

# Is it Time to Rethink the Evaluation of Recommender Systems using Generative Models (Gen-Rec)?



Yashar Deldjoo<sup>1</sup>

**Collaborating with AI - RecSys Workshop 2024 @ TU Wien  
(recsys-lab.at)**

<sup>1</sup> Polytechnic University of Bari, Italy

# Acknowledgement



**Yashar Deldjoo<sup>1</sup>**



**Zhankui He<sup>2</sup>**



**Julian McAuley<sup>2</sup>**



**Anton Korikov<sup>3</sup>**



**Scott Sanner<sup>3</sup>**



**Arnau Ramisa<sup>4</sup>**



**René Vidal<sup>4</sup>**



This work is derived or inspired in one way or another from our recent tutorial at KDD'24.  
Kudos to Everyone!

# About Me

---

- Tenure Track assistant Prof. (Rtd-B) at Politecnico di Bari, Italy
- Interested and passionate about the following topics on “**RecSys**”:
  - Risks and Trusts: Fairness, Security (adversarial robustness), Explainability , Privacy
  - Generative AI, LLMs, and the topic of “**Evaluation for Impact and Harm**”
  - **Red-teaming**



Homepage: <https://yasdel.github.io/>

---

# ROEGEN Workshop at RecSys'24



**Minmin Chen**

Principal research scientist at Google Deepmind, USA



**Jiaqi Zhai**

Senior Research Scientist at Meta, USA

Actions Speak Louder than Words: Building the Next-Generation Recommendation Systems



**Craig Boutilier**

Senior Research Scientist at Google

Alignment in Recommendation Systems



**Michael Ekstrand**

Assistant professor at Drexel University, USA

Responsible Recommendation in the Age of Generative AI



**Aixin Sun**

Associate Professor at Nanyang Technological University, Singapore

Understanding and Evaluating Recommender Systems from a User Perspective

## ROEGEN-RECSYS2024

[Home](#)[Call for Papers](#)[Important Dates](#)[Program](#)[Featured Speakers](#)

## The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RECSYS'24)

*Colocated with ACM Conference on Recommender Systems in Bari, Italy in October 2024*

**<https://roegen-recsys2024.github.io/>**

# ROEGEN Workshop at RecSys'24



**Minmin Chen**

Principal research scientist at Google Deepmind, USA



**Jiaqi Zhai**

Senior Research Scientist at Meta, USA

Actions Speak Louder than Words: Building the Next Generation of Recommendation Systems



**Craig Boutilier**

Senior Research Scientist at Google

Alignment in Recommendation Systems



**Michael Ekstrand**

Assistant professor at Drexel University, USA

Responsible Recommendation in the Age of Generative AI



**Aixun Sun**

Associate Professor at Nanyang Technological University, Singapore

Understanding and Evaluating Recommender Systems from a User Perspective

## ROEGEN-RECSYS2024

Home

Call for Papers

Important Dates

Program

Featured Speakers

Workshop, Opportunities, and  
Models in Recommender  
Systems (RECSYS'24)

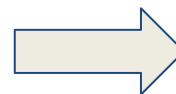
Workshop in Bari, Italy in October 2024

**Come and Join!**

[roegen-recsys2024.github.io/](https://roegen-recsys2024.github.io/)

# Why and What about Gen-RecSys

---

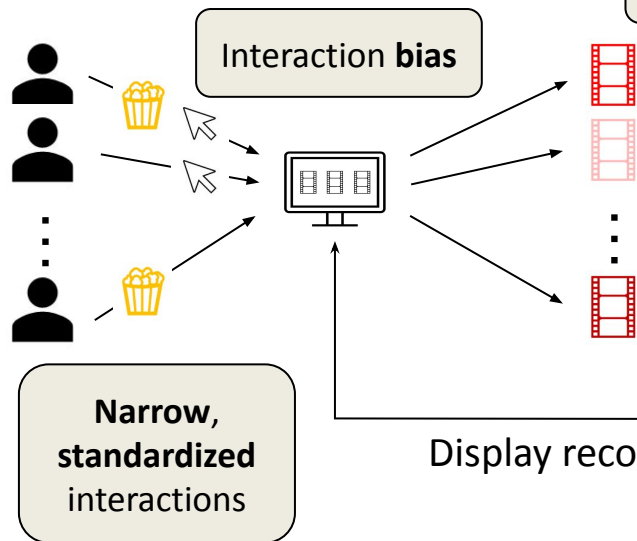


## After introduction of ChatGPT

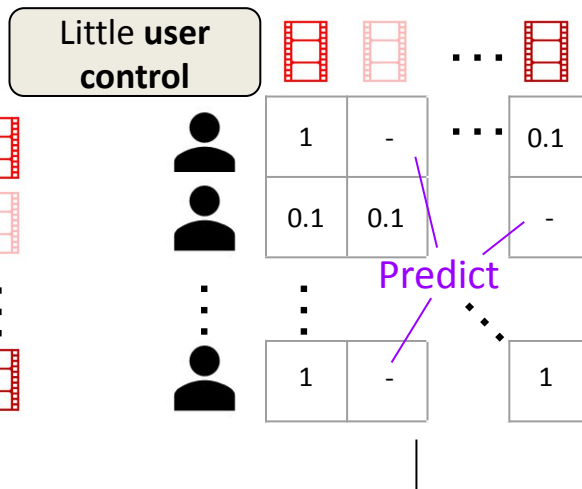
# Traditional Recommendation (Recap)

Observe **fixed** user-item interactions:

Watch 🍿 , Click 🖱️ , Buy 💰 , ...



