A Systematic Analysis of Methods for Uncertainty Quantification in Distribution-based Recommender Systems

Joel Machado Pires[1], **Eduardo Ferreira da Silva**[1], **Frederico Araújo Durão** [1]

[1]Institute of Computing – Federal University of Bahia (UFBA)
Caixa Postal 40.170-110 – Salvador – BA – Brazil

{joelpires, eduardoferreira, fdurao}@ufba.br

Abstract. *Recommender systems (RS) have become necessary tools for personalizing content and services across digital platforms, yet their confidence estimation remains a critical concern. Confidence estimation in RS aims to answer how confident the model is in its predictions. Despite the advances, integrating confidence into RS models often degrades performance, with outcomes strongly dependent on both dataset characteristics and architectural suitability. To investigate these issues, we conducted a comprehensive empirical study over established distribution-based confidence-aware models, OrdRec, CPMF, CBPMF, LBD, PRGAT, and PRLIGHTGCN. We further performed systematic capacity ablations on deeper models. The methodology combined expected calibration error, confidence-error correlation, rating distribution comparisons, and capacity-vs-performance ablation. Our results show that integrating confidence estimation does not consistently improve and can degrade point-estimate performance in recommendation models. Increasing model capacity fails to yield gains and can negatively impact performance, even in the training set, indicating optimization and data constraints. Moreover, calibration quality is not reliably aligned with predictive accuracy. These models may exhibit meaningful confidence–error correlation while remaining poorly calibrated in absolute terms.*

1. Introduction

Recommender systems (RS) are important tools for helping users navigate large volumes of information by providing personalized content across domains. By leveraging machine learning techniques, these systems capture complex user–item interaction patterns, often through collaborative filtering approaches [da Silva et al. 2025]. Among these, model-based methods have demonstrated strong performance in prediction and ranking tasks.

Despite these advances, the reliability of recommender systems remains a critical concern. Traditional RS primarily focuses on optimizing point-prediction accuracy or ranking quality, often neglecting the uncertainty inherent in user preferences and sparse data [Wu 2024, Chen et al. 2025]. As a result, models may produce predictions without indicating their trustworthiness, leading to unreliable recommendations. Confidence estimation has emerged as a complementary objective, aiming to quantify the certainty of model predictions and provide users or downstream systems with an indication of reliability [Coscrato and Bridge 2023, Pires et al. 2026].

In recent years, research has highlighted several challenges in recommender systems that can directly affect model reliability. The limited availability of diverse

and explicitly labeled datasets constrains model learning [dos Santos et al. 2025]. This can lead to biased representations and unreliable output estimates. Fairness and exposure bias further distort interaction data, concentrating information on popular items [de Lourdes M. Silva et al. 2025]. Then, it reduces reliability by increasing uncertainty in long-tail predictions. Additionally, non-IID data distributions introduce distributional shifts that degrade performance [Negrão et al. 2025]. Together, these factors compromise both predictive reliability and highlight the need for a well-defined confidence estimation method.

In the literature, several approaches have been proposed to incorporate confidence estimation into collaborative filtering models. A common strategy is to impose probabilistic assumptions on model outputs, such as Gaussian, Gamma, or Beta distributions, to capture uncertainty. Notable examples include OrdRec (OrdRec) [Koren and Sill 2011], confidence-aware probabilistic matrix factorization (CPMF), confidence-aware bayesian probabilistic matrix factorization (CBPMF) [Wang et al. 2018], learned beta distributions (LBD) [Knyazev and Oosterhuis 2023], probabilistic recommender graph attention networks (PRGAT), and probabilistic recommender light graph convolutional networks (PRLIGHTGCN) [Pires et al. 2026], which extend an embedding-based model with distribution-based confidence modeling. Other works adopt heuristic strategies to estimate confidence and use the resulting estimates for calibration. These approaches generally interpret confidence as the model's belief in the correctness of its predictions, offering an additional layer of information beyond predicted ratings.

However, existing methods exhibit several important limitations [Coscrato and Bridge 2023, Pires et al. 2026]. For instance, prior work reports that incorporating confidence estimation often degrades recommendation performance, as observed with OrdRec, CPMF, CBPMF, and LBD. These, including PRGAT and PRLIGHTGCN, are highly dependent on dataset characteristics and model design. Some methods, such as CBPMF and LBD, suffer from practical issues, including convergence instability. More importantly, the literature lacks a systematic evaluation of confidence models, particularly by addressing why they degrade accuracy, why they produce uncalibrated confidence, and why they depend on the dataset or architectural design. Consequently, it remains unclear whether current confidence modeling techniques truly provide meaningful and trustworthy uncertainty estimates.

