Towards Trustworthy Recommender Systems through Uncertainty Quantification

Mateus Toledo da Silveira Leao, Bapi Dutta, and Luis Martínez

University of Jaén, Jaén, Spain
mtleao@ujaen.es

PhD start date: January, 2026. Expected defence: January 2030.
Thesis director: Dr. Bapi Dutta; tutor: Prof. Luis Martínez.

Abstract. Recommender systems are increasingly expected not only to predict user preferences accurately, but also to signal when those predictions should be trusted. A key conceptual tool for this purpose is the distinction between aleatoric uncertainty (AU, the irreducible variability of user behaviour) and epistemic uncertainty (EU, which arises from limited data and model knowledge and is in principle reducible). My research starts from the observation that the assumptions under which this distinction is operationalised in dense benchmark domains do not transfer cleanly to recommendation, where interaction data is extremely sparse and every item effectively forms its own class. Two empirical studies I conducted show that Dirichlet-based signals in both sequential recommendation and collaborative filtering track item popularity rather than genuine model ignorance. On a MultiVAE-based model the concentration parameter correlated with item popularity rank at $\rho = -0.989$ (Spearman), and the same method applied on a SASRec-based model improved long-tail exposure by 108% on ultra-sparse data while degrading it by 28% on dense data. Building on this diagnosis, and rather than assuming that a clean separation is attainable in this regime, my research takes a broader view of Uncertainty Quantification in Recommender Systems: it asks what drives uncertainty in this setting, examines its entanglement with item popularity and the conditions under which a meaningful aleatoric/epistemic separation is achievable and investigates how uncertainty can be exploited beneficially. To this end, I will study second-order tools such as credal sets and conformal prediction, and examine whether reliable uncertainty signals can naturally improve exposure of niche items without requiring explicit fairness constraints. My goal is to provide both practical tools and a reproducible evaluation standard for uncertainty quantification in recommender systems.

Keywords: Uncertainty Quantification · Recommender Systems · Credal Sets · Epistemic Uncertainty · Aleatoric Uncertainty · Fairness · Sequential Recommendation

1 Introduction and Background

Recommender systems predict which items a user is most likely to engage with next and rank them accordingly. While most research optimises for ranking accuracy, an equally important but less studied question is how confident the model should be in each prediction—and, crucially, why it is or is not confident. This is the domain of uncertainty quantification (UQ) in recommender systems [6].

There are two qualitatively different uncertainty sources usually considered in Machine Learning models. Aleatoric uncertainty (AU) reflects genuine randomness in user behaviour, for instance, a user whose interactions span unrelated item categories is inherently harder to predict, and no additional data will remove this difficulty. Epistemic uncertainty (EU), by contrast, reflects a lack of information. So, a user who has interacted with very few items is difficult to predict simply because the model has not observed enough of their behaviour, and this difficulty can be reduced as more interactions accumulate [6].

The practical implications of this distinction are significant. For example, high AU suggests the model should focus recommendations on the narrowed space-item preferences of the user; high EU suggests it should instead diversify before committing to a narrow ranking.

Standard probabilistic predictors map user histories to a single distribution over items, $h : \mathcal{X} \to \mathfrak{P}(\mathcal{I})$. This first-order prediction captures AU because the spread of the distribution reflects preference variability. But it cannot represent EU, since a single distribution leaves no room to express uncertainty about which distribution is correct. A second-order predictor $H : \mathcal{X} \to \mathfrak{Q}(\mathcal{I})$ maps each user to a set of distributions rather than to a single one [7]. This set of distributions could also be called a credal set. The size of this set encodes EU: a small set means the model is confident about which distribution is correct. The internal structure of the set encodes AU independently because even a singleton credal set may contain a high-entropy distribution, reflecting irreducible preference variability.

The standard approach to handling multiple hypotheses is Bayesian model averaging. It usually computes a weighted average over candidate distributions and uses the result as the prediction [6]. This collapses second-order information back to first order, averaging out the epistemic component in the process.

