Multilingual Semantic Retrieval for OTT Streaming Services

Theresia Purnomo

Supervised by Prof. D. Y. Yeung · HKUST CSE

Introduction & Objectives

Problem

OTT streaming platforms serve multilingual audiences, but catalogue metadata is often:

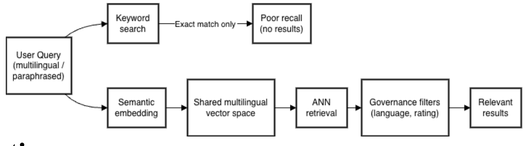
- Sparse
- Inconsistently localised
- English-dominant

As a result, keyword search fails for:

- Cross-language queries
- Paraphrased editorial prompts
- Non-English discovery

The hardest case is *cold-start discovery*, where no user behaviour data exists.

Core Idea: *Multilingual semantic retrieval enables cold-start discovery in a shared embedding space, with governance enforced at retrieval time.*



Motivation

Semantic retrieval is promising, but real deployment requires more than relevance alone:

- Results must respect governance constraints (language, age rating)
- Latency must be predictable for editorial workflows
- Evaluation must be reproducible without behavioural ground truth

This project evaluates and validates these requirements in a realistic OTT retrieval system.

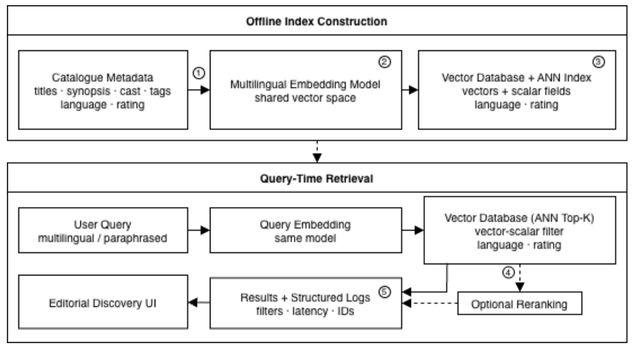
Objectives

- Enable multilingual semantic retrieval using catalogue metadata
- Enforce language and age constraints during retrieval
- Empirically evaluate effectiveness, latency, and limitations under cold-start conditions

System Architecture & Retrieval Pipeline

System Architecture

Retrieval Architecture (Offline → Online)



① Metadata scope defines embedding inputs and governance constraints
② Shared multilingual embeddings align catalogue items and queries
③ Governance is enforced during search via ANN Top-K retrieval
④ Optional reranking adjusts order without violating constraints
⑤ Audit logs ensure traceability and reproducibility

Governance & Auditability

Key Design Principle: Governance at Retrieval Time

- Governance enforced at retrieval time: Language and age limits are applied *before ranking*, preventing invalid results.
- Reproducible evaluation: each request logs query + filters + latency + returned IDs for auditing, debugging, and repeatable experiments

Retrieval Flow

1. Embed catalogue metadata into a shared multilingual space
2. Embed editorial queries using the same model
3. Retrieve Top-K results via ANN with vector-scalar filtering
4. Optionally rerank results (ordering only)
5. Log query, filters, latency, and IDs

Evaluation & Analysis

Evaluation Setup

- **Data:** 12,438 catalogue items (titles, synopses, genres, cast)
- **Languages:** English, Chinese (Traditional), Bahasa Indonesia
- **Queries:** 50 curated multilingual editorial prompts (cold-start)
- **Ground truth:** Silver relevance set + editorial mini-benchmark (N=15)
- **Constraints:** No behavioral data; governance enforced at retrieval time

Retrieval Effectiveness (Cold-Start)

Metrics: Recall@K, MRR@K, NDCG@K
Configurations: ANN retrieval ± reranking (semantic pipeline only)

Configuration	Recall@10	MRR@10	NDCG@10
ANN only	0.36	0.248	0.274
ANN + rerank	**0.44**	**0.295**	**0.318**

Key insight: Under cold-start conditions with no behavioural data, ANN-based multilingual semantic retrieval delivers stable, governance-correct discovery, while reranking improves early precision (↑MRR, ↑NDCG) without violating latency or policy constraints.

Performance & Scalability

- **Latency (ANN)**: p50 308–375 ms; p95 < 1 s at concurrency 20
- **Error rate**: 0% under sustained load
- **Behavior**: Gradual tail-latency degradation (no failure spikes)

Scope & Limitations (Explicit)

- Metadata-only retrieval (no behavioural signals)
- Silver relevance set; no true lexical baseline in deployment
- Claims restricted to system-level behaviour under fixed configuration