To address these gaps, we propose a comprehensive experimental study of confidence-aware recommender systems across both traditional and graph-based architectures. We further conduct systematic capacity ablations on deep architectures, including multi-head attention mechanisms and deep graph attention networks, to investigate how model expressiveness affects both performance and confidence quality. The evaluation framework combines standard recommendation metrics with detailed diagnostic analyses, including convergence behavior, calibration assessment via expected calibration error, distributional alignment between predicted and observed ratings, and capacity–performance trade-offs. Through this study, we aim to provide a deeper understanding of the interplay among accuracy, model capacity, and confidence estimation, and to assess whether current approaches reliably quantify uncertainty in recommender systems.

The remainder of this article is as follows. We compare prior works in the literature in Section 3. The experimental evaluation comprises the methodology, research

questions, datasets, metrics, and results in Section 4. We discuss and answer the research questions in the Section 5. Section 6 concludes the work and presents future directions.

2. Background

This section presents some confidence methods for recommendation from the literature: OrdRec [Koren and Sill 2011], CPMF, [Wang et al. 2018], and LBD [Knyazev and Oosterhuis 2023], PRLIGHTGCN [Pires et al. 2026]. Interestingly, these methods minimize the negative log loss function using a gradient descent-based algorithm,

$$\mathcal{L} = -\frac{1}{|\mathcal{R}|} \sum_{(u,i)\in\mathcal{R}} \log P(r_{ui} \mid \Theta)\,, \tag{1}$$

where $\mathcal{R}$ is the set of observed user-item interactions and Θ is the models' parameters.

2.1. OrdRec

Koren and Sill (2011) propose OrdRec, a collaborative filtering (CF) method that models user ratings as ordinal rather than numerical outcomes. For each user–item interaction, a latent preference score predicted by a matrix factorization (MF) model is assumed to follow a Normal distribution and is mapped to discrete rating levels through user-specific, monotonically increasing thresholds parameterized by learned scalars. The probability that a rating does not exceed a given level is modeled using a logistic function:

$$P(r_{ui} \leq r \mid \Theta) = \frac{1}{1 + e^{y_{ui} - t_r^u}}, \tag{2}$$

where y_{ui} is the predicted latent score and Θ denotes all model parameters. Model training minimizes the negative log-likelihood of observed ratings using stochastic gradient descent.

OrdRec also introduces a post-training calibration step to improve ranking quality. After fitting the ordinal model, a vector of ranking weights is learned for each rating level. For item pairs where the observed user preference is known, the difference between their predicted rating probability distributions is computed, and the weights are optimized to maximize a pairwise ranking objective:

$$F(\mathcal{T} \mid \mathbf{w}) = \sum_{(i,j)\in\mathcal{T}} f(\mathbf{w}^T \Delta P_u(i,j)), \tag{3}$$

where $\mathcal{T}$ denotes the set of correctly ordered item pairs and (f) is the sigmoid function.

During inference, the model outputs a full probability distribution over rating levels. The predicted rating is given by the expected value of this distribution, while a confidence estimate is derived from its standard deviation. For ranking tasks, items are scored using the calibrated ranking function $s_{ui} = \mathbf{w}^T \mathbf{p}_{ui}$ and sorted accordingly.

2.2. Confidence-Aware Matrix Factorization for Recommender Systems

This study presents two implementations of a confidence-aware MF model: CPMF and CBPMF. We describe the CPMF variant. Assuming the rating r_{ui} follows a normal distribution, the mean and variance are given by

$$\mu_{ui} = \mathbf{e}_u \cdot \mathbf{e}_i \text{ and } \nu = \frac{1}{\gamma_{U_i} \cdot \gamma_{V_j} \cdot \alpha_p}\,, \text{ respectively.} \tag{4}$$

In which γ_{Ui}, γ_{V_j}, and α_p are learned scalars, for each user and each item, respectively. Instead of modeling the likelihood of an exact rating value, the model defines a prediction interval within which the rating is expected to lie with a given confidence level p:

$$\left[\mu_{ui} - Z(p/2) \cdot \sqrt{\nu},\ \mu_{ui} + Z(p/2) \cdot \sqrt{\nu}\right] , \tag{5}$$

where $Z(p/2)$ denotes the quantile of the standard normal distribution corresponding to the two-tailed confidence level p.

2.3. Learned Beta Distributions

The proposed model represents each user–item interaction as a Beta distribution whose parameters are produced by a matrix factorization-like architecture. The mean parameter, μ_{ui}, is obtained from the cosine similarity between user and item embeddings, while the confidence term is defined as

$$\nu_{ui} = \|\mathbf{e}_u + \mathbf{e}_i\|_2. \tag{6}$$

The Beta shape parameters are constructed by scaling the mean and its complement by the confidence, and are further adjusted using learned global, user-specific, and item-specific bias terms, with a small constant added to ensure numerical stability. The predicted rating corresponds to the expected value of the resulting Beta distribution, rescaled to the original rating range.