Hüllermeier and Waegeman [7] illustrate the problem with a very simple example: a fair coin and a coin of completely unknown bias both yield a first-order probability of $1/2$ for heads, yet they represent entirely different epistemic states. A second-order predictor can distinguish them; a first-order predictor cannot. In recommendation, this ambiguity is consequential. For instance, a system that cannot tell whether a user is genuinely hard to predict or simply under-observed will respond identically in both cases, even though the appropriate actions differ fundamentally.

We began by examining existing Dirichlet-based UQ methods, which approximate second-order prediction by placing a distribution over probability distributions. The idea is to analyse whether they successfully capture the AU/EU distinction in recommender systems. Initially we developed two empirical studies to address this question, and both reach the same conclusion: they do not.

The mechanism behind this, and the theoretical reason why it resists correction within the Dirichlet framework, motivate the next phase of this research. Whether this failure is specific to the Dirichlet parameterisation or structural to the recommendation regime—with its extreme sparsity and item-as-class label space—is one of the questions that shape this thesis.

2 Research Questions

The central research question of this PhD is: *What drives uncertainty in recommender systems; under which conditions can aleatoric and epistemic uncertainty be meaningfully separated in this regime; and how can uncertainty—whether disentangled or expressed as a single total uncertainty (TU) signal—be exploited to make recommendation more trustworthy and fairer, without requiring explicit fairness constraints?*

Rather than assuming that a clean AU/EU separation is attainable, I take a broader view of UQ in recommendation, in line with my research plan. I hypothesise that (i) the extreme sparsity and item-as-class label space of recommendation structurally entangle uncertainty estimates with item popularity; (ii) under these conditions, TU may be a more robust and useful signal than a strict AU/EU decomposition, a meaningful separation being achievable only under specific conditions that have yet to be characterised; and (iii) reliable uncertainty signals, disentangled or total, can be exploited beneficially, supporting trustworthiness and fairer long-tail exposure. Second-order tools—Dirichlet models, credal sets, ensembles, and conformal prediction—are the instruments of this investigation, and four research questions organise the work.

RQ1: Defining and Evaluating Meaningful UQ in Recommender Systems. Before proposing a new method, we need a testable standard for what "good" UQ looks like. This sub-question formalises two evaluation criteria. The first is *behavioural coherence*: a model should report high EU for cold-start users with very short interaction histories and low EU for users with abundant, consistent evidence. Likewise, AU should be high for users whose histories reflect diverse and seemingly unrelated preferences, and low for users with highly consistent interaction patterns. The second criterion is *calibrated error prediction*: EU scores should be informative about how likely the model is to produce an error for each user. Users assigned high EU should yield measurably lower ranking accuracy at evaluation time than users assigned low EU. These criteria will account for both a disentangled AU/EU representation and a single TU signal, and—motivated by the diagnosis of our preliminary studies—any candidate signal must remain informative after controlling for item popularity, so that popularity proxies are not mistaken for uncertainty.

RQ2: Credal Sets for Sequential Recommendation. This sub-question applies the credal set framework to sequential recommendation, using SASRec [11] as the base model and evaluating on the same datasets as the preliminary studies

to allow direct comparison. We test whether EU is higher for cold-start users than for users with rich histories. And AU will be higher for users with diverse item attribute profiles. We also intend to examine whether the density-dependent fairness reversal observed in prior work disappears under the credal set construction, which would confirm that the earlier effect was an artifact of the Dirichlet parameterisation rather than a fundamental property of data sparsity.

RQ3: Scalable Credal Set Construction. Item catalogues in practice contain millions of entries, so construction strategies must be computationally tractable. This sub-question investigates two approaches. The first is ensemble-based: train K independent instances of the base recommender and define the credal set for each user as the convex hull of the K ranking distributions those instances produce. The second approach uses conformal prediction [9]: given a single trained model and a held-out calibration set, construct for each user the smallest set of distributions that contains the true preference distribution at a specified coverage level. Both approaches treat the base model as a black box and are compatible with any existing architecture, including transformer-based sequential models.

