of each instance, we adaptively re-weight the optimization process, thereby preventing premature convergence on trivial patterns and ensuring persistent engagement with hard-to-learn samples. Finally, we propose *explicit confidence alignment*, which empowers the model to articulate confidence scores alongside recommendations, externalizing its internal certainty into quantifiable risk signals to support reliable decision-making. Together, these mechanisms elevate the optimization objective from simple outcome fitting to robust uncertainty-aware learning.

Extensive experiments validate the effectiveness of UGR, revealing several critical and counter-intuitive findings. Notably, we observe that standard preference optimization, lacking uncertainty awareness, often underperforms the Supervised Fine-Tuning (SFT) baseline. This degradation arises because coarse rewards treat all errors equally, introducing noisy signals when the model is uncertain. In contrast, analyzing the training dynamics reveals that UGR stabilizes the learning trajectory: it prevents premature convergence on trivial samples while sustaining effective exploration. Moreover, the aligned confidence serves not only as a regularization mechanism to boost ranking accuracy but also provides reliable signals

for downstream risk-aware applications, such as list truncation and reranking.

The main contributions of this work are summarized as follows:

- We identify the issue of *uncertainty blindness* in current preference optimization. We analyze how neglecting uncertainty undermines training stability and limits the model’s capability.
- We propose **Uncertainty-aware Generative Recommendation (UGR)**, a unified framework that leverages uncertainty for adaptive optimization. Through joint training, UGR simultaneously optimizes recommendation quality and endows the model with explicit confidence expression.
- Extensive experiments demonstrate that UGR yields superior performance and fundamentally stabilizes training dynamics, preventing the degradation observed in standard methods. Furthermore, the learned confidence enables reliable downstream risk-aware applications.

Taken together, our work suggests that in generative recommendation, modeling uncertainty is not merely a calibration task, but an indispensable component of the preference optimization loop itself.

2 Preliminary

In this section, we explicitly formulate the generative recommendation task and clarify the distinction between implicit generation uncertainty and explicit confidence, laying the theoretical groundwork for our proposed method.

2.1 Generative Recommendation Formulation

Let \mathcal{U} and \mathcal{I} denote the sets of users and items, respectively. For a user $u \in \mathcal{U}$, we represent their historical interactions as a time-ordered sequence $H_u = [i_1, i_2, \dots, i_k]$, where each $i_j \in \mathcal{I}$ denotes an interacted item.

In the generative recommendation paradigm, the recommendation task is formulated as a conditional generation problem. Given the user context x derived from H_u , the model aims to generate an output sequence y that represents the recommended item. This process is modeled by a parameterized generative policy $\pi_\theta(y | x)$, which produces the output in an autoregressive manner:

$$\pi_\theta(y | x) = \prod_{j=1}^L P(t_j | x, y_{<j}; \theta), \quad (1)$$

where t_j denotes the j -th generated token and L is the sequence length.

The final recommendation is obtained by decoding from the model’s generative distribution. In practice, each item may be represented by a sequence of discrete tokens from the vocabulary. In this work, we specifically consider SIDs, where the token sequence encodes the categorical hierarchy of the item.

To align the generative policy with user preference, the model is typically optimized to maximize a reward function $R(x, y)$ that evaluates the quality of the generated hypothesis against the ground truth y^* . Standard approaches predominantly employ outcome-based rewards [11]. The most common form is the binary reward, which strictly penalizes any mismatch:

$$R_{\text{bin}}(y, y^*) = \mathbb{I}(y = y^*). \quad (2)$$

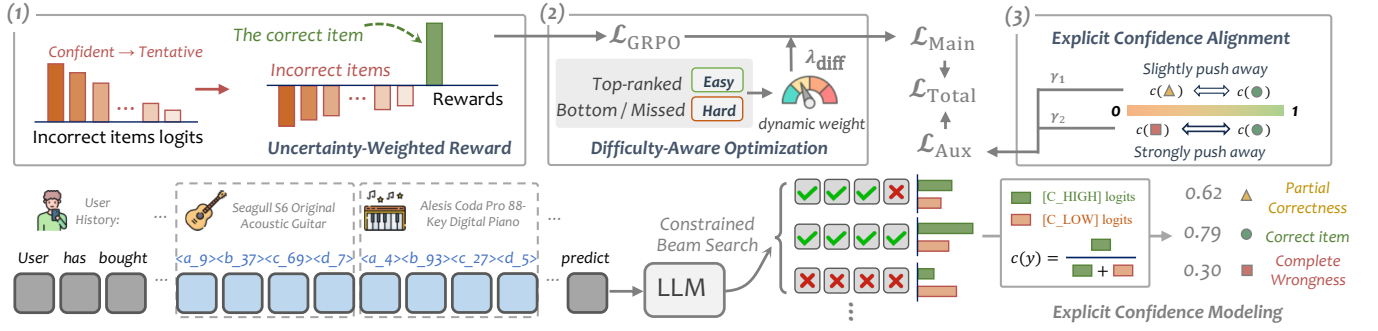


Figure 2: Overview of the proposed UGR framework. Given the user history, the model generates a candidate set via constrained beam search. These candidates effectively drive three synergistic mechanisms: (1) Uncertainty-weighted reward penalizes confident hallucinations based on logit intensity; (2) Difficulty-aware optimization adaptively re-weights the gradient budget based on ranking difficulty; (3) Explicit confidence alignment externalizes internal certainty into quantifiable risk signals.

To address the uniform penalty applied to all incorrect generations by the binary reward, ranking-based rewards have been introduced [6, 22, 35]. This paradigm aims to differentiate error severity by imposing stricter penalties on hard negatives—incorrect items assigned high rankings by the model. Formally, this can be formulated as a rank-aware penalty:

$$R_{\text{rank}}(y, y^*) = \mathbb{I}(y = y^*) - \frac{\mathbb{I}(y \neq y^*)}{\log(\text{rank}(y) + 1)}, \quad (3)$$

where $\text{rank}(y)$ denotes the ranking position of the generated item y within the model’s probability distribution.

However, both identity or rank paradigms treat the reward as a static signal derived solely from discrete outcomes, neglecting the model’s intrinsic generation uncertainty and the varying difficulty of training instances.

2.2 Confidence and Uncertainty in Recommendation

Generative recommendation fundamentally differs from factoid QA or logical reasoning tasks, which typically possess a unique, objectively correct ground truth. In recommendation scenarios, user interaction is inherently subjective and stochastic. Consequently, uncertainty stems not merely from model capability limitations but serves as an intrinsic property characterizing the diversity of user interests and the distribution of potential behaviors [20]. We explicitly distinguish two forms of uncertainty:

2.2.1 Implicit Generation Uncertainty. Mainstream LLMs generate recommendations via autoregressive decoding, where uncertainty is implicitly encoded in the probability distribution over the vocabulary [23]. Formally, for a generated recommendation y under context x , implicit uncertainty $\mathcal{U}_{\text{imp}}(y | x)$ is often characterized by the negative log-likelihood of the generated sequence [43]:

$$\mathcal{U}_{\text{imp}}(y | x) = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log P(y_t | y_{<t}, x). \quad (4)$$

where y_t denotes the token generated at step t , and $y_{<t}$ represents the prefix sequence. A higher \mathcal{U}_{imp} indicates lower prediction probability, reflecting greater uncertainty. Such implicit uncertainty

provides a fine-grained reflection of the model’s internal generation dynamics, capturing relative preference and ambiguity at the token level. It is particularly effective for monitoring generation states (e.g., distinguishing determination from hesitation) and assessing the relative difficulty of different inputs. However, tied to token-level objectives, implicit uncertainty is not explicitly designed to convey decision-level confidence or to directly support risk-aware recommendation. Moreover, due to the normalization of probability distributions, implicit uncertainty limits its absolute comparability across samples, constraining its expressiveness in highly ambiguous scenarios.