To train the model, the normalized rating interval ([0,1]) is discretized into (S) adaptive bins whose boundaries depend on both the user and the item. Bin widths are determined by learned user and item embeddings through an exponential weighting scheme that is normalized to form a valid partition of the interval. The probability mass assigned to each bin is computed as the difference of the Beta cumulative distribution function evaluated at consecutive bin boundaries.

2.4. PRGAT and PRLIGHTGCN

These are a variants that applies the probabilistic recommender confidence integration into the graph convolutional neural network architecture. The authors combines the techniques, mainly from CPMF, with some improvements for obtaining a more stable version.

They optimize the probability of observing a rating, given a pair (u, i), as the same in CPMF, by assuming the ratings follow a normal distribution. The standard deviation of this distribution is

$$\sigma_{\mathbf{ui}} = \frac{1}{\sqrt{\gamma_{\mathbf{u}} \cdot \gamma_{\mathbf{i}} \cdot \nu'}} , \tag{7}$$

where γ_u, γ_i, and ν' are learned parameters for users, items, and global, respectively. The mean μ is estimated by a graph attention networks (GAT)-based or light graph convolutional networks (LIGHTGCN) model. The confidence estimation as the probability of observing the predicted is μ_{ui},

$$P(r_{ui} = \mu_{ui}|\Theta) = P(r_{ui} <= \mu_{ui} + \delta_r) - P(r_{ui} <= \mu_{ui} - \delta_r) , \tag{8}$$

where we define δ_r as a small tunable scalar. The final goal is to maximize the probability of observing the current observed rating and reduce the mean squared error (MSE) error

as well, changing the loss function to

$$\mathcal{L}_{PRGAT} = \frac{1}{|\mathcal{R}|} \sum_{(u,i)\in\mathcal{R}} \left(-\log(P(r_{ui} = \mu_{ui}|\Theta)) + \lambda \cdot \text{MSE}(r_{ui}, \mu_{ui})\right) , \tag{9}$$

where λ is a tunable parameter. This parameter controls the trade-off between the standard recommendation objective and the confidence-aware regularization.

3. Related Works

Coscrato and Bridge (2023) conducted an extensive empirical study comparing several classes of uncertainty estimators for recommender systems, among them, OrdRec and CPMF. Their results revealed important concerns regarding the reliability and consistency of uncertainty estimates for rating prediction. Many models showed inconsistent correlations across different uncertainty–error evaluation metrics. For instance, a method could exhibit a strong Pearson correlation with prediction error while performing poorly under alternative measures such as Spearman correlation or UPI. In addition, the authors introduced two new approaches: ENSEMBLE and EB-Linear. However, the experimental results indicated that they did not outperform simpler, information-based estimators in most evaluation metrics [Coscrato and Bridge 2023].

Pires, Silva, and Durão (2025) explored ORDREC, CBPMF, CPMF, and LBD, identifying convergence issues, architecture suitability, and dataset dependency, confidence miscalibration, and degraded accuracy resulting from confidence integration in the datasets they evaluated. They then proposed a confidence-integration approach to address these issues, introducing two variants: PRGAT and PRLIGHTGCN. However, their solution still exhibits confidence miscalibration and is dataset-dependent [Pires et al. 2026].

Conversely, in this work, we explore why these prior-distribution-based methods often harm accuracy, why there is consistent dataset dependency across these models, and why confidence calibration in the rating prediction task remains poor. These issues remain unexplored by these aforementioned works.

4. Experimental Evaluation

Our proposal is to experiment with the prior distribution-based approaches, exploring the aspects of the prior works that have not been explored. For instance, we design an experimental evaluation to identify the causes of the issues. The following section presents the experimental evaluation, including the research questions, method configurations, datasets, and metrics. The source code of this work is publicly available at: https://github.com/joel021/distribution-conf-rating-recsys.

4.1. Methodology

This experimental evaluation is designed to answer the following research questions:

1. Do current confidence modeling techniques provide meaningful and reliable uncertainty estimates that correlate with prediction error?
2. Why does the integration of distribution-based confidence estimation methods degrade recommendation accuracy in collaborative filtering models?

3. Why do existing confidence estimation approaches fail to produce well-calibrated confidence scores in rating prediction tasks?
4. To what extent are confidence-aware recommender models dependent on dataset characteristics and architectural design?
5. How does model capacity affect the trade-off between recommendation performance and confidence quality?