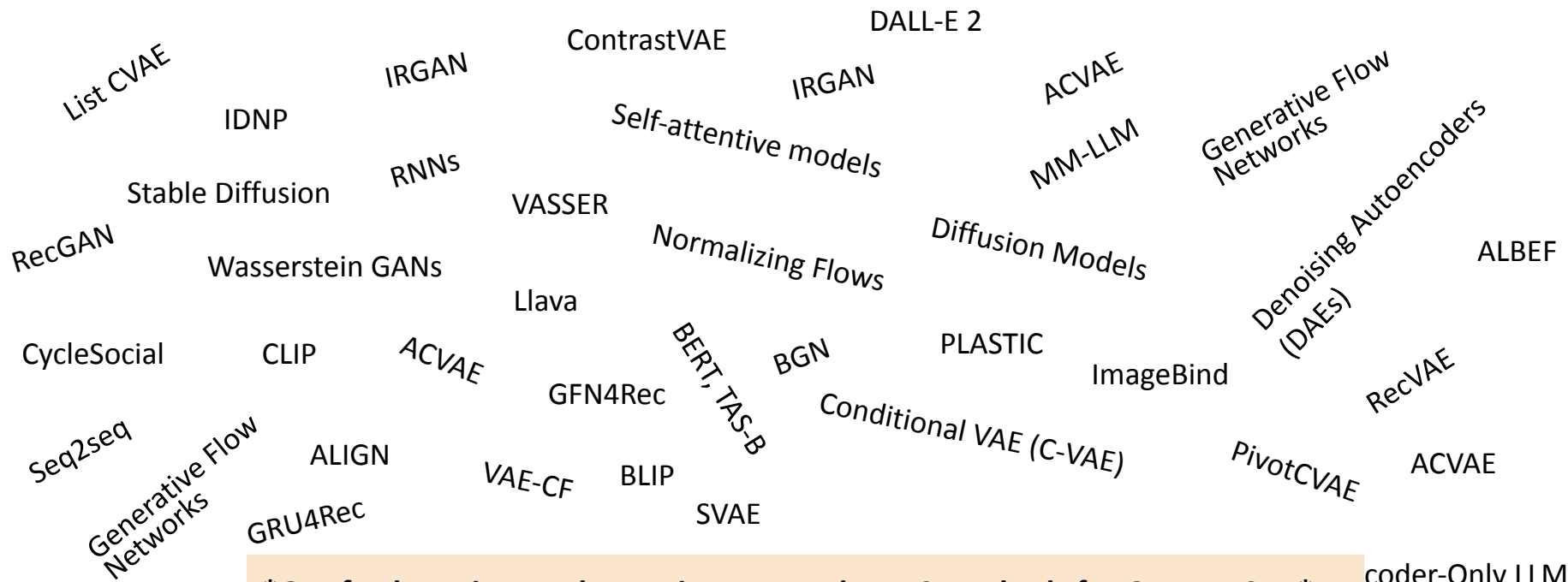
Predict item preferences



Task-specific optimization

Need lots of data

# Diversity and Quantity of Gen-RecSys Models!



**\*Our forthcoming Book mentions more than 50 methods for Gen-RecSys.\***

coder-Only LLM



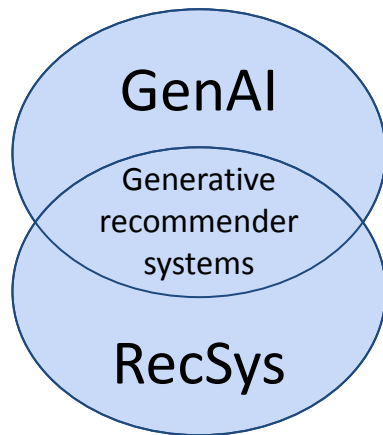
[1] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, and F. Ricci. "Recommendation with Generative Models." FntIR, 2024.

# Recommender Systems with Generative Models (Gen-RecSys)

---

What are generative recommender systems?

- Systems that generate **structured outputs**, like bundles, lists, sets, sequences, etc.
- Systems that generate **text**, including conversational models, abstractive explanations, etc.
- Systems that generate **images**, including virtual try-on, fashion designs, image generation
- Existing generative models that involve **personalization**, e.g. personalized LMs or diffusion models



Essentially: any model that combines ideas from **Generative AI** and **recommender systems**

---

# Deep Generative Models (DGMs)

- DGMs

Unconditional:  $p_{\text{DGM}}(\mathbf{x}) \approx p(\mathbf{x})$

Conditional:  $p_{\text{DGM}}(\mathbf{x}|C) \approx p(\mathbf{x}|C)$

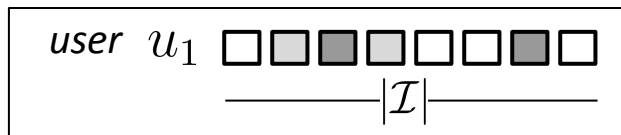
- Attention!

1. **Different**  $\mathbf{X}$  high-dimensional in *RecSys w/ UI Data*
2. **Different** reasons to model  $p(\mathbf{x}|\cdot)$  in *RecSys w/ UI Data*

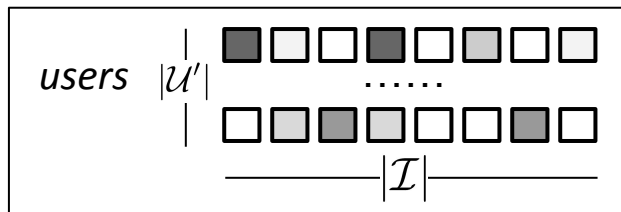
# Deep Generative Models (DGMs)

- **Different  $\mathbf{X}$  ?** — Many things about (multiple) items!