2.2.2 Explicit Confidence. Distinct from implicit statistics, explicit confidence $C_{\text{exp}}(y)$ refers to the model’s direct assessment of the overall quality or plausibility of a generated recommendation, typically articulated through dedicated output heads or confidence tokens [8]. In logic-intensive tasks, confidence learning often targets absolute calibration, aiming to align confidence values with empirical accuracy [13]. In recommendation settings, however, such absolute calibration is less meaningful due to the extreme data sparsity and the generally low true click-through rate, which would force calibrated scores towards zero. Instead, the relative ordering of confidence across different recommendations more naturally reflects the model’s comparative belief in alternative decisions. This ranking-based confidence complements implicit uncertainty by providing a decision-level signal aligned with recommendation objectives, facilitating robust preference modeling and risk-aware applications.

3 Method

3.1 Overview

In this section, we introduce the **Uncertainty-aware Generative Recommendation (UGR)**, designed to address the pervasive issue of uncertainty blindness in generative recommendation. Diverging from traditional preference optimization methods that focus solely on outcome correctness, UGR reformulates preference alignment as an uncertainty-aware closed-loop system. As illustrated in Figure 2, the framework comprises three synergistic components.

To ensure that uncertainty signals are statistically measurable and comparable within this loop, UGR is grounded in two structural prerequisites. First, by representing items as SIDs [31], we leverage their fixed-length discrete structure to eliminate length-induced bias, rendering logits directly comparable across sequences while naturally preserving fine-grained partial correctness signals. Second, we employ constrained beam search during rollout to generate a set of distinct candidates confined to the high-likelihood space [35]. This ensures that logit discrepancies reflect the model’s intrinsic hesitation regarding hard negatives, rather than stochastic noise or redundant sampling.

Building upon this solid foundation, UGR injects uncertainty awareness through three core mechanisms:

- **Uncertainty-aware Rollout Reward:** Addressing the limitation of uniform penalties, this mechanism distinguishes the nature of errors during the rollout phase. By analyzing logit intensity, we impose stricter penalties on confident hallucinations, thereby providing denoised, high-quality signals for the subsequent optimization.
- **Difficulty-aware Optimization Dynamics:** Recognizing the variation in learning difficulty, this module dynamically modulates the gradient budget based on instance-level generation difficulty. This prevents overfitting on trivial patterns while ensuring sustained and effective exploration for hard-to-learn samples.
- **Explicit Confidence Alignment:** Leveraging the hierarchical structure of SIDs, this component aligns the model’s explicit confidence with generation quality (i.e., complete vs. partial correctness). This endows the model with interpretable risk-aware capabilities, transforming internal uncertainty into actionable decision signals.

3.2 Uncertainty-aware Rollout Reward

Existing reward mechanisms are limited by their coarse-grained characterization of error severity. While basic binary rewards impose uniform penalties, improved ranking-based rewards, despite incorporating ordinal information, remain constrained by a rigid penalty scale. This discrete ranking signal obscures the model’s intrinsic probability landscape: it fails to differentiate between errors stemming from highly confident hallucinations versus low-confidence hesitations, assigning identical penalties solely based on rank positions.

To break the bottleneck, we propose a transition from position-aware to distribution-aware optimization, leveraging the logit intensity of rollouts to construct a continuous uncertainty-aware reward. Specifically, given a context x , we employ constrained beam search to sample a candidate set $\mathcal{G} = \{y_1, \dots, y_G\}$. We define the cumulative logit intensity of a sequence as $s(y) = \sum_t \log P(y_t | y_{<t}, x)$, and design a reward $r(y_i)$ based on $s(y)$ to adaptively penalize the set of incorrect generations $\mathcal{N} = \mathcal{G} \setminus \{y^*\}$, where y^* denotes the ground-truth item sequence. The proposed reward $r(y_i)$ is as below:

$$r(y_i) = \begin{cases} 1.0, & \text{if } y_i = y^*, \\ -\frac{\exp(s(y_i))}{\sum_{y_j \in \mathcal{N}} \exp(s(y_j))}, & \text{if } y_i \in \mathcal{N}, \end{cases} \quad (5)$$

This design augments ranking-based comparison with continuous modeling of error intensity. Intuitively, the higher the model’s

certainty in a specific error, the larger the magnitude of the corresponding negative reward. Finally, adhering to the Group Relative Policy Optimization (GRPO) framework [33], we normalize these rewards within each sampled group to compute the relative advantage:

$$\hat{A}_{i,t} = \frac{r(y_i) - \mu_r}{\sigma_r + \epsilon}, \quad (6)$$

where μ_r and σ_r denote the mean and standard deviation of rewards $\{r(y_j)\}_{j=1}^G$ in the current group, and ϵ is a small constant for numerical stability. This normalized advantage $\hat{A}_{i,t}$ serves as the input for the subsequent optimization objective.

3.3 Difficulty-aware Optimization Dynamics

While the uncertainty-aware rollout reward addresses intra-sample error severity, generative recommendation still suffers from significant inter-sample imbalance. User contexts vary drastically in inference difficulty, yet standard training paradigms typically optimize a uniform objective across all instances. This approach inherently overlooks the disparity in learning difficulty [28], often causing the model to overfit trivial patterns while under-exploring hard samples that contain rich structural information. In generative recommendation, learning difficulty directly reflects generation uncertainty, since a sample is difficult precisely when the model cannot confidently locate the correct item.

To address this, we propose difficulty-aware optimization dynamics, which incorporates an adaptive re-weighting mechanism to shift the learning focus toward harder samples. Since sample difficulty lacks explicit supervision, we adopt the ground truth’s ranking position within the beam search candidates as a lightweight proxy, which effectively reflects the model’s current capability to locate the target item.

Specifically, for a context x and ground-truth y^* , let $\text{rank}(y^*)$ denote the position of y^* within the candidate set \mathcal{G} sorted by logits (0-indexed). If y^* is missed by the beam search ($y^* \notin \mathcal{G}$), indicating a high-uncertainty region, we assign the maximum difficulty $\text{rank}(y^*) = G - 1$. The dynamic weight $\lambda_{\text{diff}}(x)$ is constructed as:

$$\lambda_{\text{diff}}(x) = \alpha + (1 - \alpha) \cdot \frac{\text{rank}(y^*)}{G - 1}, \quad (7)$$

where $\alpha \in [0, 1]$ serves as a hyperparameter to ensure basic learning stability for easy samples (i.e., the lower bound of λ_{diff} when $\text{rank}(y^*) = 0$).

