Item Identifiability from Large-Language-Models-Based Embeddings in Recommender Systems

Ing. Linda Beková **Supervisor: Dr. rer. nat. Rodrigo Augusto da Silva Alves**

Czech Technical University in Prague — Faculty of Information Technology

Problem & Motivation

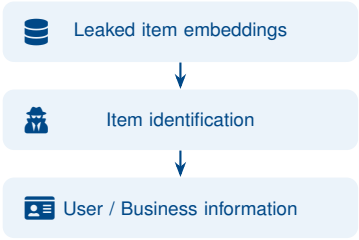
Recommender Systems increasingly use language embeddings to improve recommendation quality. These embeddings:

- Represent items using, e.g., titles, descriptions, and reviews.
- Can be shared instead of explicit item identifiers
- Are often perceived as privacy-preserving

But are these embeddings really privacy-preserving?

- High-dimensional vectors appear difficult to interpret
- Metadata from items are frequently publicly available
- Public metadata can provide attackers with a candidate pool

Potential privacy risk



Main Research Question

Given anonymized item language embeddings and auxiliary/publicly available information, can an attacker recover which real items those vectors represent?

Key Contributions

- **Threat Model:** We formalize item-identification attacks under different levels of attacker knowledge.
- **Item Identification:** We show that anonymized embeddings can reveal item identities using publicly available data.
- **Privacy-Preserving Representation:** We show that reducing identifying information can substantially weaken attacks.
- **User Re-identification:** In an *extended version* of this work, we show that leaked item embeddings, combined with publicly available user interactions, can reveal user identities.

Methodology

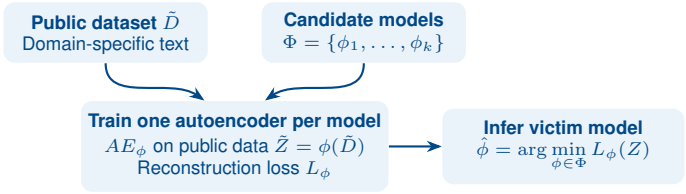
Attacker knowledge

We consider only realistic attack scenarios that can arise in real-world recommender systems. The main assumptions are:

- Access to anonymized item embeddings
- Knowledge of the application domain
- Access to publicly available auxiliary data
- The embedding model may be unknown

Stage 1: Which model generated the embeddings?

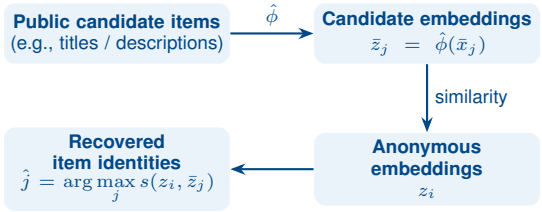
We test whether an attacker can identify the embedding model by training autoencoders on embeddings $\tilde{Z}$ derived from publicly available corpora $\tilde{D}$ and evaluating reconstruction on anonymized embeddings Z.



The model whose learned embedding-space structure best explains the anonymized embeddings is selected. **If the actual model $\phi \in \Phi$, we successfully identify it in all tested scenarios.**

Stage 2: Which items do the embeddings represent?

We then test whether an attacker can recover item identities by matching anonymized embeddings against publicly available candidate items.



Candidate items are embedded with the inferred model $\hat{\phi}$. Their embeddings are compared with the anonymous embeddings, and the most similar candidate is selected as the predicted item identity.

Results

We evaluate the attack on two benchmark recommendation datasets, **MovieLens** and **Goodreads**, across different attacker knowledge scenarios and candidate-set sizes.

Strongest recovery scenario

When the victim and attacker have access to overlapping identifying information (e.g., titles, character names, or cast), item identities can be recovered with very high accuracy, even from large candidate sets.

MovieLens	Goodreads
98%	**84%**
Hit@1	Hit@1
MRR: 0.99	**MRR:** 0.88

Best-performing model with $1{,}000$ victim items and $10{,}000$ public candidates. See Chapter 4 of the thesis for all attack scenarios.

Key takeaways

- **Item embeddings can strongly reveal identity:** in some scenarios, almost all items can be recovered.
- **Identifying text matters:** titles and distinctive details substantially increase attack success.
- **Large candidate pools are not sufficient protection:** high identification rates persist even when the attacker must search among $10{,}000$ possible candidate items.
- **Defence helps, but introduces trade-offs:** removing identity-revealing information substantially reduces attack success. RMSE remains competitive, while extended experiments show degradation in ranking quality.

Research Dissemination

An extended version of this original research work is currently under review at the **3rd Workshop on Evaluating and Applying Recommender Systems with Large Language Models**, co-located with the **35th ACM International Conference on Information and Knowledge Management (CIKM 2026, CORE: A)**.