Post-Hoc Dirichlet Fine-Tuning of Variational Autoencoders for Collaborative Filtering: Accuracy Gains and Popularity Costs

Mateus Toledo da Silveira Leao, Bapi Dutta(✉), Raciel Yera, and Luis Martínez

Department of Computer Science, University of Jaén, Spain
bdutta@ujaen.es

Abstract. Dirichlet-based uncertainty objectives have demonstrated accuracy and calibration gains in recommendation, but their structural interaction with popularity bias in VAE-based collaborative filtering is unknown. We propose a post-hoc fine-tuning protocol that attaches a lightweight evidence head to a frozen MultiVAE backbone and optimises it under two Dirichlet objectives: Evidential Deep Learning (EDL) and Belief Matching (BM). Both methods consistently improve ranking accuracy across three datasets spanning sparsity levels from 96.96% to 99.999%: gains reach +10.06% NDCG@10 on MovieLens-1M (EDL) and +18.23% NDCG@10 on Amazon Office Products (BM). However, accuracy improvement is structurally coupled to popularity concentration. We show that the Dirichlet KL regulariser creates an asymmetric gradient: popular items accumulate evidence proportional to their interaction frequency, while long-tail items are penalised in nearly every minibatch and collapse toward the Dirichlet prior. On dense data, the mean per-item concentration $\bar{\alpha}_k$ correlates with popularity rank at $\rho = -0.989$, and item coverage drops by up to 24% relative to the backbone. On ultra-sparse data, the evidence head fails entirely to learn item-level discrimination; recommendations default to the backbone's pre-trained logit ordering, amplifying its embedded popularity bias without any corrective capacity. These two failure modes define qualitatively distinct operating regimes whose boundary is determined by interaction density. Monitoring the $\bar{\alpha}_k$ distribution serves as a lightweight diagnostic that practitioners can compute at no additional cost to assess item bias risk before deployment.

Keywords: Recommender Systems · Variational Autoencoders · Evidential Deep Learning · Belief Matching · Uncertainty Quantification · Popularity Bias

1 Introduction

Variational Autoencoders (VAEs) have established themselves as a highly effective standard for collaborative filtering (CF) by capturing complex user-item relationships [12]. Despite the effectiveness of VAEs, models trained to maximise

point-estimate ranking metrics inherit a structural popularity bias: items observed by highly active users accumulate disproportionate training signals, causing the model to systematically under-recommend the long tail [1,18]. Concurrently, to improve the trustworthiness of models dealing with natural data uncertainty, researchers have developed Uncertainty Quantification (UQ) methods[7]. Foundational Dirichlet-based techniques, such as Evidential Deep Learning (EDL) [14] and Belief Matching (BM) [8], replace softmax point estimates with distributions over class probabilities. Recently, researchers have adapted these UQ principles to recommendation via aleatoric modelling in CF [16] and post-hoc Dirichlet wrappers in sequential architectures [3].

Unlike standard recommenders that output a single relevance score per item, a Dirichlet objective trains the model to output an amount of evidence for each item. This evidence yields both a point estimate of relevance and a measure of confidence. Frequently interacted items accumulate high evidence, whereas items with sparse histories fall back to a default, low-confidence state. As long as making the model's uncertainty explicit is highly appealing for recommendations, it introduces the core question this paper studies: in catalogues dominated by rare items, does tying relevance to accumulated evidence inadvertently bias the system toward the popular few?

While these Dirichlet objectives successfully improve calibration and predictive accuracy, their behavioural impact on the popularity dynamics of VAE-based architectures remains, to the best of our knowledge, unexplored. The core problem is that recommendation catalogues are already heavily skewed by popularity bias [1], a structural problem that standard CF architectures struggle to correct. Applying uncertainty objectives to this domain carries two structural risks. First, because Dirichlet regularisation inherently penalises unobserved data, it may systematically suppress niche items. Second, in highly sparse environments, these probabilistic heads risk failing to learn item-level discrimination altogether, further amplifying the pre-existing biases of the underlying architecture.

This research gap motivates the following research question: To what extent do post-hoc Dirichlet objectives alter the accuracy and popularity dynamics of VAE-based CF across different sparsity regimes, and what theoretical mechanisms drive these changes?