We integrate this adaptive weighting mechanism into the GRPO framework. Standard GRPO optimizes the policy π_θ by maximizing the expected advantage of sampled generations. Based on the normalized advantage $\hat{A}_{i,t}$ derived in Eq. (6), the unweighted objective for a single context x is formulated as:

$$\mathcal{L}_{\text{GRPO}}(x; \theta) = -\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \frac{\pi_\theta(y_{i,t} | x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} | x, y_{i,<t})} \hat{A}_{i,t}, \quad (8)$$

In practice, this objective is typically augmented with clipping or KL-regularization constraints to stabilize updates; we present the simplified form here to highlight the core optimization logic.

Finally, we propose the difficulty-aware instance loss by modulating the standard objective with the derived difficulty weight:

$$\mathcal{L}_{\text{Main}}(x; \theta) = \lambda_{\text{diff}}(x) \cdot \mathcal{L}_{\text{GRPO}}(x; \theta). \quad (9)$$

This formulation effectively propagates uncertainty awareness from the reward level (via $\hat{A}_{i,t}$) to the optimization dynamics level (via λ_{diff}). It allows the model to automatically balance the consolidation of simple patterns with the aggressive exploration of hard samples, laying a robust foundation for the subsequent confidence alignment task.

3.4 Explicit Confidence Modeling and Alignment

While the aforementioned modules effectively optimize training dynamics, the utilized uncertainty remains implicit and inaccessible for inference. To bridge this, we propose explicit confidence alignment. This module serves a dual purpose: it externalizes intrinsic certainty into interpretable signals for risk-aware decision-making, and acts as semantic regularization to reinforce the model’s comprehension of the hierarchical SID structure.

Verbalized Confidence Tokenization. We expand the vocabulary with a dedicated set of tokens: $V_{\text{conf}} = \{[\text{C_HIGH}], [\text{C_LOW}]\}$. Upon generating a recommendation sequence y given context x , we perform a lightweight forward pass using an “Append-and-Query” strategy. The explicit confidence score $c(y)$ is defined as the normalized probability of the high-confidence token:

$$c(y) = \frac{P([\text{C_HIGH}] | x, y)}{P([\text{C_HIGH}] | x, y) + P([\text{C_LOW}] | x, y)}. \quad (10)$$

This formulation renders the confidence score a fully differentiable internal representation that updates synchronously with the generation policy.

Hierarchical Error Taxonomy. To provide high-quality alignment signals, we leverage the inherent tree-structured semantics of SIDs to structure the negative sample space into two distinct types:

- **Partial Correctness** (N_{partial}): The generated sequence y shares a non-empty prefix with the ground truth y^* but deviates at the fine-grained item level. These samples typically correspond to hard negatives within the high-likelihood space.
- **Complete Wrongness** (N_{wrong}): The sequence y mismatches y^* at the root level, indicating a significant semantic drift.

Hierarchical Confidence Alignment. Based on this taxonomy, we construct a multi-granular pair-wise ranking loss to map the semantic hierarchy onto the confidence space:

$$\mathcal{L}_{\text{Aux}} = \gamma_1 \cdot \mathcal{L}_{\text{rank}}(y^*, N_{\text{partial}}) + \gamma_2 \cdot \mathcal{L}_{\text{rank}}(y^*, N_{\text{wrong}}), \quad (11)$$

where $\mathcal{L}_{\text{rank}}$ adopts the logistic margin form:

$$\mathcal{L}_{\text{rank}}(y^*, S) = \frac{1}{|S|} \sum_{y \in S} \log(1 + \exp(c(y) - c(y^*))). \quad (12)$$

By setting $\gamma_2 > \gamma_1$, we impose a stronger gradient penalty on complete errors, forcing their margin from the ground truth in the probability space to be significantly larger than that of partially correct samples. This mechanism implicitly induces a confidence hierarchy, $c(y^*) > c(y_{\text{partial}}) > c(y_{\text{wrong}})$, thereby endowing the model with the ability to distinguish among different types of negative samples.

3.5 Overall Objective and Training

We model the preference alignment of Generative Recommendation as a unified multi-task optimization framework. Given that both tasks share the same context inputs and generation process, for a batch \mathcal{B} in a single iteration, we define the total training objective as the direct summation of the two losses:

$$\mathcal{L}_{\text{Total}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} (\mathcal{L}_{\text{Main}}(x; \theta) + \mathcal{L}_{\text{Aux}}(x; \theta)). \quad (13)$$

Here, $\mathcal{L}_{\text{Main}}(x; \theta)$ refers to the difficulty-aware GRPO loss (Eq. (9)), optimizing generation correctness and diversity; $\mathcal{L}_{\text{Aux}}(x; \theta)$ corresponds to the confidence ranking loss derived from the same rollout samples (Eq. (11)). Since the auxiliary loss is intrinsically scale-compatible with the normalized advantage-based main loss, and internally balanced by γ_1, γ_2 , we do not introduce an additional global hyperparameter, thereby maintaining the simplicity and robustness of the optimization framework.

To efficiently optimize this objective, UGR adopts an end-to-end training strategy. In each training step, the workflow strictly follows three steps:

- **Step 1 (Joint Rollout & Evaluation):** The model executes constrained beam search under the current policy π_θ to generate a candidate set \mathcal{G} . Crucially, this single set is simultaneously utilized to compute the reward for the main task and to construct hierarchical negative pairs for the auxiliary task, ensuring consistent variance across objectives.
- **Step 2 (Dynamic Weighting):** Based on the relative ranking position of the ground truth y^* within \mathcal{G} , we compute the instance-level dynamic weight $\lambda_{\text{diff}}(x)$ to adaptively regulate the gradient contribution of $\mathcal{L}_{\text{Main}}$.
- **Step 3 (Unified Backpropagation):** A forward pass is performed on Eq. (13) to output both the likelihood of recommendation tokens and the distribution of confidence tokens. Finally, gradients from $\mathcal{L}_{\text{Main}}$ and \mathcal{L}_{Aux} are aggregated to update the shared parameters θ .

Through this joint optimization mechanism, UGR ensures that confidence modeling is deeply embedded in the policy learning dynamics, realizing the synergistic evolution of generative capability and risk awareness.

4 Experiments

In this section, we conduct experiments to answer the following research questions (RQs):

- **RQ1:** How does UGR perform compared to existing baselines in generative recommendation tasks?
- **RQ2:** How does each uncertainty-aware component contribute to the overall effectiveness of the proposed framework?
- **RQ3:** Can UGR effectively stabilize the training process and mitigate the issue of uncertainty blindness?
- **RQ4:** Does the externalized confidence signal serve as a reliable indicator for downstream risk-aware applications?

Table 1: Overall performance comparison on three datasets. The baselines are categorized into traditional models (SASRec), LLM-based models (covering text-based methods: BIGRec, SPRec; and SID-based generative methods: TIGER, MiniOneRec), and reasoning-enhanced models (ReaRec, R²ec). The best results are highlighted in bold, and the second-best results are underlined.