RQ4: Uncertainty Estimation and Long-Tail Fairness. A well-documented problem in recommender systems is that popular items receive a disproportionate share of recommendations, leaving niche items chronically underexposed even when they would be highly relevant for specific users [1, 5]. An alternative hypothesis, supported by recent work on uncertainty and fairness [18, 4, 19], is that unfairness in ranking arises primarily from uncertainty about item merits: when a system is uncertain about a niche item's relevance (because it has few interactions), it defaults to safe, popular choices. If that uncertainty is properly represented and separated from aleatoric preference variability, niche items that genuinely fit a user's interaction profile should receive higher estimated relevance bounds within the credal set and appear in recommendations without any popularity correction. Tahir et al. [19] showed that explicitly decoupling AU and EU improves the fairness-utility trade-off without dedicated fairness objectives in classification tasks; we investigate whether the same mechanism operates in recommendation.

3 Research Progress

Research conducted since January 2026 has focused on understanding the behaviour of existing Dirichlet-based UQ methods in recommender systems, resulting in two submitted studies.

3.1 Reproducibility and Fairness Study of ExUA (RQ1, RQ2)

We investigated Explainable Uncertainty Attribution (ExUA) [2], a recently proposed framework for uncertainty-aware sequential recommendation published at SIGIR 2024. ExUA integrates gradient-based saliency with post-hoc Dirichlet

training (Evidential Deep Learning and Belief Matching) and reported accuracy improvements exceeding 6% NDCG@20. To determine whether the attribution mechanism drove these gains, we conducted a rigorous multi-seed ablation study decoupling the attribution logic from the training schedule.

Replacing the attribution with uniform weights or disabling it entirely left all metrics statistically indistinguishable from the full method across 27 tested configurations. The performance gains originated entirely from the post-hoc training objective. Extending the evaluation to fairness and diversity metrics across datasets of varying density, we found a clear pattern: the Dirichlet UQ method improved long-tail exposure by 108.1% (ACLT@20) on ultra-sparse data (1.37 interactions per user) while degrading it by 28.0% on dense datasets (165 interactions per user). This reversal indicates that the method's uncertainty signal responds to the statistical structure of the interaction data rather than to the model's knowledge about individual users [13].

Key Contributions: (i) Multi-seed ablation demonstrating that ExUA's attribution mechanism does not drive uncertainty reduction or accuracy gains. (ii) To the best of our knowledge, the first beyond-accuracy evaluation of UQ-enhanced sequential recommendation across data-density regimes. (iii) Open-source code, data, and hyperparameter configurations. Submitted to an international conference (2026).

3.2 Post-Hoc Dirichlet Fine-Tuning of VAEs (RQ1, RQ3)

To examine the same problem in a different architecture, we developed a post-hoc fine-tuning protocol for VAE-based collaborative filtering [14]. We attached a lightweight evidence head to a frozen MultiVAE backbone [12] and optimised it under EDL and BM objectives. Both methods improved ranking accuracy consistently across three datasets (+18.23% NDCG@10 on Amazon Office Products) while degrading beyond-accuracy metrics in all cases.

To identify the cause, we analysed the mean Dirichlet concentration $\bar{\alpha}_k$ per item, which quantifies how much evidence the model assigns to item k on average. On dense data, $\bar{\alpha}_k$ correlated with item popularity rank at $\rho = -0.989$ (Spearman): the model concentrated almost all of its evidence on the most popular items, treating the rest as unknown regardless of their relevance to any particular user. On ultra-sparse data, the evidence head learned essentially nothing, leaving the system to rely on the backbone's pre-trained popularity ordering. Both failure modes reflect the same structural cause: the KL regulariser penalises evidence on non-target items in proportion to their frequency in training, which is simply the popularity distribution [14]. The result is a first-order signal masquerading as a second-order one.

Key Contributions: (i) Multi-label generalisation of EDL and BM objectives for the collaborative filtering setting. (ii) Concentration analysis ($\bar{\alpha}_k$ vs. popularity rank) identifying the structural mechanism behind beyond-accuracy degradation. (iii) Characterisation of two operating regimes determined by interaction density. Submitted to CAEPIA 2026.

Ongoing work extends this diagnosis beyond Dirichlet models, applying the RQ1 criteria to ensemble-, credal-, and conformal-based estimators across backbones and density regimes.