Our goal is to characterise the accuracy and beyond-accuracy behaviour of post-hoc Dirichlet fine-tuning on VAE-based CF. Concretely, we apply EDL and BM objectives to a pre-trained MultiVAE[11] backbone, evaluate the resulting accuracy–popularity outcomes across datasets of varying interaction density, and analyse the underlying mechanism through Dirichlet concentration analysis [15]. We make the following contributions:

1. We formulate a multi-label generalisation of the EDL and BM reconstruction and Kullback-Leibler (KL) terms for the full user–item interaction matrix setting.
2. We quantify a structural accuracy–popularity trade-off: post-hoc Dirichlet fine-tuning improves NDCG@10 by up to +18.23% over MultiVAE, but

reduces item catalogue coverage by up to 24% (in five of six configurations) and inflates popularity concentration on every dataset.
3. We trace this degradation to the asymmetric gradient pressure of the Dirichlet KL regulariser, validated through a $\rho = -0.989$ Spearman correlation between item popularity rank and $\bar{\alpha}_k$ on dense data.
4. We characterise two distinct operational regimes determined by interaction density: an evidence-driven regime, where the Dirichlet head learns meaningful item discrimination, and a decoder-prior-dominated regime, where it collapses to the uniform prior and amplifies the backbone's embedded popularity bias.

2 Related Work

Popularity bias in collaborative filtering. The systematic over-recommendation of popular items has been shown to impact fairness [4], item under-exposure [18], and beyond-accuracy metrics such as catalogue coverage [10]. Mitigation strategies fall into two broad categories: post-processing re ranking [2], which adjusts recommendation lists after training, and regularisation applied directly during the training phase [1]. Our work does not propose a mitigation strategy. Instead, we characterise how a probabilistic fine-tuning objective, applied for uncertainty quantification, interacts with the pre-existing popularity bias of the backbone and explain the structural mechanism behind this interaction.

Uncertainty quantification in recommendation. Joo et al. [9] proposed placing a Dirichlet prior over the VAE latent space to improve representation learning in general generative modelling; their approach modifies the latent prior and was not applied to recommendation. By contrast, we attach a post-hoc Dirichlet evidence head to the output layer of a pre-trained MultiVAE, following the EDL [14] and BM [8] objectives for uncertainty quantification over item relevance. Wang et al. [16] proposed the aleatoric uncertainty-aware recommendation (AUR) framework to model the randomness of unobserved interactions in VAE-based CF, showing that skewed training signals in sparse interaction matrices create a mislabelling problem that systematically disadvantages long-tail items. Fan et al. [5] modelled user history as a distribution for sequential recommendation. Balsells-Rodas et al. [3] established that Dirichlet objectives improve uncertainty calibration in the sequential setting. To the best of our knowledge, we are the first to apply post-hoc Dirichlet fine-tuning to VAE-based CF, evaluate it against beyond-accuracy criteria, and characterise its operating regimes across datasets of varying interaction density.

3 Methodology

3.1 MultiVAE for Collaborative Filtering

MultiVAE [11] frames collaborative filtering as variational inference over a neural generative model. For user u with binary interaction history $\mathbf{r}_u \in \{0,1\}^{|\mathcal{I}|}$, an

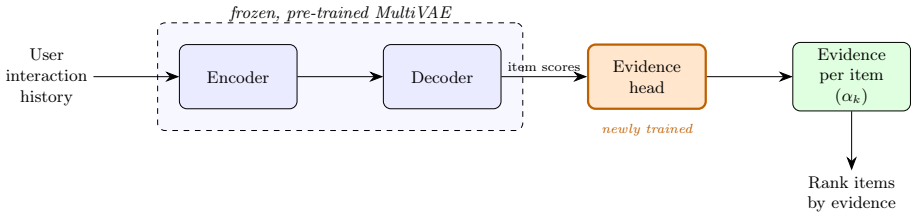


Fig. 1. Overview of the post-hoc fine-tuning protocol. The pre-trained MultiVAE encoder and decoder are kept frozen; only a small *evidence head* is added and trained. It turns each item's score into a Dirichlet evidence value α_k, and items are ranked by this evidence. During an initial warmup only the head is trained, after which the backbone is unfrozen for fine-tuning.

MLP encoder produces an approximate Gaussian posterior $q_\phi(\mathbf{z}_u \mid \mathbf{r}_u)$, and a decoder maps the sampled latent $\mathbf{z}_u$ to item logits $\mathbf{f}_\theta(\mathbf{z}_u) \in \mathbb{R}^{|\mathcal{I}|}$, whose k-th component f_k scores item k. The network is trained by maximising the standard β-ELBO with a multinomial reconstruction likelihood and a linearly annealed KL weight [11]. We train the backbone using the RecBole implementation [17] without modification and use the resulting checkpoint as the starting point for post-hoc Dirichlet fine-tuning.

