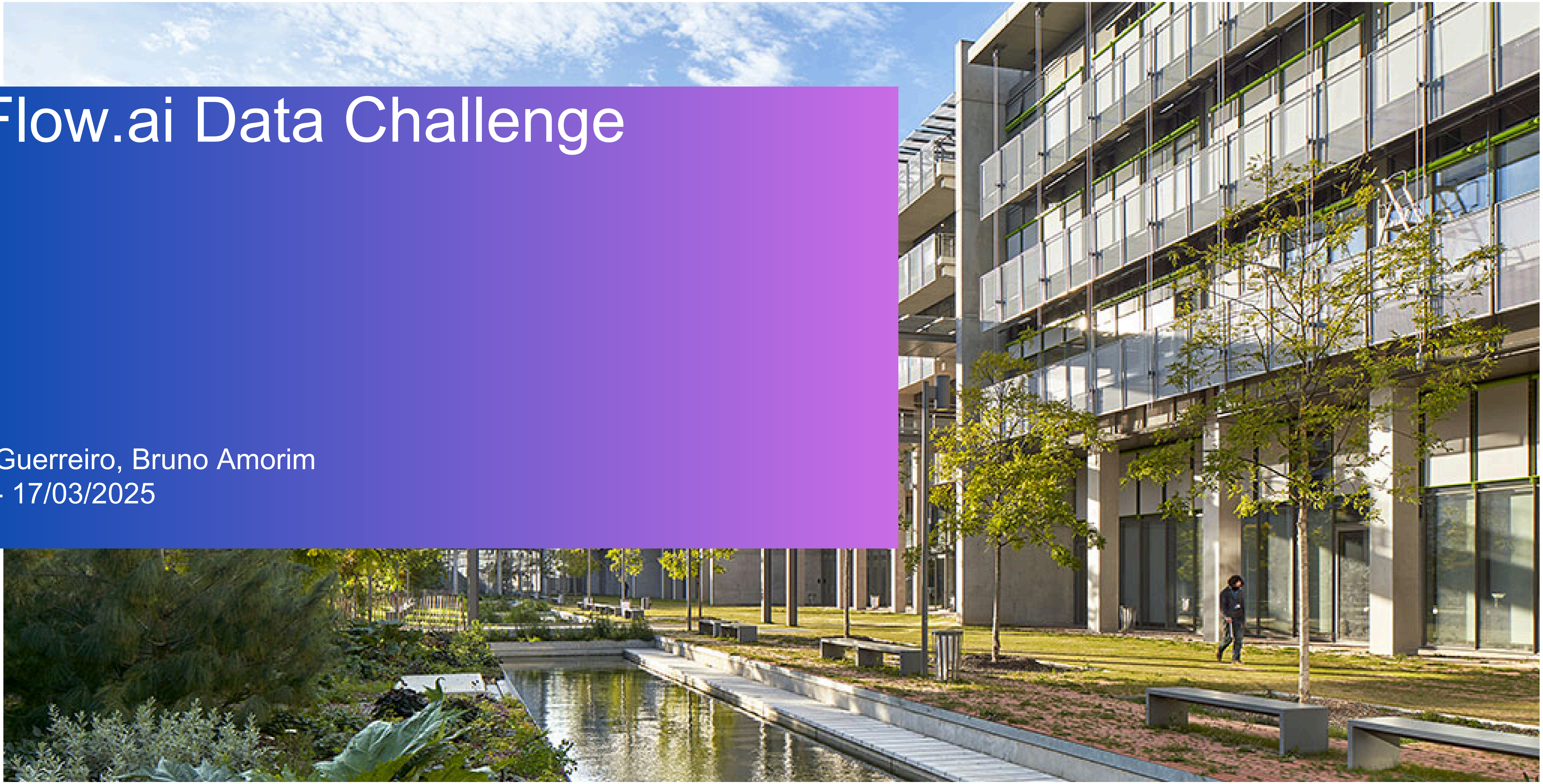
MATHÉMATIQUES
VISION
APPRENTISSAGE

# HrFlow.ai Data Challenge

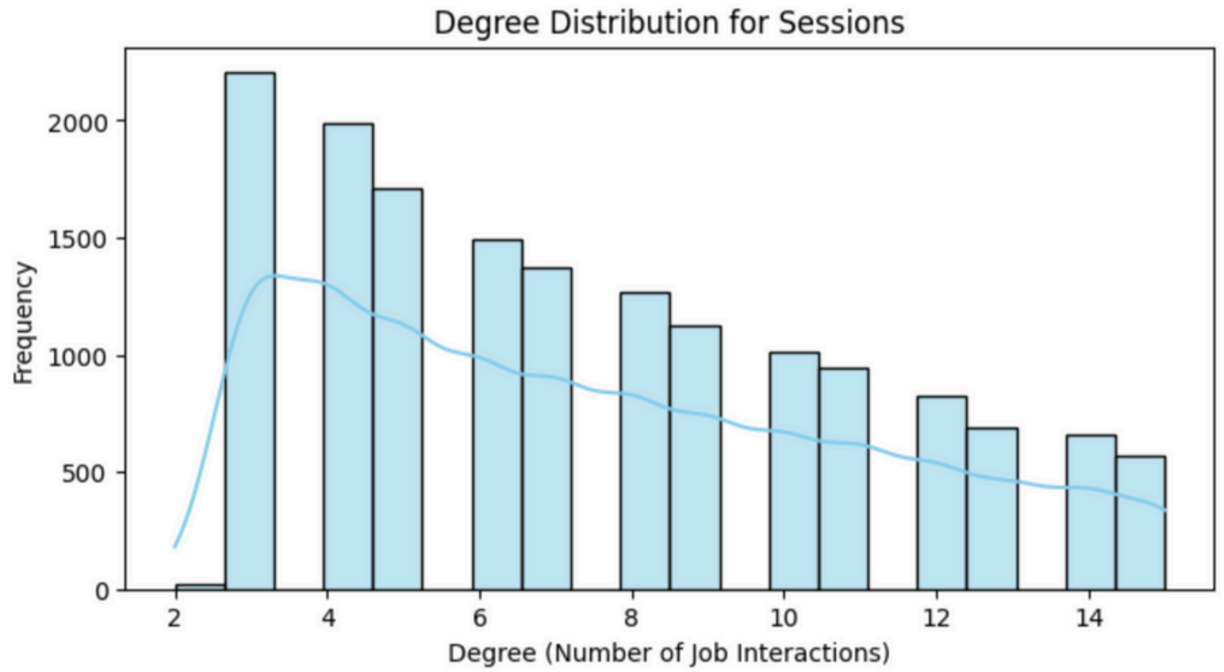
Joao Guerreiro, Bruno Amorim
MVA - 17/03/2025