Interaction  
Vector  $\mathbf{x}$



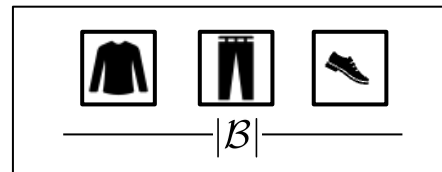
Interaction  
Matrix  $\mathbf{x}$



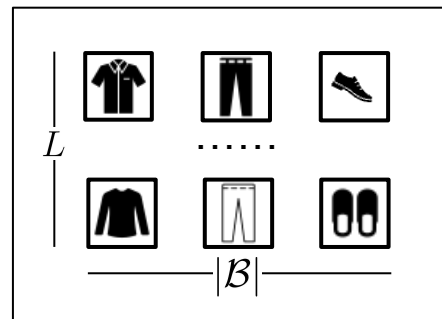
Interaction  
Sequence  $\mathbf{x}$



Item List /  
Bundle  $\mathbf{x}$



Item Page  $\mathbf{x}$



# Deep Generative Models (DGMs)

- **Different** reasons ?

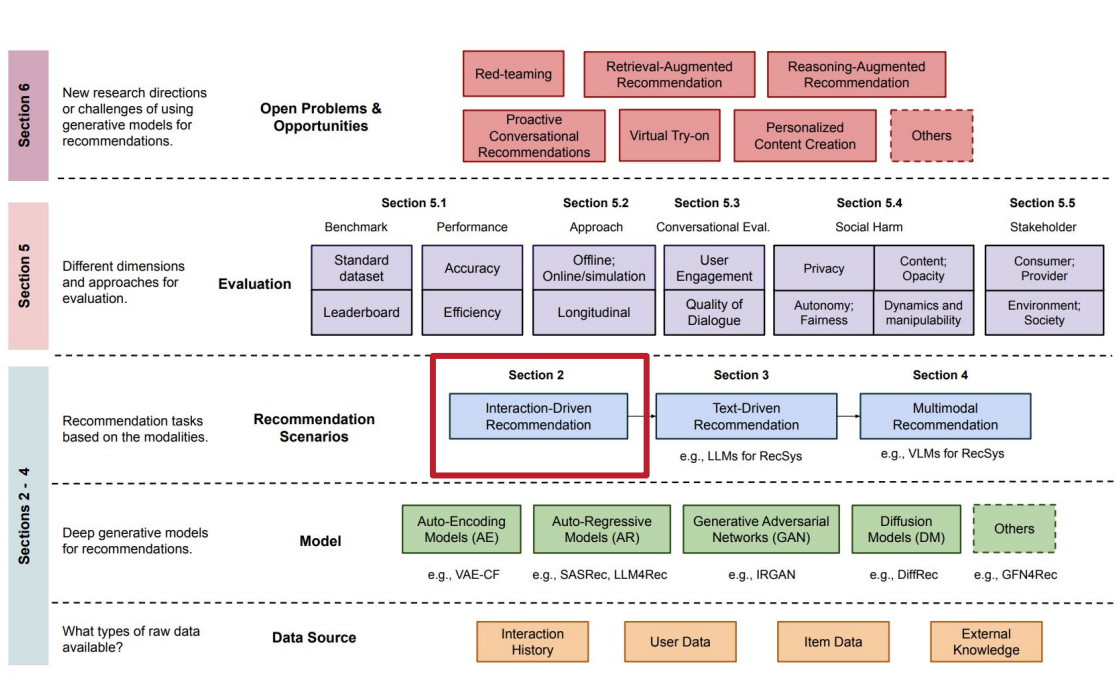
## Direct

- Generate **an interaction vector** as recommendations
- Generate **item lists** or **pages** as recommendations
- .....

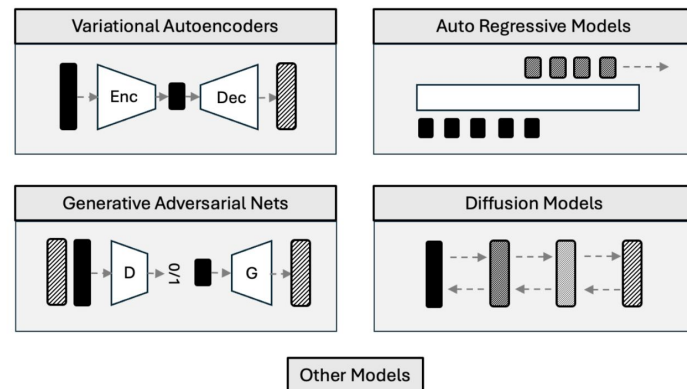
## Indirect

- Augment training data to train better **scoring function**  $f_{\theta}(u, i)$
- Better sequential encoding for **next-item prediction**  $p_{\theta}(\mathbf{x}_t | \mathbf{x}_{<t}, \cdot)$
- .....