We evaluate a set of distribution-based and confidence-aware recommendation models, including OrdRec, CPMF, CBPMF, LBD, PRGAT, and PRLIGHTGCN. For all models, the embedding dimensions were set according to the configurations reported in their original papers: 512 for OrdRec, 20 for both CPMF and CBPMF, 522 for LBD, and 64 for both PRGAT and PRLIGHTGCN.

All models are trained using the Adam optimizer with a learning rate of 0.001, except for CBPMF, which is optimized via Gibbs sampling as described in [Wang et al. 2018]. Early stopping is used as a regularization strategy, with a patience of 5 epochs for no improvement in the evaluation error. This threshold is chosen based on empirical observations that performance typically stagnates after 2 epochs, with an additional 3 epochs to ensure stability. Model evaluation is conducted using Monte Carlo cross-validation with five runs: the first run uses a sorted data split, while the remaining four runs use randomized splits. The expected calibration error (ECE) error is computed using a threshold $\tau = 0.5$, corresponding to the midpoint between consecutive rating values in the adopted rating scale 1, 2, 3, 4, 5.

After analyzing the relationship between confidence estimates and absolute error in these models, we conduct additional experiments with deeper architectures that do not explicitly incorporate confidence. These include deep MF, a deeper DGAT variant, and a multi-head attention model, enabling a comparative analysis of whether increased model capacity alone can address the observed limitations

4.2. Datasets

The experiments are conducted on three publicly available datasets widely used in recommender systems research: Amazon Movies and TVs [He and McAuley 2016], Jester Joke [Goldberg et al. 2001], and MovieLens 1M [Harper and Konstan 2015]. These datasets are selected based on prior work to capture diverse recommendation scenarios across different domains, feedback types, and sparsity levels. All datasets are tabular and include user identification (ID), item ID, explicit feedback ratings, and timestamps.

The Amazon Movies and TV dataset captures large-scale user interactions in the movie and television domain, including both purchase and rating information. In its original form, it contains over 2 million users, 200 thousand items, and more than 4.6 million interactions, with discrete ratings typically ranging from 1 to 5, resulting in an extremely sparse user–item matrix. After preprocessing steps, including outlier removal and cold-start filtering, the dataset is reduced to 1,316 users and 7,164 items, with approximately 1.79 million interactions and a sparsity level of 0.9998.

The Jester Joke dataset represents a humor recommendation scenario with continuous-valued explicit feedback. It contains approximately 4.1 million interactions, where 7,699 users rate 158 jokes using real-valued scores in the range $[-10.00, +10.00]$.

After preprocessing, the dataset includes 2,498 users and 99 items, with about 1.8 million interactions and a significantly lower sparsity level of 0.2720.

The MovieLens 1M dataset is a standard benchmark for collaborative filtering in the movie domain, containing over 1 million ratings from 6,040 users on 3,416 movies, with ratings ranging from 1 to 5. Due to the large number of possible user–item pairs, the data exhibit high sparsity. After preprocessing, the dataset retains 999,611 interactions with a sparsity of 0.9515. These data are split into training, evaluation, and test sets using ratios of 75%, 12.5%, and 12.5%, respectively

4.3. Metrics

To quantify the alignment between model certainty and accuracy, this analysis uses the Pearson correlation as the primary metric to assess the association between confidence levels and observed errors.