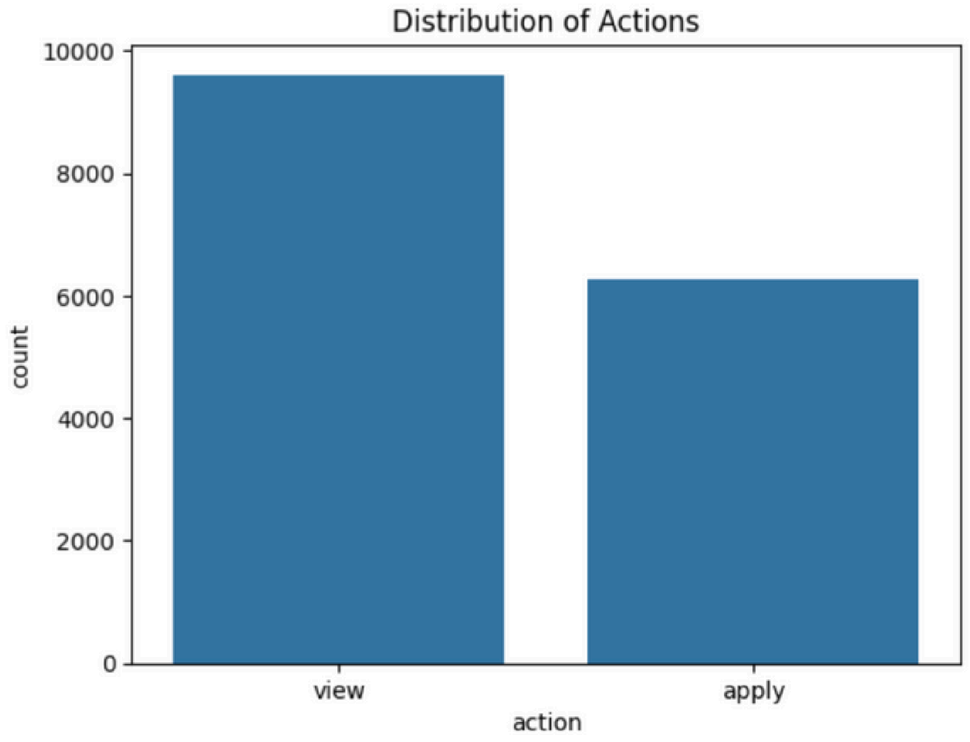


# The challenge

# Data Overview

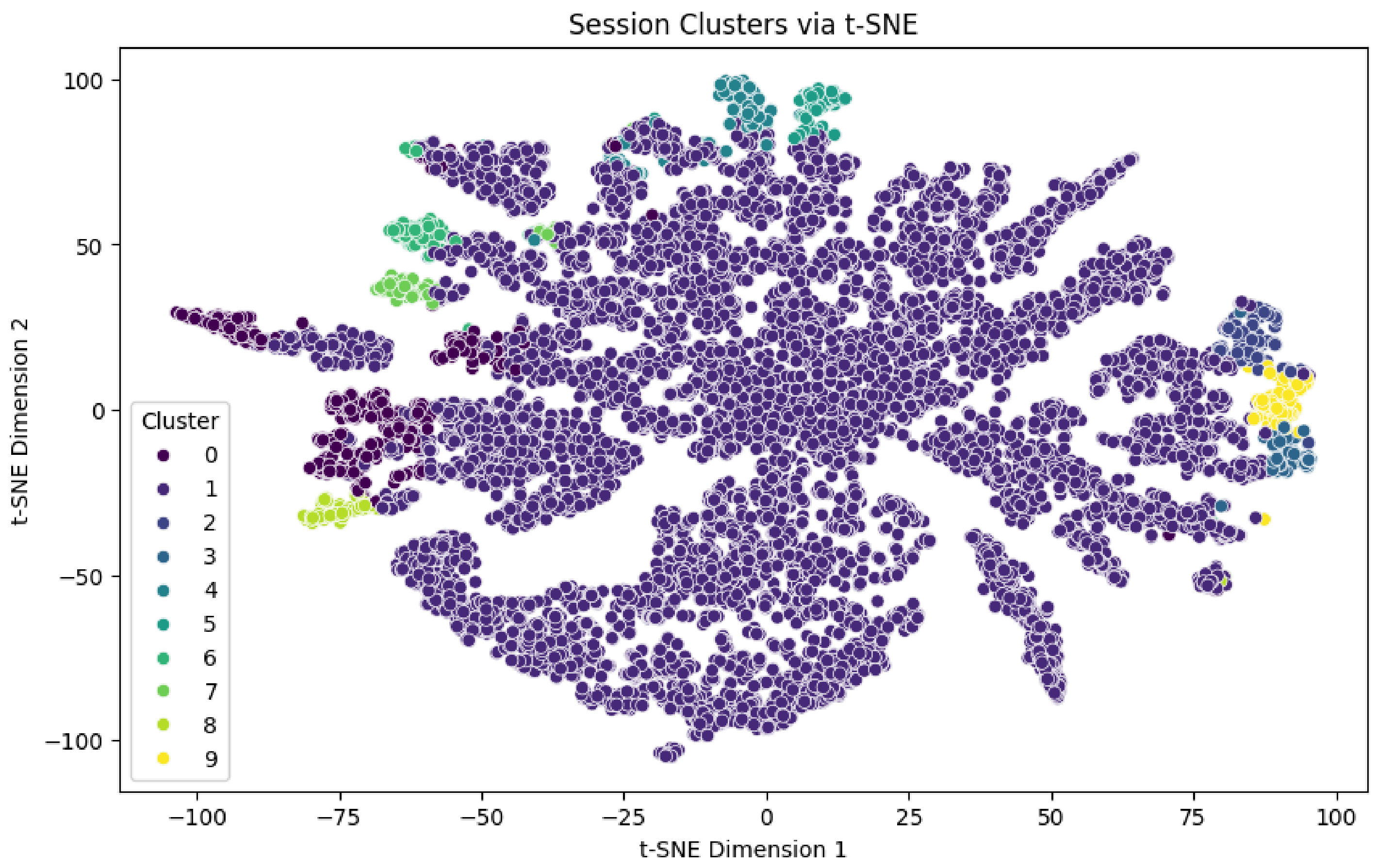
Jo descriptions are long text (21k jobs with average 1k tokens)
21M tokens to be analyzed



# Collaborative Filtering

- Singular Value Decomposition (SVD)
- Interaction Matrix
- Collaborative filtering score




| | Collaborative |
|---|---|
| MRR | 0.032 |
| Action Score | 42.5% |
| Final Score | 0.15 |

# Planning

# Planning

1. Features Extraction
2. Features Embedding
3. Content-Based Filtering
4. Hybrid Recommendation System
5. Action prediciton via MLP

# Features Extraction

Challenge: analyze 21M tokens per feature (multiple features per prompt is too complex for smaller models)