3.2 Post-Hoc Fine-Tuning Protocol

Given a pre-trained MultiVAE checkpoint, we freeze the original encoder and decoder weights and attach a lightweight *evidence head* (Figure 1). The head is applied element-wise to each scalar decoder logit $f_k \in \mathbb{R}$:

$$\mathrm{EvidenceHead}(f_k) = W_2 \tanh\big(W_1 f_k + b_1\big) + b_2, \tag{1}$$

where $W_1 \in \mathbb{R}^{16\times 1}$, $b_1 \in \mathbb{R}^{16}$, $W_2 \in \mathbb{R}^{1\times 16}$, and $b_2 \in \mathbb{R}$. The Dirichlet concentration for item k is then

$$\alpha_k = \mathrm{softplus}\big(\mathrm{EvidenceHead}(f_k)\big) + 1. \tag{2}$$

The additive constant ensures $\alpha_k > 1$ for all items, which guarantees the existence of the Dirichlet mode and keeps every evaluation of the digamma function away from its singularity at zero. Applying the head to each logit independently via an unsqueeze–squeeze operation preserves the $O(|\mathcal{I}|)$ complexity of the original decoder and introduces no cross-item dependencies beyond those already encoded in $\mathbf{f}_\theta$.

Fine-tuning proceeds in two sequential phases. During the warmup phase ($E_w = 5$ epochs), the encoder and decoder parameters are frozen and only the evidence-head parameters are updated via the Adam optimiser at the full base learning rate. This isolation ensures that the concentration vector $\boldsymbol{\alpha}$ develops a meaningful distribution before the backbone is exposed to the Dirichlet objective.

Once the warmup concludes, the backbone is unfrozen, and training continues under different optimiser configurations for each loss function.

For BM, all parameters (encoder, decoder, and evidence head) are updated with a uniform Adam learning rate of $\eta = 10^{-4}$ throughout the main phase ($E = 30$ epochs). For EDL, we scale the learning rate of the pre-trained encoder and decoder down by a factor of 100 (to $\eta_{\text{bb}} = 10^{-6}$) while retaining the full rate for the evidence head ($\eta_{\text{head}} = 10^{-4}$). This asymmetric schedule is necessary for EDL because the log-form marginal likelihood loss (Eq. 3 of [14]) can yield large gradient magnitudes that destabilise pre-trained representations when applied without damping. The checkpoint with the highest validation NDCG@20 is retained as the final model.

3.3 Multi-Label Dirichlet Losses

The original EDL and BM formulations target a single positive class. In collaborative filtering, each user u possesses a set of positive items $\mathcal{R}_u = \{k : r_{uk} = 1\}$, which is typically non-singleton. We generalise both reconstruction terms to this multi-label setting by averaging the per-item loss over the observed positives.

Belief Matching. Substituting the multi-label interaction structure into the BM ELBO reconstruction term [8], the expected log-likelihood of all observed interactions under the predicted Dirichlet is

$$\mathcal{L}_{\text{recon}}^{\text{BM}} = \frac{1}{|\mathcal{R}_u|} \sum_{k \in \mathcal{R}_u} \big[\psi(\alpha_0) - \psi(\alpha_k)\big]. \tag{3}$$

This is a loss to be minimised; the sign is the negation of the ELBO reconstruction term so that minimising Eq. (3) corresponds to maximising $\sum_{k \in \mathcal{R}_u} [\psi(\alpha_k) - \psi(\alpha_0)]/|\mathcal{R}_u|$. When $|\mathcal{R}_u| = 1$ the expression reduces exactly to the single-label BM [8].

Evidential Deep Learning. Substituting the multi-label structure into the Type-II Maximum Likelihood objective [14], the negative log marginal likelihood averaged over positives is

$$\mathcal{L}_{\text{recon}}^{\text{EDL}} = \frac{1}{|\mathcal{R}_u|} \sum_{k \in \mathcal{R}_u} \big[\log(\alpha_0) - \log(\alpha_k)\big]. \tag{4}$$

Unlike Eq. (3), this form uses log rather than ψ; it is the exact marginal likelihood with no digamma approximation. When $|\mathcal{R}_u| = 1$ it reduces to the single-label EDL Type-II loss [14].