Method	Office					Industrial					Yelp				
	HR@1	HR@5	HR@10	NG@5	NG@10	HR@1	HR@5	HR@10	NG@5	NG@10	HR@1	HR@5	HR@10	NG@5	NG@10
SASRec	0.0176	0.0771	0.1028	0.0489	0.0572	0.0276	0.0732	0.0894	0.0498	0.0551	0.0059	0.0253	0.0441	0.0153	0.0213
BIGRec	0.0714	0.1053	0.1198	0.0898	0.0944	0.0272	0.0907	0.1143	0.0635	0.0710	0.0036	0.0089	0.0119	0.0063	0.0073
SPRec	0.0665	0.0885	0.1013	0.0787	0.0828	0.0305	0.0970	0.1244	0.0689	0.0780	0.0066	0.0096	0.0113	0.0081	0.0087
TIGER	0.0715	0.1146	0.1361	0.0946	0.1015	0.0489	0.0954	0.1217	0.0744	0.0828	0.0057	0.0283	0.0460	<u>0.0167</u>	0.0223
MiniOneRec	<u>0.0823</u>	<u>0.1314</u>	<u>0.1479</u>	<u>0.1084</u>	<u>0.1137</u>	<u>0.0745</u>	<u>0.1080</u>	<u>0.1276</u>	<u>0.0917</u>	<u>0.0980</u>	<u>0.0087</u>	0.0243	0.0420	0.0165	0.0226
ReaRec	0.0782	0.1164	0.1380	0.0990	0.1060	0.0529	0.0952	0.1148	0.0760	0.0823	0.0074	0.0258	<u>0.0470</u>	0.0163	<u>0.0232</u>
R ² ec	0.0496	0.1006	0.1364	0.0758	0.0873	0.0372	0.0886	0.1274	0.0635	0.0761	0.0066	0.0253	0.0464	0.0161	0.0227
UGR	0.0904	0.1407	0.1598	0.1169	0.1231	0.0747	0.1179	0.1393	0.0965	0.1033	0.0096	<u>0.0281</u>	0.0473	0.0188	0.0250

Table 2: Ablation study of UGR on three datasets. (a) is the base model with SFT only. (b1)-(b3) denote variants using GRPO with different reward functions: 0/1 Reward, Rank-based Reward, and Uncertainty-Weighted Reward (UW-Reward). (c) adds Difficulty-aware Optimization upon (b3), and (d) is the full UGR model further incorporating Explicit Confidence Alignment. Bold indicates the best performance.

Method / Variant	Office				Industrial				Yelp			
	HR@5	HR@10	NG@5	NG@10	HR@5	HR@10	NG@5	NG@10	HR@5	HR@10	NG@5	NG@10
(a) Only SFT	0.1374	0.1563	0.1152	0.1213	0.1104	0.1378	0.0923	0.1010	0.0264	0.0423	0.0165	0.0217
(b1) + GRPO (0/1 Reward)	0.1283	0.1485	0.1056	0.1120	0.1041	0.1303	0.0845	0.0931	0.0249	0.0405	0.0170	0.0220
(b2) + GRPO (Rank Reward)	0.1291	0.1468	0.1079	0.1135	0.1076	0.1321	0.0878	0.0958	0.0232	0.0423	0.0157	0.0219
(b3) + GRPO (UW-Reward)	0.1349	0.1522	0.1137	0.1193	0.1144	0.1338	0.0967	0.1029	0.0234	0.0424	0.0160	0.0222
(c) + Difficulty-Aware Opt.	0.1378	0.1579	0.1158	0.1223	0.1151	0.1349	0.0967	0.1030	0.0264	0.0454	0.0178	0.0239
(d) UGR (Full)	0.1407	0.1598	0.1169	0.1231	0.1179	0.1393	0.0965	0.1033	0.0281	0.0473	0.0188	0.0250

4.1 Experimental Setup

4.1.1 Datasets. We conduct experiments on three real-world benchmarks: Office Products (Office) and Industrial and Scientific (Industrial) from the Amazon Review Dataset¹, along with the Yelp dataset². Following previous works [22] we partition the data into training, validation, and test sets with an 8:1:1 ratio, strictly adhering to the chronological order of interactions. Detailed preprocessing steps and dataset statistics are provided in Appendix A.

4.1.2 Evaluation Setting and Metrics. We adopt two widely used metrics to evaluate top- K recommendation performance: Hit Ratio (HR@ K) and Normalized Discounted Cumulative Gain (NG@ K), with $K \in \{1, 5, 10\}$. During the inference phase, we employ beam search with a beam size of 10 to generate the ranking list.

4.1.3 Baselines. We compare UGR against a comprehensive set of baselines categorized into three groups: (1) Traditional Models, represented by SASRec [19]; (2) LLM-based Models, which are further divided into two streams: text-based methods (BIGRec [2], SPRec [11]) that rely on grounding mechanisms, and SID-based generative methods (TIGER [31], MiniOneRec [22]) that utilize

SID for direct generation; and (3) Reasoning-enhanced Models, including ReaRec [36] and R²ec [45]. Further details are provided in Appendix B.

4.1.4 Implementations. All experiments are conducted on 8 NVIDIA A100 GPUs. We utilize Qwen3-8B [44] as the backbone LLM, representing items via 4-layer SIDs. The training pipeline consists of two stages: First, we perform SFT for up to 10 epochs with early stopping. Second, initialized from the SFT checkpoint, we conduct the proposed preference optimization as described in Section 3.5 for 2 epochs. Crucially, during the training rollout phase, we utilize beam search with a group size of $G = 16$ to construct candidate sets for reward calculation and alignment. Further implementation details are available in Appendix C.

4.2 Overall Performance Comparison (RQ1)

In this section, we present the comparative results of UGR against all baselines across three datasets. The performance metrics are summarized in Table 1.

The results reveal four key observations: (1) UGR consistently achieves the best performance across all datasets and metrics, establishing a new state-of-the-art. The significant margin over the

¹<https://jmcauley.ucsd.edu/data/amazon>

²<https://www.yelp.com/dataset>

strongest baselines validates the robustness of our uncertainty-aware framework. (2) SID-based generative methods (TIGER, MiniOneRec, UGR) generally surpass traditional (SASRec) and text-based (BIGRec) models, confirming that combining LLM semantics with structured representations captures complex preferences better than shallow embeddings or open-ended generation. (3) Crucially, UGR significantly outperforms MiniOneRec. Given their shared backbone, this gain is strictly attributed to our resolution of uncertainty blindness, where weighted rewards and difficulty-aware dynamics provide superior optimization signals over binary feedback. (4) UGR also outperforms reasoning-enhanced models (ReaRec, R²ec), indicating that explicit uncertainty modeling offers a more direct and robust signal for ranking refinement.

4.3 Component Analysis and Ablation (RQ2)

To evaluate the contribution of each component in UGR, we conduct a detailed ablation study in Table 2, incrementally introducing different reward functions and optimization dynamics starting from the SFT baseline.

4.3.1 Impact of Uncertainty-Weighted Reward. First, we observe a critical phenomenon: standard preference optimization (Row b1-b2) fails to surpass the SFT baseline. This confirms our hypothesis of uncertainty blindness—coarse binary or ranking rewards introduce noisy signals by treating confident and tentative errors equally. In contrast, introducing the uncertainty-weighted reward (Row b3) significantly recovers performance, effectively matching SFT. This validates that distinguishing error severity based on generation confidence is a prerequisite for effective optimization. However, reward-level improvements alone are insufficient to significantly surpass SFT, necessitating advanced optimization dynamics.

4.3.2 Effectiveness of Difficulty-aware Optimization. Building upon (b3), incorporating difficulty-aware optimization (Row c) yields consistent gains across datasets. By dynamically reweighting instances based on difficulty and uncertainty, this strategy prevents overfitting to easy samples and sustains exploration, and its specific impact on training stability will be further analyzed in Sec. 4.4.