Solution: Optimized GPU implementation

Model: Llama 3.2 Instruct 1B (fits in one 16gb RAM with a reasonable output rate)

| title_section | seniority | company | industry | location |
|---|---|---|---|---|
| Architecte messagerie M365 | Senior | LeHibou | Banking | Bordeaux, France |
| UX DESIGNER SENIOR | Senior | Computer Futures | IT | London, England |
| Chef de projet technique média | Mid | "MediaCorp" | Media | New York, USA |
| Développeur Full Stack Java/ Angular | Senior | Développeur Full Stack Java/Ang | Full Stack | Paris, France |
| Développeur BO BI4 | Senior | SAP | Industry: Finance | Paris, France |
| Architecte GCP | Senior | Google Cloud Platform | IT | Paris, France |
| Product Manager PIM/DAM/MDM | Senior | Carrefour Group | Industry: Retail | Paris, France |
| Consultant project manager Senior | Senior | Recueillir et comprendre les besoins | IT | New York, USA |
| un architecte cloud Azure F/H | Senior | Infogene | Cloud | Paris, France |
| Lead Tech - Javascript (h/f) | Senior | LeHibou | IT | Bordeaux, France |
| PMO / Product Owner | Senior | TRSB | EDI | Paris, France |
| Chef de Projet Logiciel | Mid | "Toulouse" | IT | Toulouse, France |
| Développeur PHP / Symfony | Senior | HR Team | PHP/ Symfony | Paris, France |
| Chef de projet data supply (H/F) | Senior | Insitoo | Data Supply | Paris, France |
| Tech Lead Java & Scrum Master (H/F) | Senior | Accelite IT & Business Consulting | Banking & Finance | Paris, France |
| Développeur Java (Luxembourg) | Senior | GBTO/MAR/REG/ITS | IT | Luxembourg, Luxembourg |
| ADMINISTRATEUR SAP BASIS H/F | Senior | SAP | Information Systems | Paris, France |
| Consultant Testeur Confirmé H/F | Mid | Hexateam | Business Intelligence | Paris, France |
| ır Etudes et Développement C++ / Sophis | Senior | Ingénieur Etudes et Dé | Finance | Paris, France |
| Devops (mob) freelance | Senior | FreelanceRepublik | DevOps | Paris, France |

# Features Embedding

1. Hugging Face Sentence Transformers - all-MiniLM-L6-v2 with 384 dimensions

```
# Load pre-trained sentence transformer model
from sentence_transformers import SentenceTransformer
from transformers import pipeline, AutoTokenizer, AutoModelForQuestionAnswering, BigBirdTokenizer

model = SentenceTransformer('all-MiniLM-L6-v2').to(device)
```

2. OpenAI API - text-embedding-3-small with 1024 dimensions

Same results in Content-Based Filtering and Hybrid Recommendation System

# Content-Based Filtering

User preference creation

$$u_f = \sum_{past\ jobs\ j} w_{action} . j_f$$

Cosine Similarity with all jobs to rank and recommend the top 10

$$< u, j >= \sum_{features} w_f . \| u_f , j_f \|$$

# Content-Based Filtering

Learning the feature weights

| | Title | Location | Seniority | Company | Industry |
|---|---|---|---|---|---|
| Feature Weight | 5.63 | 0.00 | 0.00 | 0.00 | 0.89 |

Table 1: Learned Feature Weights

No improvements on the MRR result

| | Content-Based |
|---|---|
| MRR | 0.010 |

# Hybrid Filtering

$$S^{ij}_{Hybrid} = \alpha \cdot S^{ij}_{CF} + (1 - \alpha) \cdot S^{ij}_{CBF}$$

# Action Prediction via MLP

$$h = \sigma(W_1 \cdot x + b_1)$$

$$\hat{y} = \sigma(W_2 \cdot h + b_2)$$

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

$$\text{Final Score} = 0.7 \times \text{MRR} + 0.3 \times \text{Action Accuracy}$$

# Results

# Results

| | Collaborative | NN |
|---|---|---|
| Action Score | 42.5% | 67.1% |

| | Content-Based | Collaborative | Hybrid |
|---|---|---|---|
| MRR | 0.010 | 0.032 | 0.041 |
| Final Score | 0.15 | 0.15 | 0.23 |

Table 2: Best score obtained for each the model.

# Next Steps

# References

[1] M. cakir, sule gunduz oguducu, and resul tugay. A deep hybrid model for recommendation systems, 2020.

[2] X. Chen, L. Yao, J. McAuley, G. Zhou, and X. Wang. A survey of deep reinforcement learning in recommender systems: A systematic review and future directions, 2021.

[3] R. Glauber and A. Loula. Collaborative filtering vs. content-based filtering: differences and similarities, 2019.

[4] V. Kharidia, D. Paprunia, and P. Kanikar. Lightfusionrec: Lightweight transformers-based cross-domain recommendation model, 2024.

[5] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui. Graph neural networks in recommender systems: A survey, 2022.

[6] E. Çano and M. Morisio. Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6):1487–1524, Nov. 2017.