# Uncertainty-aware conversational recommendation

Naman Garg

## 1 · PROBLEM

what and why

Cold-start conversational recommenders must infer intent from sparse dialogue, often when the user's request is ambiguous or exploratory. Our VCE study showed that learned expansion helps most in cross-modal, low-coherence sessions, but is not universally superior: on TalkPlayData VCE reaches NDCG@10 = 0.0250 vs. 0.0244 for WPRF ($p = 0.098$), while WPRF dominates text-aligned datasets. We will rebuild VCE from the ground up to model intent uncertainty, learn which evidence modality to trust, and adapt retrieval breadth accordingly.

## 2 · APPROACH

model / method, and why this one

C1 · Distributional intent: infer a distribution over plausible intents rather than a single centroid; jointly learn intent direction and calibrated uncertainty. C2 · Modality routing: dynamically combine lexical, semantic, and multimodal retrieval signals based on modality compatibility and uncertainty. C3 · Adaptive retrieval: high confidence → focused retrieval; moderate uncertainty → multi-hypothesis exploration; high uncertainty → broader retrieval or clarification. C4 · Human evaluation: test whether uncertainty-aware behavior improves recommendation quality and user control during ambiguous requests. The existing vMF formulation is a starting hypothesis; architecture, uncertainty objective, and sampling will be redesigned.

## 3 · DATASET

source, size, preprocessing

TalkPlayData (15,199 train / 1,000 test / 46,579 tracks) as primary cross-modal benchmark; MovieLens-1M, Last.fm 2K, and Amazon Music as text-aligned / sparse controls; plus ≥1M additional samples for large-scale retraining. TalkPlayData uses 512-D CLAP embeddings; other benchmarks use 384-D Sentence-BERT. L2 normalization, leave-one-out evaluation, full-catalog ranking, and anchor exclusion. New data undergoes deduplication, leakage checks, and strict train/validation/test separation.

## 4 · METRICS

how you'll know it worked

how you'll know it worked Retrieval: NDCG@10, Recall@10, MRR. Ambiguity: performance by coherence/uncertainty. Calibration: ECE, Brier, uncertainty–error correlation. Robustness under anchor corruption. Exploration diversity/novelty/coverage. Human study: success, relevance, satisfaction, control. Efficiency: ms/query and memory. Key preliminary signal: VCE already gives +90.3% vs. uniform centroid and +5.4% vs. WPRF in low-coherence TalkPlayData sessions.

## 5 · COMPUTE

training time, memory, dependencies

Staged experimental design: expensive training reserved for the most promising architectures; embeddings, retrieval indices, and multimodal features cached. Data processing + embedding on 1× A100/H100 (~0.5 day); architecture search and final training on GPU with wall-clock measured rather than assumed.

## 6 · DELIVERABLES

model, paper, demo

- · Open-source uncertainty-aware conversational retrieval system.
- · Trained checkpoints and experiment configs.
- · Human-study protocol and results for ambiguous sessions.
- · Research paper documenting methodology, modality routing, and failure cases.

## 7 · RISK

what breaks first, and plan B

what breaks first, and plan B Uncertainty estimates may be poorly calibrated. Plan B: post-hoc calibration and explicit reporting of ECE/Brier before claiming adaptive gains. Gains may be dataset-specific to TalkPlayData. Plan B: require consistent directional improvements on at least one text-aligned control before claiming generality. Human evaluation may disagree with offline metrics. Plan B: treat human preference as the primary ambiguous-session outcome and analyze disagreements.