4.3.3 Benefit of Explicit Confidence Alignment. Finally, the full UGR framework (Row d) achieves the best performance by jointly optimizing with explicit confidence alignment. Beyond enabling confidence expression, the auxiliary ranking loss acts as a semantic regularizer, reinforcing the model’s structural understanding of SIDs to boost ranking accuracy by enforcing self-assessment. This module also enables the risk-aware applications explored in Sec. 4.5.

Beyond the incremental ablation above, we further investigate the mutual synergy among the three modules through a complete decomposition over all seven combinations, validate our binary confidence token design against alternatives, and compare with calibration-based uncertainty methods (e.g., RLCR [8]). All analyses are deferred to Appendix D.

4.4 Training Dynamics Analysis (RQ3)

To investigate the impact of difficulty-aware optimization on training stability and exploration capability, we visualize the training trajectories of HR@1 (ranking accuracy) and HR@16 (candidate

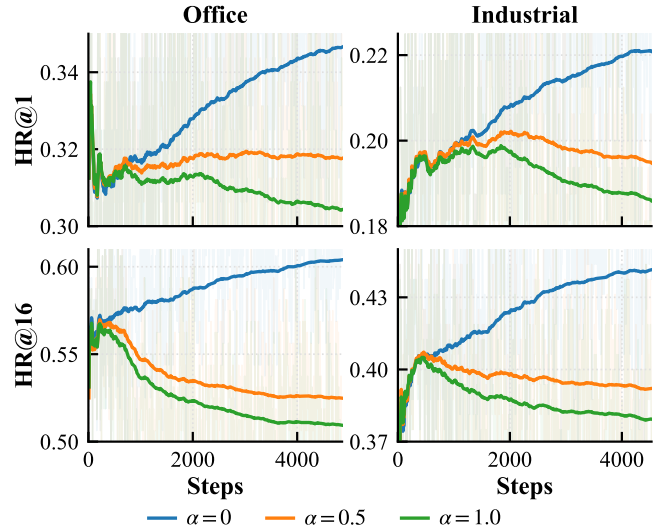


Figure 3: Training dynamics of HR@1 and HR@16 on Office and Industrial datasets under varying difficulty coefficients α . Standard uniform weighting ($\alpha = 1.0$) leads to performance collapse due to overfitting, whereas UGR’s difficulty-aware optimization ($\alpha = 0$) prevents degradation and enables sustained learning.

coverage, aligned with the training beam width $G = 16$) under varying difficulty coefficients α in Figure 3.

4.4.1 Instability of Uniform Optimization. Standard uniform weighting ($\alpha = 1.0$) exhibits severe instability. As shown, it suffers from a distinct pattern of early convergence followed by regression. After an initial peak, both HR@1 and HR@16 degrade significantly. This indicates that the model overfits to trivial samples, causing the effective search space to shrink prematurely (dropping HR@16) and hampering the retrieval of hard items.

4.4.2 Sustained Exploration via Difficulty Awareness. In contrast, UGR ($\alpha = 0$) stabilizes training, maintaining a robust, monotonic upward trend. By suppressing gradients from easy samples, the optimization is compelled to persist in refining hard instances. This mechanism ensures sustained candidate coverage (High HR@16), which effectively translates diversity into precise top-tier recommendations (High HR@1), extending the effective learning horizon.

4.4.3 Analysis of Candidate Collapse. Figure 4 exposes the underlying failure mode of standard preference optimization, which we term *candidate collapse*. The uniform variant ($\alpha=1.0$) exhibits a steady erosion in the number of unique valid items retrieved per rollout. This degradation stems from an *overconfidence trap* on easy samples: the uniform objective forces the model to push the probability of trivial ground-truth tokens toward the extreme, yielding an excessively peaked distribution. As a result, beam search is left with progressively fewer alternative paths to explore, and the pool of valid candidates available for contrastive learning steadily shrinks.

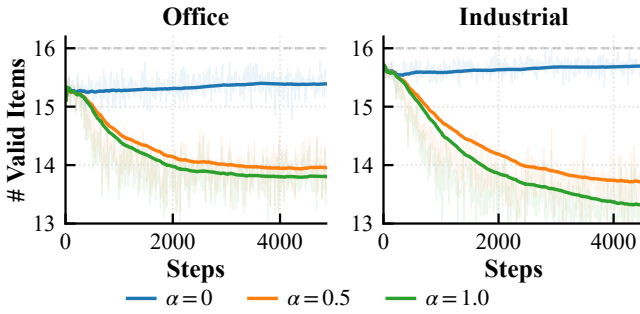


Figure 4: Average number of unique valid items per rollout (beam size $G=16$). As the standard variant ($\alpha=1.0$) becomes overconfident on easy samples, candidate diversity gradually erodes, whereas UGR ($\alpha=0$) sustains a near-saturated search space.

Table 3: Performance comparison of Confidence-based Re-ranking on the Office dataset.

Method	NDCG@1	NDCG@3	NDCG@5	NDCG@10
UGR (Original Rank)	0.0904	0.1102	0.1169	0.1231
UGR + Conf. Re-rank	0.0933	0.1121	0.1185	0.1245
Gain (Δ)	+0.0029	+0.0018	+0.0016	+0.0014

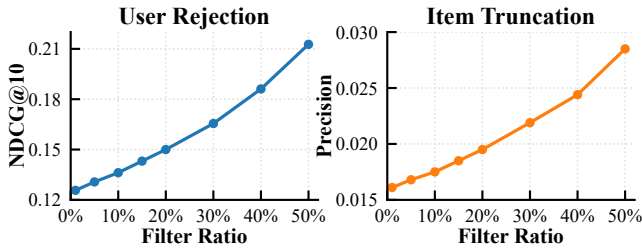


Figure 5: Performance of risk-aware applications on the Office dataset. User-level Rejection (Left): Filtering low-confidence requests steadily improves NDCG@10. Item-level Truncation (Right): Dynamically removing low-confidence tail items effectively enhances Precision.

In contrast, UGR ($\alpha=0$) resolves this pathology by dynamically down-weighting the gradients of easy samples, which directly suppresses such overconfidence and prevents the model from overfitting to trivial patterns. By maintaining a reasonable rather than degenerate distribution, UGR preserves a healthy search space, ensuring that beam search consistently retrieves a diverse set of hard candidates to fuel stable optimization.

4.5 Utility of Explicit Confidence Signals (RQ4)

Having established the training benefits in Section 4.3, we now focus on the practical utility of explicit confidence signals during the inference phase. Using the Office dataset, we evaluate their reliability as a decision basis across two dimensions: ranking enhancement and risk-aware recommendation.

Table 4: Sensitivity of margin hyperparameters γ_1 and γ_2 .

γ_2	γ_1	Office		Industrial		Yelp	
		HR@10	NG@10	HR@10	NG@10	HR@10	NG@10
0.5	0.0	0.1587	0.1227	0.1389	0.1032	0.0473	0.0250
	0.1	0.1598	0.1231	0.1393	0.1033	0.0457	0.0246
	0.5	0.1579	0.1221	0.1382	0.1030	0.0464	0.0243
1.0	0.0	0.1595	0.1226	0.1386	0.1034	0.0469	0.0249
	0.1	0.1600	0.1227	0.1393	0.1036	0.0466	0.0247
	0.5	0.1581	0.1217	0.1373	0.1029	0.0465	0.0244

4.5.1 Confidence-based Ranking Enhancement. First, we investigate the impact of incorporating confidence signals to re-rank candidates generated by beam search. As shown in Table 3, this integration yields consistent improvements across all metrics compared to using logits alone. This suggests that explicit confidence provides an additional assessment of logical correctness complementary to generation probability, effectively correcting high-confidence but factually incorrect hallucinations.