4 Methodology and Work Plan

4.1 Proposed Methodology

Both completed studies converge on the same diagnosis: Dirichlet-based methods seem unable to properly distinguish AU from EU in recommender systems because they collapse the model's epistemic state into a single distribution. The theoretical basis for this failure was established by Jürgens et al. [10], who proved that any second-order predictor trained end-to-end by minimising empirical risk degenerates to a point prediction with zero EU; any EU that appears is then controlled by the regulariser in an arbitrary way. Building credal sets externally, by observing how much independently trained models disagree about the same user, sidesteps this impossibility and produces a genuinely second-order signal without requiring changes to the base architecture.

Ensemble-based credal sets. Given K independently trained instances of a base recommender f_θ, the credal set for user u is:

$$Q_u = \mathrm{conv}\big(\{p_{\theta_1}(\cdot \mid x_u), \ldots, p_{\theta_K}(\cdot \mid x_u)\}\big), \tag{1}$$

the convex hull of the ranking distributions the K models produce for that user. EU is quantified as the maximum pairwise divergence among ensemble members using the L-divergence from Sale et al. [16]: $\mathrm{EU}(Q_u) = \max_{p,p' \in Q_u} D_{\mathcal{L}}(p, p')$. When all models agree (a user with a rich, consistent history) Q_u collapses toward a singleton and EU is low. When they disagree (a cold-start user) Q_u is wide and EU is high.

Conformalized credal sets. Training K separate models is expensive for large catalogues. Conformal prediction [9] offers a more efficient alternative: given a single trained model and a held-out calibration set, construct for each user the smallest set of distributions that contains the true preference distribution at a specified coverage probability. The design of the nonconformity score (measuring how surprising each item would be for a given user) is the key methodological contribution of RQ3 and requires no retraining of the base model.

Evaluation. All constructions are evaluated against the criteria from RQ1. Behavioural coherence is verified through controlled user profiles of varying interaction density and item attribute diversity. Calibrated error prediction is assessed by plotting EU deciles against mean NDCG@20 at test time, producing a reliability diagram. In all analyses, item popularity is explicitly controlled for, so that a popularity proxy cannot pass as an uncertainty signal. Beyond-accuracy metrics (ACLT, APLT, PopREO, PopRSP, Gini Index, and Item Coverage [13]) evaluate RQ4.

Table 1. Work plan. P1 and P2 produced in Phase 1 (under peer review).

Period	Phase	Milestones
Jan–Jun 2026	Phase 1: Diagnosis	P1 (ExUA reproducibility) submitted; P2 (MultiVAE Dirichlet study) submitted; credal set baseline implemented on SASRec.
Jul 2026–Sep 2027	Phase 2: Core	P3: credal sets for sequential recommendation (RQ2+RQ3) submitted; P4: fairness emergence study (RQ4) submitted.
Oct 2027–Jun 2029	Phase 3: Extension	P5: credal sets for generative retrieval architectures (RQ4 extension) submitted; P6: efficiency study comparing ensemble and conformal constructions at scale submitted.
Jan 2026–Jan 2030	Phase 4: Thesis	Thesis writing, revision, and defence.

4.2 Work Plan

The PhD is structured across four phases, summarised in Table 1. Phase 1, completed in June 2026, established the empirical diagnosis through the two submitted studies. Phases 2 and 3 develop and validate the proposed second-order constructions; Phase 4 consolidates the work into the thesis.

5 Future Directions

5.1 Credal Sequential Recommendation (RQ2, RQ3)

The immediate next step is to apply both construction strategies to the SASRec backbone across the datasets used in the preliminary studies, enabling direct comparison with the Dirichlet baselines. The central test is whether $\mathrm{EU}(Q_u)$ satisfies both criteria from RQ1: it should be higher for cold-start users than for users with dense interaction histories, and it should correlate with actual ranking error at test time rather than with item interaction frequency. We will further test whether the density-dependent fairness reversal observed in prior work disappears under the credal construction. If it does, this confirms that the reversal was an artifact of how the Dirichlet regulariser interacts with the popularity distribution, and not a fundamental property of data sparsity in recommender systems. If it does not, the entanglement between uncertainty and popularity generalises across estimator families, in line with my working hypotheses, and total uncertainty will be evaluated as the operational signal. Either outcome is informative for the thesis. We will also examine whether the uncertainty representation should be placed over ranking (preference) space rather than over the full item simplex, since the item-as-class training structure is a suspected source of the entanglement.