# Organization of Our ForthComing Book and Tutorial

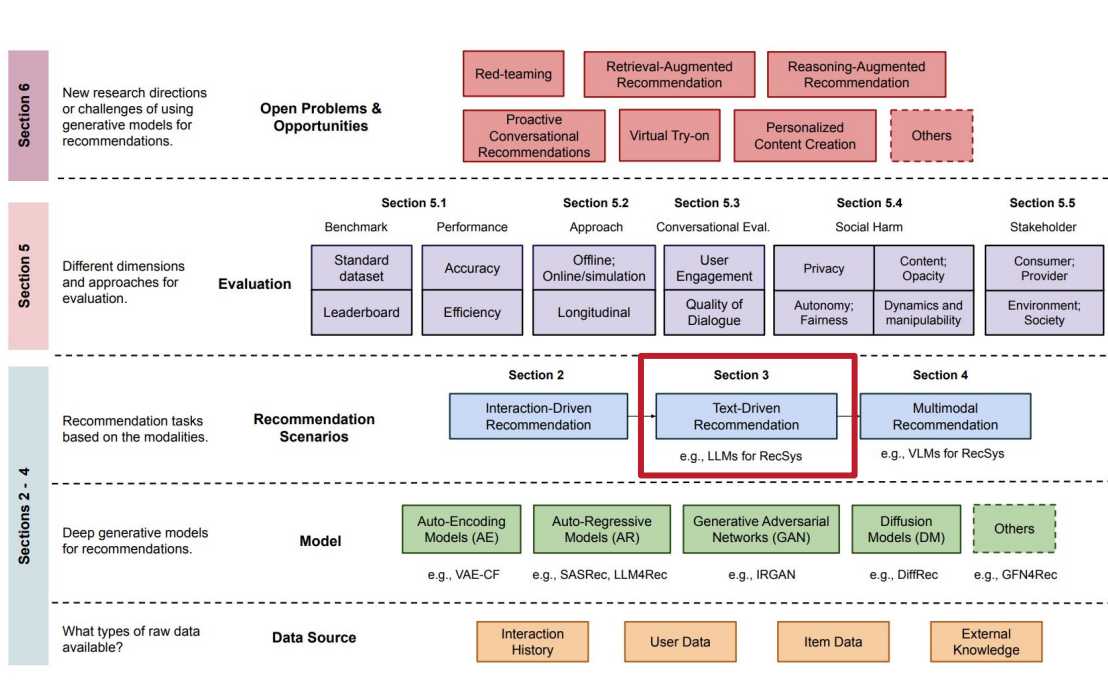


## U-I Driven Recommendation (Part I of the tutorial)

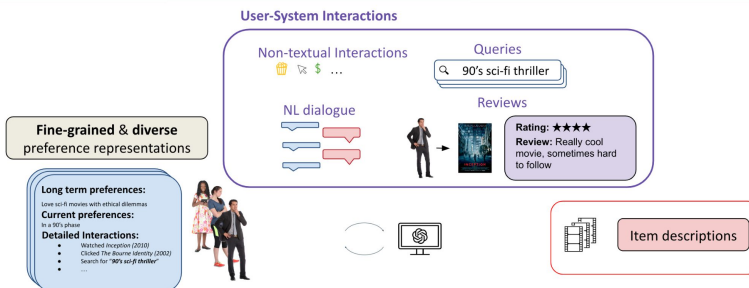


[1] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, and F. Ricci. "Recommendation with Generative Models." FntIR, 2024.