4.5.2 Risk-aware Recommendation. Beyond ranking accuracy, reliable recommendation requires effective risk control. We design two scenarios to verify the risk perception capability of confidence signals at different granularities (Figure 5):

User-level Rejection. Existing systems typically enforce mandatory output for all user requests. However, for hard samples where the model is extremely uncertain, such forced recommendations often result in a poor user experience. Figure 5 (Left) shows that filtering requests based on confidence thresholds leads to a monotonic increase in NDCG@10. This validates that low confidence accurately identifies hard-to-model samples, enabling selective recommendation—rejecting risky queries to ensure system reliability.

Item-level Truncation. Simultaneously, at the item level, traditional Top- K strategies enforce a fixed-length list (e.g., 10 items), which inevitably introduces noise or fillers at the tail. Figure 5 (Right) shows that when we dynamically truncate tail items based on confidence scores, the Precision of the recommendation list increases steadily. This indicates that the confidence signal effectively identifies invalid items, supporting a shift from fixed-length to dynamic-length recommendation. This paradigm reduces invalid exposure to users while maintaining high accuracy.

4.6 Hyperparameter Sensitivity

Table 4 examines the sensitivity of the margin hyperparameters γ_1 (partial correctness) and γ_2 (complete wrongness) in Eq. (11).

Two observations emerge. First, all configurations yield comparable performance across datasets, indicating that UGR is not sensitive to the specific choice of these margins. Second, setting γ_1 to a small positive value (e.g., 0.1) while keeping γ_2 larger consistently achieves the best or near-best results. In contrast, when γ_1 approaches γ_2 (e.g., both at 0.5), performance degrades slightly, as the reward can no longer differentiate partially correct outputs from completely wrong ones. This validates our hierarchical margin design: assigning distinct penalty levels to different error severities provides a more informative learning signal than a uniform penalty.

5 Related Work

In this section, we review the literature relevant to our work from three perspectives: generative recommendation paradigms, preference alignment strategies, and uncertainty modeling in LLMs.

5.1 Generative Recommendation

Propelled by the remarkable capabilities of LLMs, recommender systems are undergoing a fundamental paradigm shift from discriminative ranking to generative modeling [12, 42]. Early explorations primarily leveraged the general text generation capabilities of LLMs to directly verbalize item titles [4, 41, 46]. However, to address the validity challenges associated with unstructured generation, existing methods often resort to external grounding mechanisms [2, 11] or restrict the generation space to input candidate lists [26], which fundamentally constrains the end-to-end potential of generative recommendation. With the advent of works such as TIGER [31] and LC-Rec [47], generative frameworks based on SIDs have gained prominence. These approaches encode items into fixed-length discrete token sequences with hierarchical structures, and further enhance representational semantic density by incorporating collaborative information [39]. Nevertheless, the current generative paradigm remains limited to simple item index prediction—merely sampling item IDs based on probability distributions—while lacking the capability to explicitly articulate confidence levels simultaneously with the recommendation output. To bridge this gap, our proposed UGR endows the model with the ability to articulate confidence alongside recommendation generation, thereby enabling reliable risk-aware decision-making.

5.2 Preference Alignment in Recommendation

Solely relying on SFT is insufficient to fully align generative models with complex recommendation objectives. To bridge the gap between next-token prediction and ranking optimization, preference alignment techniques have been widely adopted [1, 6, 21]. Given the prohibitive computational costs of online reinforcement learning methods like PPO [32], recent research has pivoted towards more efficient offline or group-wise optimization paradigms [9, 27]. Specifically, DPO [30] and its variants [11, 26] are extensively used for pairwise preference learning. Meanwhile, GRPO [33], distinguished by its critic-free design, has been employed by works such as ReRe [35] and MiniOneRec [22] to perform list-wise optimization on high-quality candidates sampled via beam search. Furthermore, FLAME attempts to transcend binary reward limitations by introducing step-wise reward shaping [10]. However, these approaches remain fundamentally constrained by an outcome-correctness perspective. Relying solely on label correctness, they treat confident hallucinations and tentative explorations indistinguishably and ignore the disparity in learning difficulty across samples, falling into the trap of uncertainty blindness. In contrast, UGR transforms uncertainty into high-value feedback signals, achieving adaptive optimization based on error nature and sample difficulty.

5.3 Confidence and Uncertainty in LLMs and Recommendation

Uncertainty estimation has become a pivotal research topic in the general LLM community [17]. Existing studies indicate that

LLMs often exhibit significant confidence bias in complex tasks like code generation, lacking accurate self-assessment capabilities [18, 34, 38]. To address this, works like RLCR introduce reinforcement learning with calibration rewards to align confidence with correctness, demonstrating that high-quality uncertainty signals can reciprocally enhance reasoning performance [8]. However, these approaches predominantly pursue absolute calibration [13], aiming to strictly align predicted probabilities with objective accuracy. As discussed in Sec. 2.2, the inherent subjectivity and sparsity of recommendation tasks render such absolute numerical values less meaningful. While some works analyze uncertainty in recommendation [23], its effective utilization in the generative paradigm remains unexplored. Our work fills this gap by leveraging uncertainty signals to enhance the robustness of preference alignment, while simultaneously endowing the model with explicit risk-awareness capabilities.

6 Conclusion

In this paper, we identified *uncertainty blindness* as a critical bottleneck in generative recommendation. To bridge this gap, we proposed UGR, a unified framework that synergizes uncertainty-weighted rewards, difficulty-aware optimization, and explicit confidence alignment. Our extensive experiments demonstrate that UGR not only establishes new state-of-the-art results but also effectively resolves the training instability issue common in standard methods. Moreover, the learned confidence scores empower the model with reliable self-assessment capabilities for risk-aware decision-making. While validated on three benchmarks of moderate scale, UGR operates on optimization dynamics and reward shaping rather than dataset-specific features, and we expect its principles to generalize to larger-scale settings.

Looking ahead, this work opens up several promising directions. First, while validated within the GRPO framework, our core insight of uncertainty-aware optimization is generalizable. Future work will explore adapting this paradigm to other preference alignment strategies, such as DPO, to further improve sample efficiency. Second, we aim to explore more fine-grained uncertainty decomposition mechanisms to provide richer interpretability for decision-making. Finally, we plan to extend UGR to online interactive settings and dynamic domains, where rapidly shifting user interests amplify the difficulty variance that UGR is designed to handle. We seek to build a truly adaptive and risk-aware recommendation system.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U24B20180, 62525211).