KL regularisation. Both BM and EDL share the same regularisation strategy, which we generalise from the single-label setting to the multi-label collaborative filtering case. In the original formulations, Sensoy et al. [14] and Joo et al. [8] remove the evidence of the single target class before computing the KL divergence, so that only non-target evidence is penalised. We extend this to the full positive item set $\mathcal{R}_u$ by constructing the modified concentration vector

$$\tilde{\alpha}_k = \begin{cases} 1 & k \in \mathcal{R}_u, \\ \alpha_k & k \notin \mathcal{R}_u, \end{cases} \tag{5}$$

Table 1. Dataset statistics after preprocessing.

Dataset	#Users	#Items	#Inter	Avg. act./user	Sparsity
MovieLens-1M [6]	6,035	3,126	574,376	95.2	96.96%
Amazon Books [13]	21,924	23,774	1,125,676	51.3	99.78%
Amazon Office [13]	675,356	111,666	913,616	1.35	99.999%

and computing the closed-form KL divergence between $\mathrm{Dir}(\tilde{\boldsymbol{\alpha}})$ and the uniform prior $\mathrm{Dir}(\mathbf{1})$ following [15] (Appendix C.3):

$$\mathrm{KL}\big(\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}) \,\|\, \mathrm{Dir}(\mathbf{1})\big) = \log\Gamma(\tilde{\alpha}_0) - \log\Gamma(K) - \sum_k \log\Gamma(\tilde{\alpha}_k) + \sum_k (\tilde{\alpha}_k - 1)\big[\psi(\tilde{\alpha}_k) - \psi(\tilde{\alpha}_0)\big]. \tag{6}$$

This construction is identical across both methods and is evaluated once per user per forward pass at no additional cost.

Annealing schedules. The two methods use different annealing strategies for the KL weight λ. For BM, we linearly anneal λ from zero to $\lambda_{\max}$ over 50,000 gradient steps, mirroring the KL annealing schedule of the original MultiVAE training [11]. For EDL, we follow the epoch-based schedule proposed by [14]: $\lambda_t = \lambda_{\max} \cdot \min(1, t/10)$, where t is the current epoch index, so that the full regularisation weight is reached by epoch 10. Both reconstruction and KL losses are averaged across the minibatch before the backward pass.

3.4 Inference

At test time, items are ranked in descending order of their Dirichlet posterior mean

$$\bar{p}_k = \frac{\alpha_k}{\alpha_0}, \tag{7}$$

which is the score used by both models. Because $\alpha_0 = \sum_k \alpha_k$ is constant across all items for a given user, this ranking is equivalent to sorting by raw α_k; the denominator cancels and leaves the ordering unchanged.

4 Experimental Setup

Datasets. We evaluate our methodology on three datasets with different density profiles (Table 1). We binarise interactions by retaining ratings ≥ 4 as [11]. On MovieLens-1M we retain users and items with at least 5 interactions, and on Amazon Books at least 20 interactions per user and per item; Amazon Office Products is used without interaction-count filtering.

Models. We compare three models sharing the identical MultiVAE backbone (hidden layers: 600, latent dim: 128, dropout: 0.5):

– **MostPop**: Non-personalised baseline that ranks items by global popularity, trained with RecBole [17].

Table 2. Performance evaluation on MovieLens-1M, Amazon Books, and Amazon Office Products.

Dataset	Model	NDCG @10	NDCG @20	Rec. @10	Rec. @20	Cov. @20	Gini @20	AvgPop @20	Tail @20
MovieLens-1M	MostPop	0.1311	0.1330	0.0754	0.1242	0.0489	0.9889	1894.6	0.0000
	MultiVAE	0.1998	0.2258	0.2006	0.2999	**0.6599**	**0.8543**	**839.4**	**0.0012**
	MultiVAE-BM	0.2146	0.2362	0.2063	0.3034	0.5761	0.8852	918.3	0.0001
	MultiVAE-EDL	**0.2199**	**0.2410**	**0.2069**	**0.3038**	0.5816	0.8874	918.6	0.0005
Amazon Books	MostPop	0.0059	0.0077	0.0077	0.0130	0.0012	0.9995	811.2	0.0000
	MultiVAE	0.0708	0.0868	0.0893	0.1359	0.9150	**0.6598**	**122.3**	**0.0439**
	MultiVAE-BM	**0.0783**	**0.0941**	**0.0958**	**0.1427**	**0.9211**	0.6905	131.4	0.0158
	MultiVAE-EDL	0.0748	0.0904	0.0925	0.1389	0.8552	0.7352	150.0	0.0150
Amazon Office	MostPop	0.0091	0.0112	0.0181	0.0265	0.0002	0.9998	1638.0	0.0000
	MultiVAE	0.0192	0.0222	0.0311	0.0430	**0.2934**	**0.9662**	**384.0**	**0.0027**
	MultiVAE-BM	**0.0227**	**0.0257**	**0.0366**	0.0483	0.2218	0.9797	488.7	0.0025
	MultiVAE-EDL	0.0198	0.0235	0.0342	**0.0488**	0.2388	0.9819	661.8	0.0020