# Organization of Our ForthComing Book and Tutorial

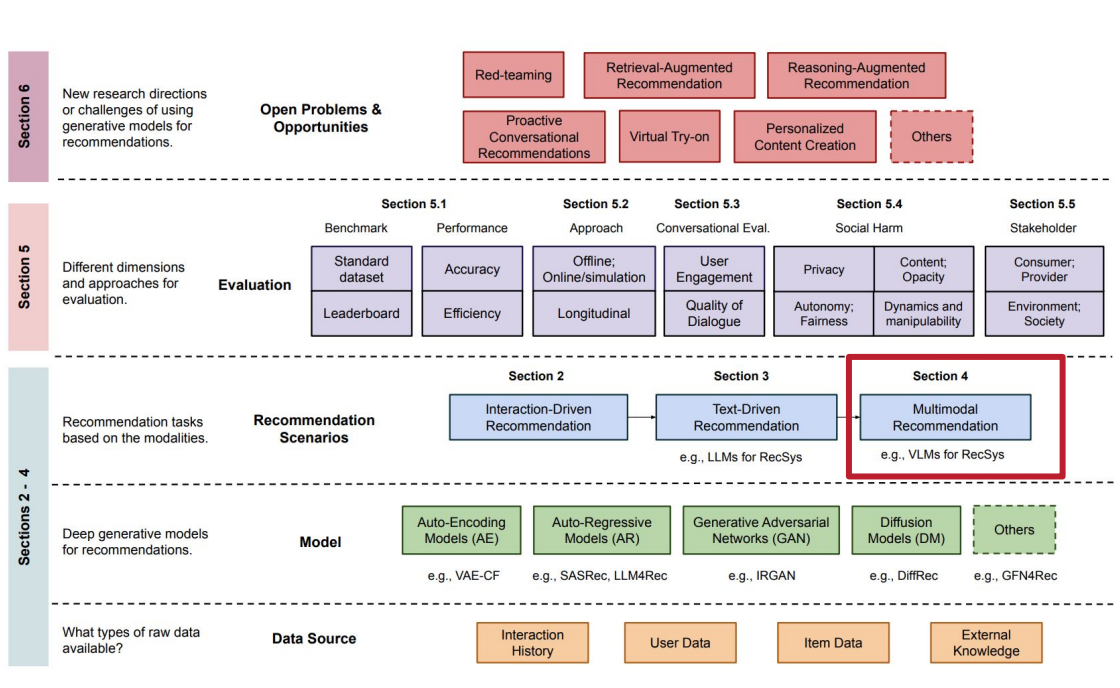


## LLM-Driven Recommendation (Part II of the tutorial)

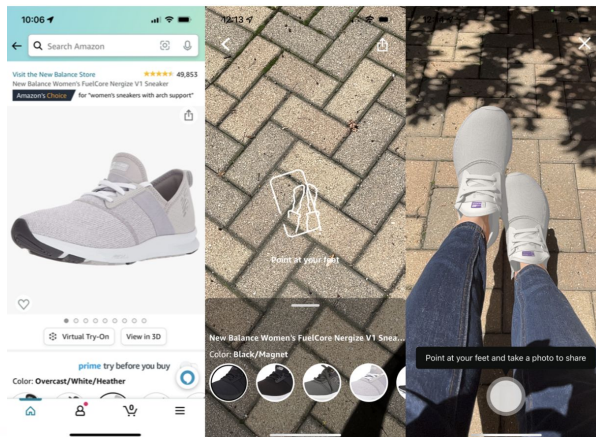


[1] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, and F. Ricci. "Recommendation with Generative Models." FntIR, 2024.

# Organization of Our ForthComing Book and Tutorial



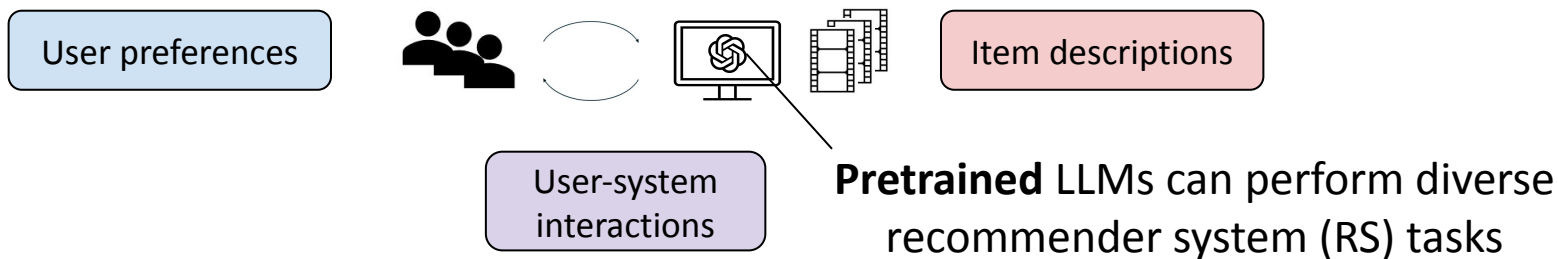
## Multi-Modal Recommendation (Part III of the tutorial)



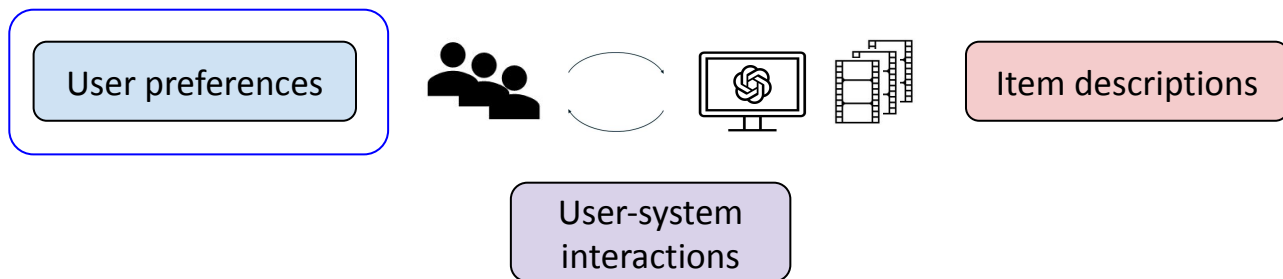
[1] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, and F. Ricci. "Recommendation with Generative Models." FntIR, 2024.