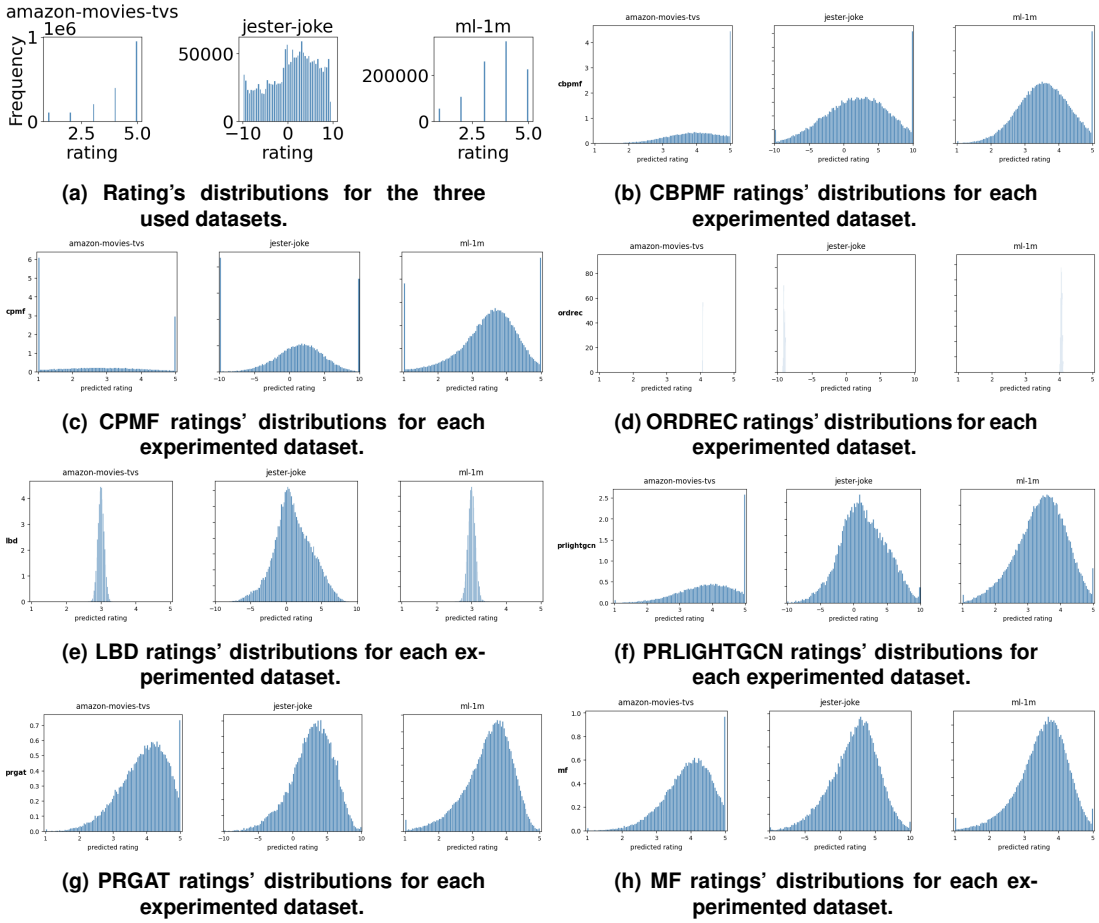
The ECE [Naeini et al. 2015] is used in this study to quantify the discrepancy between predicted confidence and empirical accuracy. Given a set of predictions with associated confidence scores, the interval ([0,1]) is partitioned into (M) equally spaced bins, and ECE is computed as a weighted average of the absolute difference between accuracy and confidence within each bin. The implementation adopted in this work follows the standard formulation but uses right-closed intervals ((b_{m-1}, b_m]) for bin assignment, ensuring that each prediction is assigned to a unique bin while excluding the left boundary to avoid overlap. Additionally, empty bins are ignored, and each bin is weighted by its sample size. To compute the accuracy per batch, it is necessary to identify correct predictions. The correct value is obtained by applying a threshold to the absolute difference between the predicted and observed ratings for each prediction.

$$\text{acc} = \begin{cases} 1, & \text{if } |\hat{r} - r| < \tau \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

We also utilize root mean squared error (RMSE) to evaluate the model's rating prediction (regression) error. normalized discounted cumulative gain (NDCG) and mean average precision (MAP) are used to evaluate the ranking quality of the models.

4.4. Results

The figures 1a, 1b, 1c, 1d, 1e, 1g, and 1f illustrate a significant divergence between the model-generated rating densities and the empirical distributions (Figure 1a) of the datasets. Therefore, it's possible to observe relatively low RMSE scores for these models, even though their predictive distributions match the observed sample distribution. For instance, in the Amazon Movies TVs dataset, the true distribution is dominated by 5-star ratings. However, models like PRGAT and MF produce Gaussian-like curves that peak between 4.0 and 4.5. LBD represents the extreme of variance capture failure, collapsing almost entirely into a Dirac-delta-like peak around the mean. This indicates that, despite the model's complexity, it treats most user-item pairs as averages.



(a) Rating's distributions for the three used datasets.

(b) CBPMF ratings' distributions for each experimented dataset.

(c) CPMF ratings' distributions for each experimented dataset.

(d) ORDREC ratings' distributions for each experimented dataset.

(e) LBD ratings' distributions for each experimented dataset.

(f) PRLIGHTGCN ratings' distributions for each experimented dataset.

(g) PRGAT ratings' distributions for each experimented dataset.

(h) MF ratings' distributions for each experimented dataset.

Figure 1. Ratings' distributions across different models and datasets.

Conversely, CPMF and CBPMF exhibit "U-shaped" distributions. While these models capture the presence of 1- and 5-star ratings, they do so by pushing the bulk of the density to the boundaries. This creates a binary-like predictive behavior that does not appear in the observed distributions of the datasets. In summary, the models are not just "predicting ratings"; they are generating a new, artificial rating space that lacks the statistical characteristics of the original data (approximating the assumed distribution). This explains why recommendation performance degrades when confidence is integrated. The models are optimizing for a distributional shape that the data itself does not support.