- **MultiVAE** [11]: Baseline, trained natively with RecBole [17].
- **MultiVAE-BM**: MultiVAE backbone configured with post-hoc BM fine-tuning.
- **MultiVAE-EDL**: MultiVAE backbone configured with post-hoc EDL fine-tuning.

Hyperparameters and Evaluation. All fine-tuned models are optimised with Adam, batch size 512, a warmup of $E_w = 5$ epochs and $E = 30$ main-phase epochs; the per-component learning rates and the KL annealing schedules are specified in Section 3. We conduct all experiments using RecBole [17], applying seed 2020, full-ranking evaluation protocols, and a cut-off of $k \in \{10, 20\}$. We compute accuracy metrics (NDCG@k, Recall@k) and beyond-accuracy metrics [10] (ItemCoverage, GiniIndex, AveragePopularity, TailPercentage).

5 Results

Table 2 reports accuracy and beyond-accuracy results for all four models. Both fine-tuned models exceed the MultiVAE baseline on every accuracy metric across the three datasets, while the non-personalised MostPop baseline is the lowest-scoring model on accuracy throughout. MultiVAE-EDL is the strongest on MovieLens-1M, improving NDCG@10 by +10.06% over MultiVAE, whereas MultiVAE-BM is the strongest on the two Amazon datasets, with NDCG@10 gains of +10.59% on Amazon Books and +18.23% on Amazon Office Products; the sole exception is Recall@20 on Amazon Office, where EDL is marginally higher.

On beyond-accuracy metrics, both fine-tuned models move toward popular items relative to MultiVAE: Gini@20 and AveragePopularity@20 rise and TailPercentage@20 falls on all three datasets, with the largest shift on the sparsest corpus (AveragePopularity@20 increases by roughly 9% on MovieLens-1M and by 72% for

MultiVAE-EDL on Amazon Office Products). MultiVAE-BM stays closer to the backbone than MultiVAE-EDL on Gini@20 and AveragePopularity@20 on every dataset, and is the only fine-tuned model to improve Item Coverage@20 over the baseline (+0.67% on Amazon Books). Across all beyond-accuracy metrics, the personalised models are less concentrated on popular items than MostPop, with lower Gini@20 and AveragePopularity@20 and higher Item Coverage@20 and TailPercentage@20.

6 Discussion

6.1 Dirichlet Concentration Analysis

To validate the mechanism driving beyond-accuracy outcomes, we compute the mean Dirichlet concentration per item across all test users,

$$\bar{\alpha}_k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \alpha_k^{(u)}. \tag{8}$$

Because items are ranked by the posterior mean $\bar{p}_k = \alpha_k / \alpha_0$, and α_0 is constant per user, $\bar{\alpha}_k$ directly quantifies the evidence the model assigns to item k on average. A value of $\bar{\alpha}_k = 1$ marks the prior floor: the evidence head has learned no item-specific discrimination for that item.

Figure 2 reveals two operating regimes separated by interaction density, with Amazon Books occupying the transition between them.

MovieLens-1M. On the densest corpus (average 95.2 interactions per user), both models establish a strict, monotonic popularity gradient ($\rho_{\text{BM}} = -0.989$, $\rho_{\text{EDL}} = -0.942$). The most popular items reach $\bar{\alpha}_k \approx 5.1$ (BM) and $\bar{\alpha}_k \approx 2.0$ (EDL), while tail items converge to the prior floor of $\bar{\alpha}_k \approx 1$ at roughly rank 1,000. BM's gradient is both steeper and higher in absolute value, reflecting its digamma-based reconstruction term's stronger pull on high-evidence items relative to the log-based EDL objective. The steeper BM gradient directly explains its larger AveragePopularity increase compared to EDL.