References

- [1] Zhuoxi Bai, Ning Wu, Fengyu Cai, Xinyi Zhu, and Yun Xiong. 2024. Aligning large language model with direct multi-preference optimization for recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 76–86.
- [2] Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. 2025. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems* 3, 4 (2025), 1–27.
- [3] Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. 2024. Decoding matters: Addressing amplification bias and homogeneity issue in recommendations for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10540–10552.
- [4] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*. 1007–1014.
- [5] Shihao Cai, Chongming Gao, Haoyan Liu, Wentao Shi, Jianshan Sun, Ruiming Tang, and Fuli Feng. 2025. Mgrfrec: Towards reinforced reasoning recommendation with multiple groundings and feedback. *arXiv preprint arXiv:2510.22888* (2025).
- [6] Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On softmax direct preference optimization for recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 27463–27489.
- [7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).
- [8] Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806* (2025).
- [9] Hongxun Ding, Keqin Bao, Jizhi Zhang, Yi Fang, Wenxin Xu, Fuli Feng, and Xiangnan He. 2026. Towards Sample-Efficient and Stable Reinforcement Learning for LLM-based Recommendation. *arXiv preprint arXiv:2602.00632* (2026).
- [10] Chenxiao Fan, Chongming Gao, Wentao Shi, Yaxin Gong, Zihao Zhao, and Fuli Feng. 2025. Fine-grained List-wise Alignment for Generative Medication Recommendation. *arXiv preprint arXiv:2505.20218* (2025).
- [11] Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. 2025. Sprec: Self-play to debias llm-based recommendation. In *Proceedings of the ACM on Web Conference 2025*. 5075–5084.
- [12] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*. 299–315.
- [13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.
- [14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.
- [15] Yupeng Hou, An Zhang, Leheng Sheng, Zhengyi Yang, Xiang Wang, Tat-Seng Chua, and Julian McAuley. 2025. Generative Recommendation Models: Progress and Directions. In *Companion Proceedings of the ACM on Web Conference 2025*. 13–16.
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.
- [17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.
- [18] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).
- [19] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [20] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).
- [21] Xiaoyu Kong, Junguang Jiang, Bin Liu, Ziru Xu, Han Zhu, Jian Xu, Bo Zheng, Jiancan Wu, and Xiang Wang. 2025. Think before Recommendation: Autonomous Reasoning-enhanced Recommender. *arXiv preprint arXiv:2510.23077* (2025).
- [22] Xiaoyu Kong, Leheng Sheng, Junfei Tan, Yuxin Chen, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. 2025. Minionerec: An open-source framework for scaling generative recommendation. *arXiv preprint arXiv:2510.24431* (2025).
- [23] Wonbin Kweon, Sanghwan Jang, SeongKu Kang, and Hwanjo Yu. 2025. Uncertainty Quantification and Decomposition for LLM-based Recommendation. In *Proceedings of the ACM on Web Conference 2025*. 4889–4901.
- [24] Doyup Lee, Chihyeon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11523–11532.
- [25] Xiaopeng Li, Bo Chen, Junda She, Shiteng Cao, You Wang, Qinlin Jia, Haiying He, Zheli Zhou, Zhao Liu, Ji Liu, et al. 2025. A survey of generative recommendation from a tri-decoupled perspective: Tokenization, architecture, and optimization. (2025).
- [26] Jiayi Liao, Xiangnan He, Ruobing Xie, Jiancan Wu, Yancheng Yuan, Xingwu Sun, Zhanhui Kang, and Xiang Wang. 2024. Rosepo: Aligning llm-based recommenders with human values. *arXiv preprint arXiv:2410.12519* (2024).
- [27] Jiacheng Lin, Tian Wang, and Kun Qian. 2025. Rec-r1: Bridging generative large language models and user-centric recommendation systems via reinforcement learning. *arXiv preprint arXiv:2503.24289* (2025).
- [28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [29] Xinyu Lin, Haihan Shi, Wenjie Wang, Fuli Feng, Qifan Wang, See-Kiong Ng, and Tat-Seng Chua. 2025. Order-agnostic identifier for large language model-based generative recommendation. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval*. 1923–1933.
- [30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.
- [31] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 10299–10315.
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [33] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).
- [34] Claudio Spiess, David Gros, Kunal Suresh Pai, Michael Pradel, Md Rafiqul Islam Rabin, Amin Alipour, Susmit Jha, Prem Devanbu, and Toufique Ahmed. 2024. Calibration and correctness of language models for code. *arXiv preprint arXiv:2402.02047* (2024).
- [35] Junfei Tan, Yuxin Chen, An Zhang, Junguang Jiang, Bin Liu, Ziru Xu, Han Zhu, Jian Xu, Bo Zheng, and Xiang Wang. 2025. Reinforced preference optimization for recommendation. *arXiv preprint arXiv:2510.12211* (2025).
- [36] Jiakai Tang, Sunhao Dai, Teng Shi, Jun Xu, Xu Chen, Wen Chen, Jian Wu, and Yuning Jiang. 2025. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. *arXiv preprint arXiv:2503.22675* (2025).
- [37] Jiayi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.
- [38] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models finetuned with human feedback. *arXiv preprint arXiv:2305.14975* (2023).
- [39] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2400–2409.
- [40] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2025. Generative Recommendation: Towards Personalized Multimodal Content Generation. In *Companion Proceedings of the ACM on Web Conference 2025*. 2421–2425.
- [41] Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Y Zhang, Qing Cui, et al. 2023. Enhancing recommender systems with large language model reasoning graphs. *arXiv preprint arXiv:2308.10835* (2023).
- [42] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60.
- [43] Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7322–7329.
- [44] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
- [45] Runyang You, Yongqi Li, Xinyu Lin, Xin Zhang, Wenjie Wang, Wenjie Li, and Liqiang Nie. 2025. R²ec: Towards Large Recommender Models with Reasoning.

- In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [46] Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. 2025. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2025).
- [47] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 1435–1448.

A Datasets

We conduct experiments on three real-world benchmarks: Office Products (Office) and Industrial and Scientific (Industrial) from the Amazon Review Dataset³, along with the Yelp dataset⁴. To ensure data quality and computational efficiency, we applied the following preprocessing steps:

- (1) **5-core Filtering:** Removing users and items with fewer than five interactions.
- (2) **Time Window Selection:** Selecting interactions within specific time ranges (as detailed in Table 5) to maintain appropriate dataset sizes and focus on relevant interaction patterns.
- (3) **Sequence Truncation:** Truncating user interaction histories to a maximum sequence length of 10.
- (4) **Chronological Splitting:** Partitioning data into training, validation, and test sets with an 8:1:1 ratio based on the strict chronological order of interactions.

The detailed statistics of the processed datasets are summarized in Table 5.

Table 5: Statistics of the preprocessed datasets.

Dataset	Time Range	# Users	# Items	Train/Valid/Test
Office	2016.10 – 2018.11	8,319	3,452	38,883 / 4,860 / 4,861
Industrial	1996.10 – 2018.11	7,702	3,689	36,294 / 4,537 / 4,537
Yelp	2018.05 – 2018.11	6,231	4,852	37,579 / 4,697 / 4,698

B Baselines

We select a comprehensive set of baselines for comparison. All baselines are implemented via official repositories, with the best validation checkpoints selected for evaluation. The baselines are categorized into three groups: traditional recommendation models, LLM-based generative models (subdivided into text-based methods and SID-based generative methods), and reasoning-enhanced models.

1. Traditional Recommendation Models

- **SASRec** [19] is a sequential recommendation model equipped with self-attention mechanisms to capture long-term semantic dependencies.

2. LLM-based Generative Models We further divide this category into two streams based on their generation targets:

(1) *Text-based Methods:*

- **BIGRec** [2] proposes a bi-step grounding paradigm that fine-tunes LLMs to generate meaningful semantic tokens, which are subsequently mapped to actual items via a grounding mechanism.