The result in Figure 1h also shows that non-distribution-based models, like MF, have a mismatch in distribution shapes. This highlights that, although these models' loss functions reduce RMSE, they should still include regularization signals to control the rating distribution. Another interesting result is that increasing capacity (e.g., moving from MF to DGAT) does not yield a distribution that better matches the ground truth. This result motivates the capacity ablation study.

From the Table 1, PRLIGHTGCN consistently achieves the strongest negative correlation between confidence and error. Yet, Table 2 shows some of the highest ECE errors. This indicates that while the model is excellent at ranking its confidence, the absolute probability values it produces are poorly scaled. Then, although it presents the best observed calibration, it is still bad.

Table 1. Confidence correlation with error, ranging from -1 to 1, where -1 is the desired correlation. 0 means no correlation, and values closer to 1 indicate a positive, undesired correlation. This correlation measures the level of confidence in calibration. Bold values mean the true best values.

Model	Amazon Movies TVs	Jester Joke	Movie Lens 1M
PRGAT	-0.1810 ± 0.0424	-0.3586 ± 0.0342	-0.2615 ± 0.0057
PRLIGHTGCN	**-0.2632 ± 0.0191**	**-0.3646 ± 0.0163**	-0.2571 ± 0.0113
ORDREC	-0.0003 ± 0.0061	0.0531 ± 0.0701	0.0189 ± 0.0055
CPMF	0.1675 ± 0.0854	-0.3296 ± 0.0985	**-0.3349 ± 0.0561**
CBPMF	-0.0179 ± 0.0067	-0.0426 ± 0.0248	-0.0128 ± 0.0035
LBD	0.0059 ± 0.0053	-0.1459 ± 0.0263	0.0145 ± 0.0098

Table 2. ECE error of the distribution-based models across the datasets. 0 means the confidence is fully calibrated, and 1 means it is not calibrated at all.

Model	Movie Lens 1M	Jester Joke	Amazon Movies TVs
PRGAT	0.088 ± 0.123	0.695 ± 0.019	0.114 ± 0.040
PRLIGHTGCN	0.358 ± 0.003	0.690 ± 0.009	0.367 ± 0.013
ORDREC	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
CPMF	0.328 ± 0.008	0.683 ± 0.006	0.278 ± 0.027
CBPMF	0.130 ± 0.005	0.340 ± 0.044	0.145 ± 0.003
LBD	0.211 ± 0.022	0.258 ± 0.102	0.096 ± 0.005

Table 3. Metrics of the base models (MF and GAT) used in prior distribution-based models in the Amazon Movies TVs dataset.

Models	RMSE	MNDCG@10	MAP@10	MNDCG@3	MAP@3
MF	0.967±0.011	0.299±0.015	0.265±0.014	0.345±0.018	0.337±0.018
DMF	1.493±0.015	0.176±0.001	0.085±0.0	0.167±0.001	0.264±0.001
DGAT	0.95±0.008	0.144±0.021	0.149±0.016	0.112±0.03	0.117±0.03
DGAT2	0.947±0.007	0.176±0.018	0.171±0.014	0.169±0.023	0.173±0.022
DGAT3	0.948±0.007	0.175±0.019	0.169±0.015	0.17±0.025	0.172±0.023

Table 4. Metrics of the base models (MF and GAT) used in prior distribution-based models in the Jester Joke dataset.

Models	RMSE	MNDCG@10	MAP@10	MNDCG@3	MAP@3
MF	4.561±0.049	0.639±0.153	0.318±0.065	0.605±0.17	0.407±0.111
DMF	6.051±0.084	0.574±0.054	0.266±0.021	0.488±0.044	0.315±0.007
DGAT	4.859±0.024	0.678±0.121	0.29±0.054	0.676±0.11	0.397±0.085
DGAT2	4.766±0.163	0.714±0.043	0.421±0.02	0.682±0.06	0.526±0.037
DGAT3	4.802±0.148	0.704±0.056	0.412±0.024	0.664±0.101	0.504±0.067

Table 5. Metrics of the base models (MF and GAT) used in prior distribution-based models in the Movie Lens 1M dataset.

Models	RMSE	MNDCG@10	MAP@10	MNDCG@3	MAP@3
MF	0.882±0.009	0.457±0.041	0.435±0.036	0.495±0.056	0.505±0.055
DMF	1.241±0.023	0.196±0.001	0.07±0.002	0.183±0.002	0.225±0.005
DGAT	0.881±0.011	0.422±0.029	0.409±0.03	0.386±0.033	0.419±0.036
DGAT2	0.885±0.007	0.426±0.048	0.403±0.044	0.427±0.061	0.441±0.064
DGAT3	0.884±0.005	0.435±0.043	0.41±0.039	0.442±0.056	0.455±0.059

ORDREC shows perfect calibration in terms of ECE, but exhibits near-zero correlation with the error. This suggests that it produces a nearly constant confidence estimate that aligns with the global error rate, yet fails to distinguish between high- and low-certainty user–item pairs. Figure 1d shows that this model's predicted ratings are highly concentrated, which helps explain the low ECE since the model's overall accuracy is also low.