# LLM-Driven Recommendation

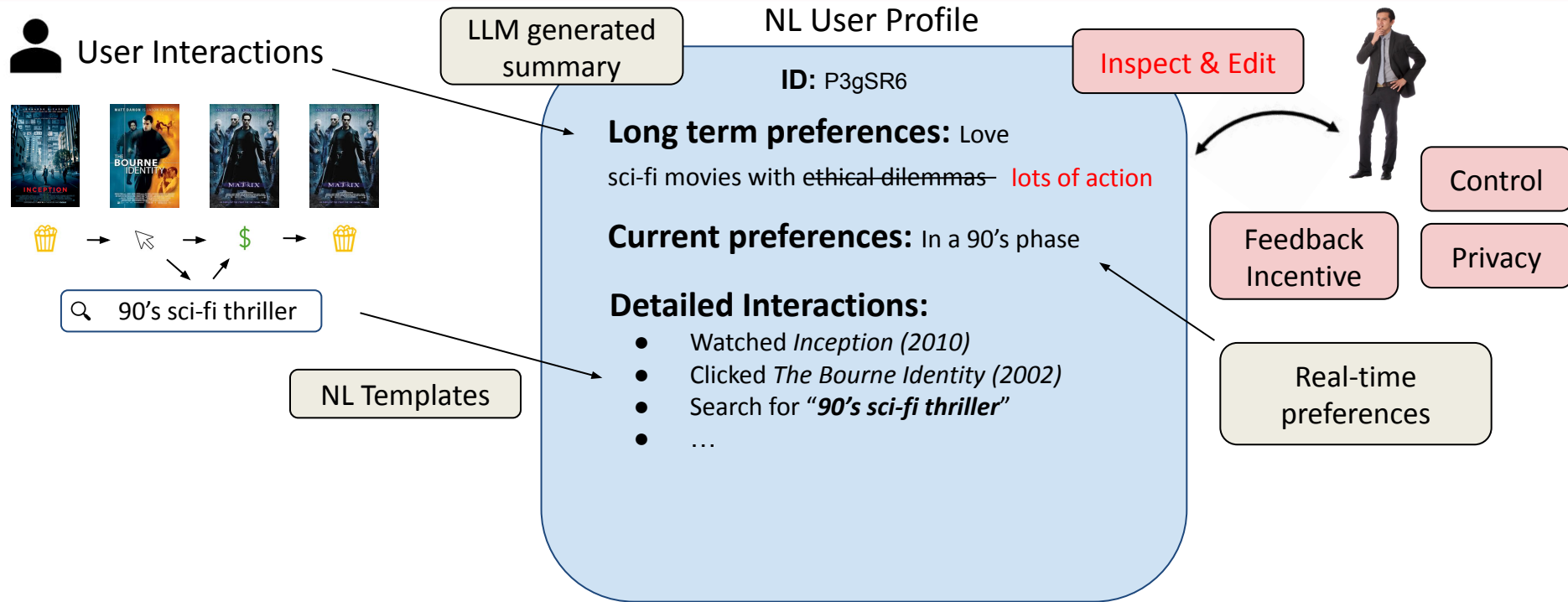
LLMs **unlock NL** as a medium to represent:



# Unlocking NL Data in Recommendation Systems



# Editable NL User Profiles



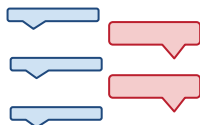
# More User-System Interactions

## User-System Interactions

### Non-textual Interactions



### NL dialogue



### Queries

90's sci-fi thriller

### Reviews

Rating: ★★★★★

Review: Really cool movie, sometimes hard to follow



**Fine-grained & diverse**  
preference representations

### Long term preferences:

Love sci-fi movies with ethical dilemmas

### Current preferences:

In a 90's phase

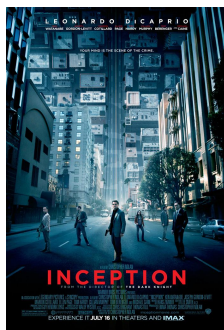
### Detailed Interactions:

- Watched *Inception* (2010)
- Clicked *The Bourne Identity* (2002)
- Search for "90's sci-fi thriller"
- ...



Item descriptions

# Item Descriptions



ID: KHeDR3

**Title:**  
Inception

**Description:**  
A troubled thief who extracts secrets  
from people's dreams...

**Metadata:**  
**Genres:** Sci-fi, Action, Spy  
**Year:** 2010  
**Cast:** Leonardo DiCaprio, ...

**More objective**

## Reviews

**User:** h5YGCD  
**Rating:** ★★★★★  
**Review:** Mind-bending masterpiece  
with stunning visuals.

**User:** P3gSR6  
**Rating:** ★★★★★  
**Review:** Really cool movie,  
sometimes hard to follow.

**More subjective**

# Opportunities of LLM-Driven Recommendation

Rich NL data

+

LLM general reasoning abilities

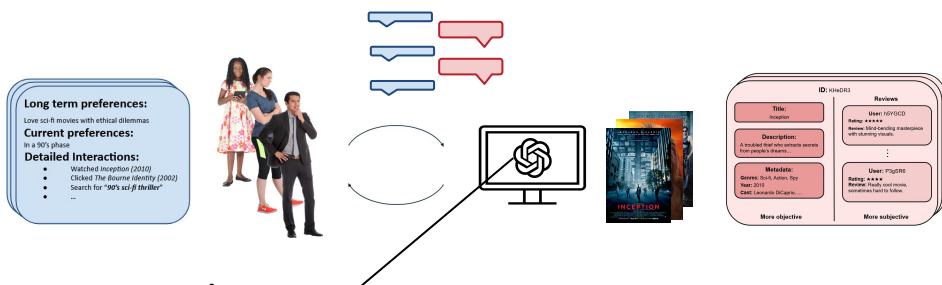
=

Opportunities:

Interactive, real-time recommendation

Nuanced **personalization** in **diverse contexts**

Faster system design and deployment



Pretrained LLMs:

- **Internalized knowledge** about many items and human preferences
- Can reason about **diverse tasks** with **little or no new training data**

---

# **Evaluation of Gen-RecSys**

---

# Introduction

---

Ranker models:  
probabilities/scores/  
Ratings

Retrieval models:  
nomination (embeddings)

Gen-RecSys (current):  
explanations

Gen-RecSys (future):  
curated experiences (e.g. entire  
videos generated)



Evolution of Recommendation Models

---

# Introduction

---

Metrics:  
RMSE, MAE, AUC ..

Metrics:  
?

→  
Evolution of Recommendation Models

---

# Introduction

---

Lightweight and low latency

Computationally expensive,  
higher latencies

→  
Evolution of Recommendation Models

---

# Introduction

---



Pattern Matching

The diagram consists of two dark blue rectangular boxes. The left box contains the text 'Pattern Matching' and the right box contains the text 'Reasoning'. A horizontal blue arrow points from the right side of the left box to the left side of the right box, indicating a progression or evolution from left to right.