Amazon Books. On the moderately sparse corpus (average 51.3 interactions per user), the gradient compresses substantially. BM peaks at $\bar{\alpha}_k \approx 1.48$ and EDL at $\bar{\alpha}_k \approx 1.11$. A structurally distinct feature appears for EDL: its smoothed curve remains nearly flat from rank 1 through approximately rank 50 before declining, whereas BM begins declining monotonically from rank 1. This plateau means that EDL assigns nearly equal concentrations across the top-50 items rather than concentrating evidence on a single most-popular item. Recommendations under EDL therefore spread more evenly within the popular head, touching a broader set of items at the cost of differentiating less among them. The strength of the popularity signal, as measured by $|\rho|$, is lower for both models than on MovieLens-1M ($\rho_{\text{BM}} = -0.879$, $\rho_{\text{EDL}} = -0.738$), confirming that data sparsity attenuates the gradient.

Amazon Office Products. On the sparsest corpus (average 1.35 interactions per user), the evidence head learns effectively no item-level discrimination. EDL

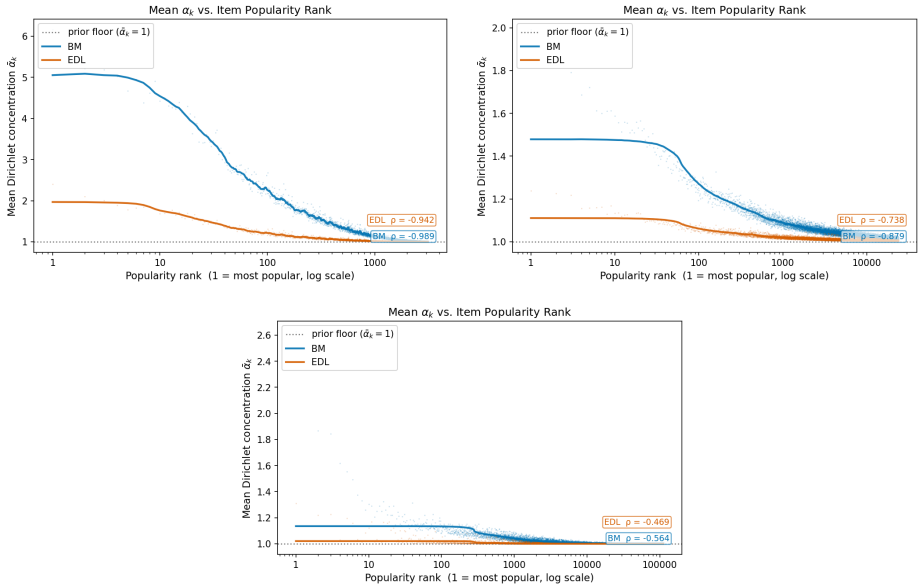


Fig. 2. Mean Dirichlet concentration $\bar{\alpha}_k$ plotted against item popularity rank (log scale; rank 1 = most popular). Smoothed curves are uniform-window moving averages; scatter points are individual items. Spearman correlations ρ between rank and $\bar{\alpha}_k$ are annotated per model. (a) MovieLens-1M (b) Amazon Books and (c) Amazon Office Products.

produces $\bar{\alpha}_k \approx 1.0$ for virtually the entire catalogue ($\rho_{\text{EDL}} = -0.469$). BM retains a marginal signal among the top 200 most popular items, with $\bar{\alpha}_k \approx 1.13$ at rank 1, but this signal decays rapidly and is absent across the long tail ($\rho_{\text{BM}} = -0.564$). The weak and noisy correlations confirm that the evidence head has essentially collapsed to the Dirichlet prior.

The $\bar{\alpha}_k$ gradient has a direct and analytically derivable consequence for epistemic uncertainty U_E. From [15], the per-item contribution of item k to the user-level epistemic uncertainty $U_E^{(u)}$ is the k-th summand before summing over items:

$$\text{ue}_k^{(u)} = -\bar{p}_k^{(u)} \cdot \left(\log \frac{\alpha_k^{(u)}}{\alpha_0^{(u)}} - \psi\left(\alpha_k^{(u)} + 1\right) + \psi\left(\alpha_0^{(u)} + 1\right) \right). \tag{9}$$

Because each item's contribution is weighted by $\bar{p}_k$, the concentration gradient in $\bar{\alpha}_k$ carries over directly to U_E. On MovieLens-1M, the mean of ue_k across users decreases monotonically with popularity rank for both models ($\rho_{\text{BM}} = -0.986$, $\rho_{\text{EDL}} = -0.957$): popular items concentrate the most probability mass and therefore receive the highest per-item U_E contributions, while tail items contribute least. This direction is set by the $\bar{p}_k$ weighting rather than by the bracketed term in Eq. (9). As an item becomes more popular its concentration α_k grows, which has two competing effects: the prefactor $\bar{p}_k = \alpha_k/\alpha_0$ rises, while the log-ratio $\log(\alpha_k/\alpha_0)$ becomes less negative and on its own would shrink the contribution.