The Jester Joke dataset consistently yields poor confidence calibration across all models, resulting in low confidence-error correlation or high ECE error. We further bring the performances of the baselines for the prior distribution-based models as an additional evaluation of whether there is or not an architectural dependence of the dataset, in Tables 3, 4, and 5. In these tables, DGAT uses the implementation in [Pires et al. 2026] with one head attention, DGAT2 uses two, and DGAT3 uses three. DMF adapts the implementation in [Xue et al. 2017], which is deeper than pure standard MF.

5. Discussion

These results are enough to answer the research questions. Q1: **Do current confidence modeling techniques provide meaningful and reliable uncertainty estimates that correlate with prediction error?** No. Although these solutions are promising and represent meaningful advances, they still require further improvement. This conclusion, supported by the analysis of ECE and the Pearson correlation between confidence and absolute error, aligns with the findings of Pires, Silva, and Durão (2025) and the observations of Coscrado *et al.* (2023). While the evaluated methods may achieve strong performance on specific datasets, their inconsistent results across the experimental datasets indicate limited generalization capability and suggest that they are not yet ready for production deployment.

Q2: **Why does the integration of distribution-based confidence estimation methods degrade recommendation accuracy in collaborative filtering models?** Distribution-based methods impose a predefined distribution over the predicted ratings, typically assuming forms such as Gaussian or Beta distributions. However, this assumption may be inappropriate when the true observed distribution does not conform to these families, as is the case in the experimental datasets. Pires, Silva, and Durão (2025) found that a primary focus on negative log likelihood loss (NLL), rather than balancing it with a pointwise error metric, constitutes an improper loss formulation and contributes to the observed performance issues. However, this limitation is only part of the problem. In addition to the loss formulation, our analysis reveals another issue: a mismatch between

the model's assumed distribution and the true distribution of observed ratings, which also degrades performance.

Prior work, along with our empirical findings, suggests a strong interaction between dataset characteristics and model architecture, indicating that certain architectures are better suited to specific datasets. This observation points to architecture–dataset alignment rather than a universally optimal model.

Under this perspective, one might expect that increasing the capacity of a well-aligned architecture would improve its ability to fit the data, particularly in the training set, potentially leading to overfitting. Furthermore, previous studies have reported that improvements in recommendation performance are often accompanied by better confidence calibration, motivating the following research question, Q3: **How does model capacity affect the trade-off between recommendation performance and confidence quality?** As shown in Tables 3, 4, and 5, increasing model capacity does not necessarily improve performance, including in the training set. This suggests that, in this context, capacity expansion introduces optimization or generalization challenges that outweigh its representational benefits. Consequently, model capacity does not monotonically improve either recommendation performance or confidence quality, even when the underlying architecture is well-aligned with the dataset.

Q5: **Why do existing confidence estimation approaches fail to produce well-calibrated confidence scores in rating prediction tasks?** A key factor underlying this limitation is the mismatch between the empirical distribution of ratings and the distribution implicitly assumed by commonly used loss functions. When the training objective enforces a distributional shape that does not align with the observed data, the model is effectively optimized to fit an incorrect target. This misalignment weakens its ability to capture the true data-generating patterns, leading not only to degraded predictive performance but also to poorly calibrated confidence estimates. However, this issue is not solely attributable to a distribution mismatch. Calibration errors may also arise from model misspecification, optimization dynamics, and the discrete and often sparse nature of rating data. Therefore, while loss-induced distribution mismatch is a central contributing factor, calibration failure in rating prediction should be understood as a consequence of multiple interacting components.

Q4: **To what extent are confidence-aware recommender models dependent on dataset characteristics and architectural design?** The results indicate that confidence-aware recommender models are strongly dependent on dataset characteristics, particularly sparsity and interaction distribution, while the impact of architectural design appears limited under these conditions. Contrary to prior expectations, we do not observe clear architectural superiority in terms of predictive performance. Instead, multiple models converge to similar performance levels, suggesting that dataset constraints dominate model behavior.

This effect is further supported by the capacity-ablation results and the training dynamics. Several models exhibit early convergence, with performance improvements concentrated in the initial epochs and little to no gain thereafter. Such behavior indicates that models quickly exhaust the available signal in the data, pointing to a limited ability to learn beyond surface-level patterns. Consequently, differences between models may

reflect marginal variations around a performance ceiling imposed by the data, rather than meaningful advances in representation learning.