5.2 Fairness Emergence (RQ4)

The hypothesis linking uncertainty to fairness has theoretical backing from Singh et al. [18], who argue that uncertainty about item merits is a primary cause of unfairness in ranking, and that explicitly modelling it leads naturally to more equitable exposure. Devic et al. [4] extend this to two-sided markets, showing that recognising uncertainty simultaneously diagnoses and addresses unfairness without external fairness constraints. In the recommendation setting, for a user whose interaction history coherently points to a specific item type, the credal set should be narrow and positioned within that type. Ranking by the minimum plausible relevance across the credal set then surfaces items in that type without any popularity penalty. This approach differs from Tahir et al. [19], who improve fairness in classification by intervening directly in the loss function; we instead use the structure of the credal set at inference time, leaving training unchanged. We evaluate using beyond-accuracy metrics from prior work [13, 5] and compare against systems where fairness is enforced externally through re-ranking.

5.3 Extension to Generative Retrieval (Phase 3)

Modern generative recommenders produce suggestions by autoregressively generating item identifiers [15] rather than scoring a fixed catalogue. These systems may generate token sequences that do not correspond to any existing item in the catalogue, and current approaches have no principled mechanism to detect when this is likely. In Phase 3, we investigate whether the credal set framework can be extended to this setting by treating disagreement among ensemble members about the next generated token as a real-time EU signal: when models diverge sharply, EU is high, and the system can fall back to a deterministic retrieval component. This extension is conditional on the results of Phase 2 and is designed as a self-contained contribution rather than a primary objective.

6 Relevance of the Proposed Research

Recommender systems mediate how people discover products, content, and opportunities; recommendations that are erroneous or overconfident can mislead users, reinforce existing biases, and unfairly limit the visibility of certain items and their providers. As these systems are deployed in increasingly complex and high-stakes settings, it is no longer sufficient for them to be merely accurate: they must also be trustworthy, providing an indication of how reliable each individual recommendation is. This is precisely what this research pursues, in line with the transparency and robustness requirements emerging in regulatory frameworks such as the EU AI Act. The conformal construction of RQ3, in particular, requires only a frozen base model and a calibration set, and is therefore designed to be deployable on top of existing production systems without retraining.

The relevance of this research is also not conditional on a single outcome. UQ methods imported into recommendation are typically validated by accuracy uplift alone; our preliminary studies show why this is insufficient, since a method

can improve accuracy while its "uncertainty" is a relabelled popularity statistic. The evaluation criteria developed in RQ1, which explicitly control for item popularity, give the community a reusable validity check that separates genuine uncertainty from popularity artifacts, independently of which construction ultimately prevails. And if a clean AU/EU separation proves unattainable in this regime, characterising the conditions under which it fails—and establishing total uncertainty as the robust operational signal—is itself an actionable result.

Finally, whether the aleatoric/epistemic dichotomy is meaningful, and how faithfully current estimators realise it, is an active debate in machine learning at large [17, 10, 3]. Recommendation is an extreme test bed for this debate—tens of thousands of item classes, extremely sparse implicit feedback, heavy popularity imbalance—so the results of this thesis can delineate boundary conditions that dense benchmarks cannot expose, connecting recommendation to the foundations of imprecise probability [20]. Moreover, because the entanglement we diagnosed is specifically with popularity, getting uncertainty right is also a fairness intervention: errors in uncertainty estimation systematically disadvantage niche items and their providers [1, 24], which makes this research directly relevant to platforms, item providers, and regulators.

7 Conclusion

Two completed empirical studies demonstrate that standard Dirichlet-based UQ methods in recommender systems produce a first-order signal that tracks item popularity rather than model ignorance, a direct consequence of collapsing aleatoric and epistemic uncertainty into a single distribution. Therefore we propose to continue our research using credal sets as a principled second-order representation: one that keeps AU and EU structurally separate, allows the model to express genuine ignorance without committing to an arbitrary prior, and may improve fairness toward niche items without external constraints. If the credal construction satisfies the behavioural coherence and calibrated error prediction criteria developed in this work, it will provide both a practical tool and a reproducible evaluation standard for UQ in recommender systems; if it does not, characterising when and why the aleatoric/epistemic separation fails in this regime—and establishing total uncertainty as the robust alternative—is an equally actionable result. In either case, the evaluation criteria and the characterisation of the recommendation regime constitute contributions the community can apply beyond the specific architectures studied here. Feedback on the construction strategies, the fairness evaluation design, and the feasibility of the generative retrieval extension would be particularly valuable as the thesis moves into its core phase.