The $\bar{p}_k$ term dominates, so popular items still receive the largest U_E contributions overall.

A second-order effect distinguishes the two objectives. BM produces *lower* mean per-item U_E than EDL across the entire rank range (approximately 1.25×10^{-4} versus 1.40×10^{-4} at rank 1), despite its higher absolute concentrations. Because BM's digamma-based reconstruction term accumulates larger α_k, the $\psi(\alpha_k + 1)$ term grows and compresses the parenthesised expression in Eq. (9); BM thus suppresses per-item epistemic uncertainty more aggressively than EDL's log-based objective, at the cost of a steeper $\bar{\alpha}_k$ gradient. The same BM/EDL asymmetry therefore appears in both quantities, with opposite sign: BM shows the higher, steeper $\bar{\alpha}_k$ curve in Figure 2 (more accumulated evidence) yet the lower mean ue_k, because the stronger evidence accumulation that raises $\bar{\alpha}_k$ is precisely what suppresses U_E.

A second-order effect distinguishes the two objectives: BM produces *lower* mean per-item U_E than EDL across the entire rank range (approximately 1.25×10^{-4} versus 1.40×10^{-4} at rank 1), despite its higher absolute concentrations. The mechanism is that BM's larger α_k inflates the $\psi(\alpha_k+1)$ term and compresses the bracketed expression in Eq. (9), so BM suppresses per-item uncertainty more aggressively than EDL's log-based objective. The two analyses are thus consistent rather than contradictory: the same evidence accumulation that gives BM the higher, steeper $\bar{\alpha}_k$ curve in Figure 2 is precisely what drives its U_E lower.

6.2 Why Dirichlet Fine-Tuning Amplifies Popularity

The asymmetry has a simple origin in how often each item is reinforced versus penalised. The reconstruction term rewards high α_k for an item only when it is a positive in the minibatch, whereas the KL term (applied to $\tilde{\boldsymbol{\alpha}}$, Eq. 5) penalises it whenever it is *not*. An item's fate therefore depends on its global interaction count c_k. A popular item appears as a positive in a large fraction of minibatches, so reconstruction repeatedly pushes its α_k up; a long-tail item ($c_k \ll |\mathcal{U}|$) is a non-target almost every time, so the KL term pulls it toward the prior while reconstruction rarely intervenes. The net gradient scales monotonically with c_k, accumulating evidence on popular items and suppressing niche ones — exactly the pattern the strong Spearman correlations in Figure 2 confirm on dense data.

This structural mechanism parallels the mislabelling problem identified by Wang et al. [16]: the asymmetry between observed and unobserved interactions systematically disadvantages long-tail items. While Wang et al. upweight uncertain negatives using aleatoric uncertainty estimates, the Dirichlet KL regulariser penalises all non-target items identically, remaining blind to whether an item is genuinely irrelevant or merely unobserved due to exposure bias.

7 Conclusion

We proposed a post hoc Dirichlet fine-tuning protocol for VAE-based collaborative filtering and characterised its accuracy–popularity trade-off across datasets

spanning three orders of magnitude in interaction density. Empirically, both EDL and BM objectives consistently improved recommendation accuracy, yielding NDCG@10 gains of up to +18.23% relative to the MultiVAE baseline without requiring modifications to the underlying architecture. Importantly, these improvements proved robust across sparsity regimes, suggesting that post hoc Dirichlet objectives can serve as a broadly effective enhancement strategy for VAE-based recommenders. However, these gains are not without cost. Improved ranking performance was accompanied by increased recommendation concentration. The Gini index rose, average recommendation popularity increased, and item coverage declined in five of six configurations (by up to 24%), indicating a shift toward head-item exposure. This shift nonetheless remained far less pronounced than the non-personalised MostPop baseline, whose popularity concentration the fine-tuned models stayed well clear of on every beyond-accuracy metric, confirming that they preserve personalisation. These findings demonstrate that the benefits of Dirichlet fine-tuning must be evaluated not only in terms of ranking accuracy but also through the lens of beyond-accuracy objectives.