The long-tail nature of user–item interactions provides additional explanation. With relatively few observations per user or item, models are unable to reliably estimate individualized preference distributions and often default to population-level tendencies. This limitation not only constrains recommendation accuracy but also hinders the learning of well-calibrated confidence estimates, leaving substantial room for improvement in uncertainty modeling.

These findings suggest that, in sparse regimes, improvements in model architecture or capacity alone are insufficient to significantly enhance performance or calibration. Instead, approaches that increase effective data density may be necessary to unlock further gains. In this context, dynamic frameworks may become more effective once sufficient interaction data is accumulated.

6. Conclusion

This work shows that current confidence-aware recommender models are not yet capable of producing fully reliable uncertainty estimates in rating prediction tasks. While some models exhibit meaningful correlation between confidence and prediction error, this relationship remains insufficient for robust calibration. A fundamental disconnect between optimization objectives and the quality of uncertainty is among the causes of the issue. The results further indicate that these limitations are not primarily due to architectural design or model capacity, but rather stem from objective misalignment and intrinsic dataset constraints, such as sparsity and long-tail interaction patterns, which restrict the models' ability to learn faithful predictive distributions.

Despite these limitations, the observed partial alignment between confidence and error suggests that current approaches are not entirely ineffective, but instead capture an incomplete notion of uncertainty. This opens a promising direction for future research based on dynamic learning paradigms. In particular, reinforcement learning and other continuously updated systems provide a natural framework to investigate how confidence quality evolves as more interaction data becomes available. Such settings would allow tracking the progression of calibration from weak or near-zero correlation toward stronger alignment with prediction error, both during online data accumulation and throughout the training process. Additionally, prior distribution-based approaches may become more suitable in these regimes, where increased data density enables more reliable estimation of user-specific distributions. Overall, advancing confidence-aware recommendations will likely require shifting from static evaluation to dynamic, data-evolving frameworks that explicitly model the temporal evolution of uncertainty.

7. Acknowledgements

We want to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Brazil) for funding this research under Grant Code 001 and 88887.310805/2026-00.

References

Chen, Q., Li, X., Fang, Y., and Wang, M. (2025). Advancing confidence calibration and quantification in medication recommendation. In *Proceedings of the 31st ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining V.1, KDD '25, page 106–117, New York, NY, USA. Association for Computing Machinery.

Coscrato, V. and Bridge, D. (2023). Estimating and evaluating the uncertainty of rating predictions and top-n recommendations in recommender systems. *ACM Trans. Recomm. Syst.*, 1(2).

da Silva, D., Pires, J., and Durão, F. (2025). Exploiting surrogate submodular and cost-effective lazy forward algorithms for calibrated recommendations. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 98–111, Porto Alegre, RS, Brasil. SBC.

de Lourdes M. Silva, M., Chaves, I., Mendonça, A. L., Neto, E. D., and Machado, J. (2025). Twix: Balancing fairness and utility in item exposure for recommendation systems. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 427–440, Porto Alegre, RS, Brasil. SBC.

dos Santos, J. V. F., Nascimento, R., Camargo, A. C., Canuto, S., Costa, G., and Sousa, D. (2025). Nova base de dados brasileira para sistemas de recomendação de artigos científicos. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 289–302, Porto Alegre, RS, Brasil. SBC.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Knyazev, N. and Oosterhuis, H. (2023). A lightweight method for modeling confidence in recommendations with learned beta distributions. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 306–317.

Koren, Y. and Sill, J. (2011). Ordrec: an ordinal model for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 117–124, New York, NY, USA. Association for Computing Machinery.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Negrão, A., Rocha, G., dos Santos, L., Malaquias, P., Pedrosa, R., Fortes, R., and Silva, P. (2025). Mitigando impactos de distribuições não-iid em aprendizagem federada para sistemas de recomendação. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 413–426, Porto Alegre, RS, Brasil. SBC.

Pires, J. M., Silva, E. F. d., and Durão, F. A. (2026). Exploiting distribution-based confidence integration in graph neural network recommenders. *Applied Intelligence*, 56(5):142.

Wang, C., Liu, Q., Wu, R., Chen, E., Liu, C., Huang, X., and Huang, Z. (2018). Confidence-aware matrix factorization for recommender systems. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 32.

Wu, L. (2024). Towards trustworthy graph neural networks and their applications in recommender systems. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8250–8252.

Xue, H.-J., Dai, X., Zhang, J., Huang, S., and Chen, J. (2017). Deep matrix factorization models for recommender systems. In *Ijcai*, volume 17, pages 3203–3209. Melbourne, Australia.