³<https://jmcauley.ucsd.edu/data/amazon>

⁴<https://www.yelp.com/dataset>

- **SPRec** [11] is a self-play framework that iteratively aligns the model via DPO using self-generated negative samples to mitigate the filter bubble effect.

(2) *SID-based Generative Methods:*

- **TIGER** [31] is a transformer-based model that utilizes RQ-VAE to acquire item SIDs for autoregressive generation.
- **MiniOneRec** [22] is an open-source generative framework that establishes an end-to-end workflow, utilizing RQ-VAE for item indexing and employing reinforcement learning with hybrid rewards for optimization.

3. Reasoning-enhanced Models

- **ReaRec** [36] is an inference-time computing framework that enhances user representations through implicit autoregressive multi-step reasoning.
- **R²ec** [45] is a unified framework featuring a dual-head architecture that jointly optimizes reasoning chain generation and item prediction.

C Implementation Details

Model Architecture and Pre-training. We employ Qwen3-8B [44] as the backbone LLM and represent items using 4-layer SIDs. The item embeddings are initialized via Qwen3-Embedding-4B, and the underlying RQ-VAE [24] is trained on single GPU using the AdamW optimizer with a learning rate of 1e-3 and a batch size of 1,024 for 10,000 epochs.

Training Protocol. The training pipeline consists of two stages performed on 8 NVIDIA A100 GPUs. First, during the SFT stage, we apply LoRA [16] ($r = 16, \alpha = 32$) and train the model for 10 epochs using a learning rate of 2e-4 with a cosine decay schedule and a per-device batch size of 8. We evaluate the model every 0.5 epochs with a patience of 3 evaluations for early stopping. Subsequently, we warm-start from the SFT checkpoint to perform Uncertainty-Aware Preference Alignment. We lower the learning rate to 2e-6 while maintaining the same LoRA configuration. For the optimization objective, we set the KL penalty $\beta = 0$ and generate $G = 16$ candidates per sample via beam search. The hyperparameters are tuned via grid search: $\alpha \in \{0, 0.5, 0.8\}$, $\gamma_1 \in \{0.0, 0.1, 0.5\}$, and $\gamma_2 \in \{0.5, 1.0\}$.

Inference and Prompting. During inference, we generate recommendation lists using beam search with a beam size of 10. The explicit confidence scores are computed post-generation via Eq. (10). We format the input as a natural language instruction with special tokens for SIDs. An example is provided below:

Input:

The user has interacted with items <a_147><b_0><c_248><d_36>, <a_147><b_0><c_248><d_36>, <a_73><b_204><c_17><d_180> in chronological order. Can you predict the next possible item that the user may expect?

Response:

<a_70><b_140><c_195><d_192>

Risk-Aware Scoring Details. For the downstream applications in Section 4.5 (i.e., re-ranking, user rejection, and item truncation), we adopt a hybrid scoring strategy that integrates both the generative

likelihood and the explicit confidence. Specifically, the final ranking score is calculated as $Score = \exp(s_{\text{gen}}) + \lambda \cdot s_{\text{conf}}$, where s_{gen} denotes the sequence log-likelihood, s_{conf} is the explicit confidence score, and λ is a balancing factor set to 0.5. For the User-level Rejection task, we represent the confidence of a user request using the maximum hybrid score among its generated candidates.

Computational Cost. Table 6 compares resource consumption on the Office dataset.

Table 6: Computational cost comparison on Office.

Method	GPU Mem	Train / Epoch	Inference / Query
SFT	34G	7 min	32.9 ms
GRPO	50G	65 min	32.9 ms
UGR	60G	80 min	37.0 ms

Compared with GRPO, UGR introduces +10G memory and +15 min per epoch, primarily due to the auxiliary confidence alignment loss and one additional forward pass for confidence estimation. At inference time, the extra 4.1 ms latency is negligible and enables the downstream risk-aware applications described in Section 4.5, which standard methods cannot support.

D Additional Ablation Studies

Complete Component-wise Ablation. Table 7 reports all seven combinations across three datasets; the first row is the SFT baseline. E alone collapses far below the baseline, confirming that confidence alignment requires the reward structure from U or D. Among two-component variants, D+E is the strongest yet still trails the full model on all metrics. Only the complete U+D+E consistently achieves the best performance, showing that the three modules provide complementary contributions.

Table 7: Complete component-wise ablation. U: Uncertainty-Weighted Reward; D: Difficulty-aware Optimization; E: Explicit Confidence Alignment.

Components			Office		Industrial		Yelp	
U	D	E	HR@10	NG@10	HR@10	NG@10	HR@10	NG@10
			0.1563	0.1213	0.1378	0.1010	0.0423	0.0217
✓			0.1522	0.1193	0.1338	0.1029	0.0424	0.0222
	✓		0.1549	0.1202	0.1353	0.1022	0.0450	0.0236
		✓	0.0710	0.0530	0.0789	0.0616	0.0415	0.0220
✓	✓		0.1579	0.1223	0.1349	0.1030	0.0454	0.0239
✓		✓	0.1530	0.1196	0.1363	0.1023	0.0430	0.0225
	✓	✓	0.1583	0.1216	0.1373	0.1028	0.0456	0.0240
✓	✓	✓	0.1598	0.1231	0.1393	0.1033	0.0473	0.0250

Confidence Token Design Choice. Table 8 compares our binary design against multi-level discrete tokens and a regression head on Office. The binary design outperforms all alternatives. We attribute this to two factors: (i) recommendation supervision is intrinsically sparse (only one ground-truth per user), which is insufficient for fine-grained calibration; and (ii) binary tokens align naturally with GRPO’s pairwise ranking as they can be directly contrasted within

the group-relative advantage, whereas multi-level or regression targets introduce competing learning signals.

Table 8: Confidence token design comparison on Office.

	Binary (Ours)	5-Token	10-Token	Regression
HR@10	0.1598	0.1587	0.1579	0.1582
NG@10	0.1231	0.1220	0.1220	0.1222

Comparison with Calibration Methods. Existing uncertainty-aware methods in recommendation are limited to post-hoc inference-time quantification [23], rather than utilizing uncertainty during training. In the general LLM domain, RLCR [8] aligns confidence with correctness via calibration rewards, but targets absolute calibration suited for tasks with deterministic answers (e.g., math, coding). To validate our ranking-based design, we compare four variants on Office (Table 9): RLCR applied to our setting; UGR without the confidence alignment module; UGR with the pairwise ranking loss replaced by a BCE-based absolute calibration loss; and the full UGR.

Table 9: Comparison with calibration-based confidence alignment on Office.

Method	HR@10	NG@10
RLCR	0.1326	0.1008
UGR w/o confidence alignment	0.1579	0.1223
UGR w/ absolute loss	0.1585	0.1225
UGR	0.1598	0.1231

RLCR substantially underperforms all UGR variants, confirming that absolute calibration is ill-suited for the ranking nature of recommendation. Even without explicit confidence alignment, UGR already surpasses RLCR by a large margin, showing that leveraging uncertainty at the reward and optimization levels is more effective than post-hoc calibration. Replacing our pairwise ranking loss with a BCE-based absolute loss slightly degrades performance, further validating that relative confidence ordering is a better objective than absolute probability alignment in this setting.