This trade-off arises from a structural asymmetry inherent to the Dirichlet KL regulariser. Because the objective penalises evidence assigned to non-target items, popular items (frequent positives) continuously accumulate evidence, whereas tail items (which act almost exclusively as non-targets) collapse toward the uniform prior. This asymmetric behaviour is not an incidental artefact, but rather an intrinsic property of the Dirichlet objective when operating under skewed interaction distributions.

In practice, the resulting $\bar{\alpha}_k$ distribution serves as a zero-cost diagnostic tool at deployment, isolating the source of beyond-accuracy degradation. A flat, catalogue-wide profile indicates a regime dominated by the decoder prior; here, the performance degradation is inherited directly from the backbone architecture, leaving the evidence head with no corrective capacity. Conversely, a steep $\bar{\alpha}_k$ curve signals an evidence-driven regime. In this case, the fine-tuning process itself concentrates evidence onto popular items, meaning the observed coverage loss is a direct consequence of the objective function.

Future work will focus on mitigating this diversity degradation without compromising recommendation accuracy. Promising directions include constraining the $\bar{\alpha}_k$ gradient during training, integrating popularity-aware debiasing techniques such as inverse propensity scoring [1], and extending these diagnostics to sequential and graph-based architectures, where distinct latent structures may alter how Dirichlet objectives interact with popularity bias.

Acknowledgements. Bapi Dutta gratefully acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities through the Ramón y Cajal Research Grant (RYC2023-045020-I). Mateus Toledo acknowledges support through a predoctoral contract from the University of Jaén under the Research and Knowledge Transfer Action Plan, Action R2.c (PIFPP_RYC_2025_2), associated with the second author's Ramón y Cajal Research Grant.

References

1. Abdollahpouri, H., Burke, R., Mobasher, B.: Controlling popularity bias in learning-to-rank recommendation. In: RecSys '17, pp. 42–46. ACM (2017)
2. Abdollahpouri, H., Burke, R., Mobasher, B.: Managing popularity bias in recommender systems with personalized re-ranking. In: FLAIRS-32, pp. 413–418. AAAI Press (2019)
3. Balsells-Rodas, C., Yang, F., Huang, Z., Gao, Y.: Explainable uncertainty attribution for sequential recommendation. In: SIGIR '24, pp. 2401–2405. ACM (2024)
4. Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., Zanzonelli, D.: Fairness in recommender systems: Research landscape and future directions. User Model. User-Adap. Int. (2023)
5. Fan, Z., Liu, Z., Wang, S., Zheng, L., Yu, P.S.: Modeling sequences as distributions with uncertainty for sequential recommendation. In: CIKM '21, pp. 3019–3023. ACM (2021)
6. Harper, F.M., Konstan, J.A.: The MovieLens datasets: History and context. ACM Trans. Interact. Intell. Syst. **5**(4), 19 (2015)
7. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. **110**(3), 457–506 (2021)
8. Joo, T., Chung, U., Seo, M.G.: Being Bayesian about categorical probability. In: ICML '20, pp. 4950–4961. PMLR (2020)
9. Joo, W., Lee, W., Park, S., Moon, I.C.: Dirichlet Variational Autoencoder. arXiv preprint arXiv:1901.02739 (2019)
10. Kaminskas, M., Bridge, D.: Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives. ACM Trans. Interact. Intell. Syst. **7**(1), 1–42 (2016)
11. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: WWW '18, pp. 689–698. ACM (2018)
12. Liang, S., et al. 2024. A Survey on Variational Autoencoders in Recommender Systems. ACM Comput. Surv. 56, 10, Article 268 (October 2024), 40 pages. https://doi.org/10.1145/3663364
13. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR '15, pp. 43–52. ACM (2015)
14. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: NeurIPS '18, pp. 3183–3193 (2018)
15. Ulmer, D.T., Hardmeier, C., Frellsen, J.: Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. Trans. Mach. Learn. Res. (2023)
16. Wang, C., et al.: Rethinking missing data: Aleatoric uncertainty-aware recommendation. IEEE Trans. Big Data **9**(6), 1607–1619 (2023)
17. Zhao, W.X., et al.: RecBole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In: CIKM '21, pp. 4653–4664. ACM (2021)
18. Zhu, Z., Wang, J., Caverlee, J.: Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: SIGIR '20, pp. 449–458. ACM (2020)