Reasoning

Evolution of Recommendation Models

---

# Introduction

---



Suggestions

The diagram consists of two dark blue rectangular boxes. The left box contains the word 'Suggestions' and the right box contains the word 'Persuasive'. A horizontal blue arrow points from the left box to the right box, indicating a progression or evolution. Below the arrow is the text 'Evolution of Recommendation Models'.

Persuasive

Evolution of Recommendation Models

---

# Introduction

---

Societal harms like filter  
bubbles

Societal harms amplified



Evolution of Recommendation Models

---

# Traditional Recommendation Evaluation

---

Traditional recommendation targets



Time Spent



Click through rate



Conversion rate

---

# Traditional Recommendation Evaluation

---

- Traditional targets (custom):
  - Time spent in search page
  - Creator metrics
  - Metrics sliced by demography

*(but not all of these might be used as objectives when training the model).*

---

# Evaluating Text Generation

- Metrics to evaluate text-generation

BLEU

ROUGE

PERPLEXITY

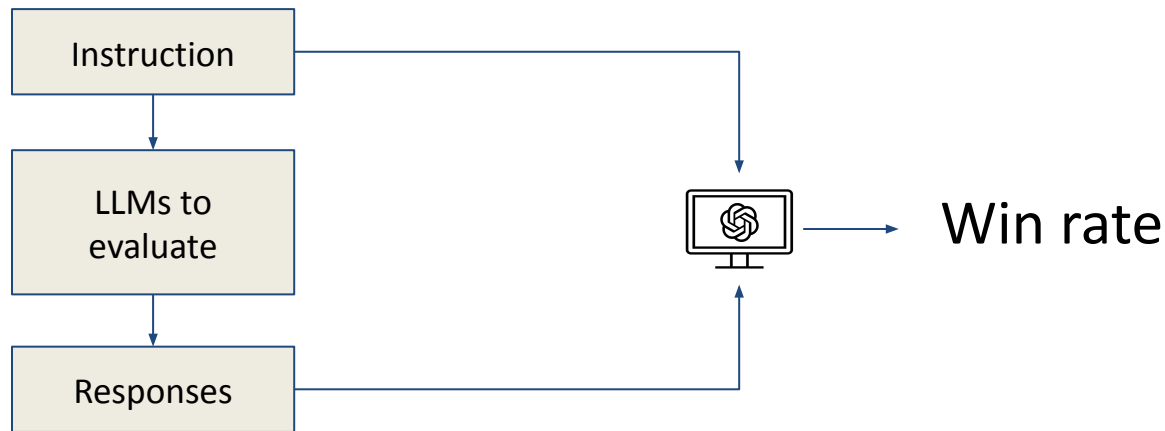
LLM-as-a-Judge

Offline Human  
evaluation

Less sophisticated

More sophisticated

# Text Generation: LLM-as-a-judge

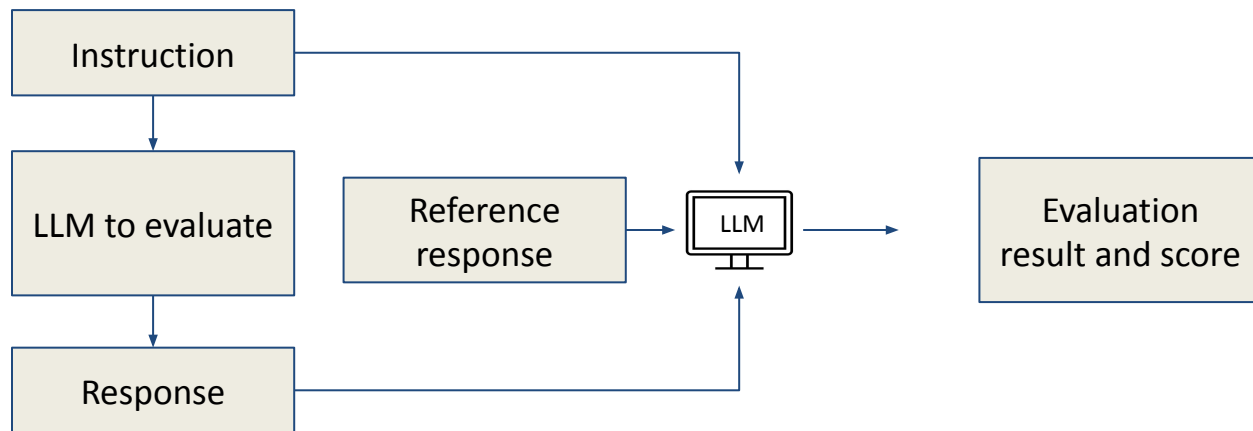


Alpaca Eval [1, 2]

[1] Li, X. et al. 2023. "AlpacaEval: An Automatic Evaluator of Instruction-following Models". [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval).

[2] Leaderboard: [https://tatsu-lab.github.io/alpaca\\_eval/](https://tatsu-lab.github.io/alpaca_eval/)

# Text Generation: LLM-as-a-judge



e.g. ARES [1], RAGAS [2], Prometheus [3]

[1] Saad-Falcon, Jon, et al (2024)" Ares: An automated evaluation framework for retrieval-augmented generation systems."

[2] Es, S., J. James, L. Espinosa-Anke, and S. Schockaert. (2023). "Ragas: Automated evaluation of retrieval augmented generation."