References

1. Abdollahpouri, H., Burke, R., Mobasher, B.: Managing popularity bias in recommender systems with personalised re-ranking. In: Proc. FLAIRS-32, pp. 413–418 (2019)

2. Balsells-Rodas, C., Yang, F., Huang, Z., Gao, Y.: Explainable uncertainty attribution for sequential recommendation. In: Proc. SIGIR 2024, pp. 2401–2405 (2024)
3. Bengs, V., Hüllermeier, E., Waegeman, W.: On second-order scoring rules for epistemic uncertainty quantification. In: Proc. ICML 2023 (2023)
4. Devic, S., Kempe, D., Sharan, V., Korolova, A.: Fairness in matching under uncertainty. arXiv:2302.03810 (2023)
5. Duricic, T., Kowald, D., Lacic, E., Lex, E.: Beyond-accuracy: A review on diversity, serendipity and fairness in recommender systems based on graph neural networks. arXiv:2310.02294 (2023)
6. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. **110**(3), 457–506 (2021)
7. Hüllermeier, E., Destercke, S., Shaker, M.H.: Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In: Proc. UAI 2022 (2022)
8. Hüllermeier, E.: On the challenge of quantifying epistemic uncertainty in machine learning. SIPTA Talk, June 2024
9. Javanmardi, A., Stutz, D., Hüllermeier, E.: Conformalized credal set predictors. In: Proc. NeurIPS 2024 (2024)
10. Jürgens, M., Bengs, V., Meinert, N., Hüllermeier, E., Waegeman, W.: Is epistemic uncertainty faithfully represented by evidential deep learning methods? In: Proc. ICML 2024 (2024)
11. Kang, W.C., McAuley, J.: Self-attentive sequential recommendation. In: Proc. ICDM 2018, pp. 197–206 (2018)
12. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: Proc. WWW 2018, pp. 689–698 (2018)
13. Leao, M., Dutta, B., Yera, R., Martinez, L.: Rethinking uncertainty attribution for sequential recommendation: A reproducibility and fairness study. Under review (2026)
14. Leao, M., Dutta, B.: Post-hoc Dirichlet fine-tuning of variational autoencoders for collaborative filtering: Accuracy gains and popularity costs. In: Proc. CAEPIA 2026 (2026)
15. Rajput, S., et al.: Recommender systems with generative retrieval. In: Proc. NeurIPS 2023 (2023)
16. Sale, Y., Bengs, V., Caprio, M., Hüllermeier, E.: Second-order uncertainty quantification: A distance-based approach. In: Proc. ICML 2024 (2024)
17. Sale, Y., Kotelevskii, N., Panov, M., Hüllermeier, E.: The aleatoric-epistemic dichotomy of uncertainty is meaningful and indispensable for machine learning. SSRN preprint 6783338 (2026)
18. Singh, A., Kempe, D., Joachims, T.: Fairness in ranking under uncertainty. In: Proc. NeurIPS 2021 (2021)
19. Tahir, A., Cheng, L., Liu, H.: Fairness through aleatoric uncertainty. In: Proc. CIKM 2023, pp. 2372–2381 (2023)
20. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, London (1991)
21. Wang, C., Feng, F., Zhang, Y., Wang, Q., Hu, X., He, X.: Rethinking missing data: Aleatoric uncertainty-aware recommendation. IEEE Trans. Big Data **9**(6), 1607–1619 (2023)
22. Wang, K., Cuzzolin, F., Moens, D., Hallez, H.: Credal ensemble distillation for uncertainty quantification. Preprint (2026)

23. Zhao, W.X., et al.: RecBole 2.0: Towards a more up-to-date recommendation library. In: Proc. CIKM 2022, pp. 4722–4726 (2022)
24. Zhu, Z., Wang, J., Caverlee, J.: Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: Proc. SIGIR 2020, pp. 449–458 (2020)