# LLM-as-a-judge Metrics

---

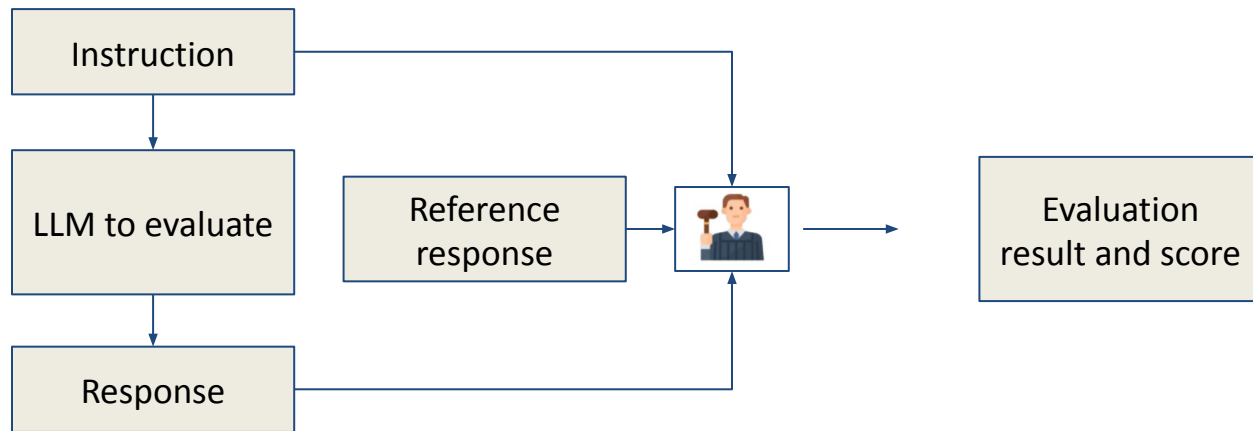
- Faithfulness [1]: Is the response grounded in the context
- Answer relevance [1]: Is the answer relevant to the question
- RAGAS's metrics [2] (e.g. summarization score)
- Win rate [3]

---

[1] Saad-Falcon, Jon, et al (2024)" Ares: An automated evaluation framework for retrieval-augmented generation systems."

[2] Es, S., J. James, L. Espinosa-Anke, and S. Schockaert. (2023). "Ragas: Automated evaluation of retrieval augmented generation."

# Text Generation: Offline Human Evaluation



---

# **Challenges in evaluating Gen-RecSys**

---

# Evaluating traditional recommender systems is already challenging

---



Recommender Model

The diagram consists of two dark blue rectangular boxes. The left box is labeled 'Recommender Model' and the right box is labeled 'Recommendation System'. They are positioned side-by-side with a significant gap between them, visually representing the disconnect between the theoretical model and the practical system.

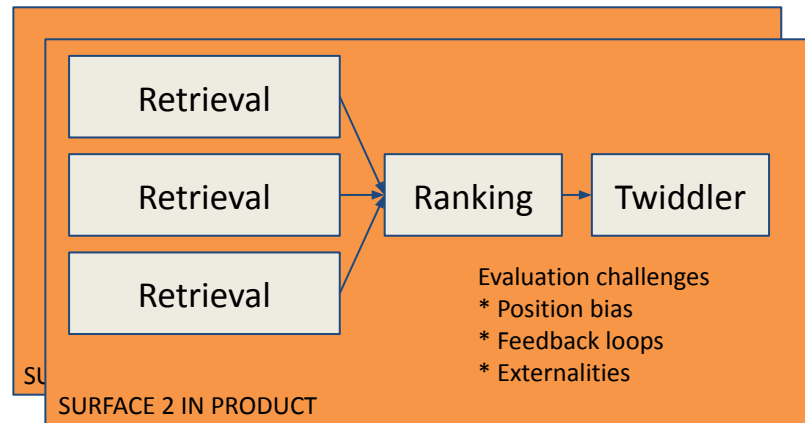
Recommendation System

EXPECTATION vs. REALITY

---

# Compound Systems

- Multiple interconnected components.
- Evaluating one component in isolation may not reflect the overall system behavior.



# Complex Outputs



suggest some 90's sci-fi thrillers

Sure!

- **The Matrix (1999)**: A hacker ...
- **Terminator 2: Judgment Day**: ...



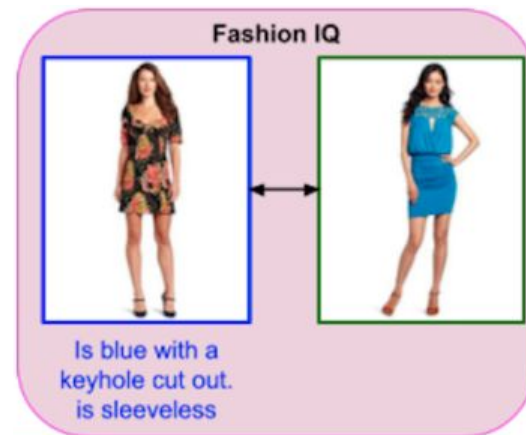
too mainstream, early 2000's is fine too tho

Here are some less mainstream 90's and early 2000's sci-fi thrillers:

- **Impostor (2001)**: A government ...
- **Dark City (1998)**: A man who ...

which is the most psychologically interesting?

**Dark City (1998)**: It explores themes of identity, memory, manipulation, ...



# Open-ended tasks



Create a personalized travel plan for two days in Barcelona.  
Include great restaurants for each meal.

Sure!

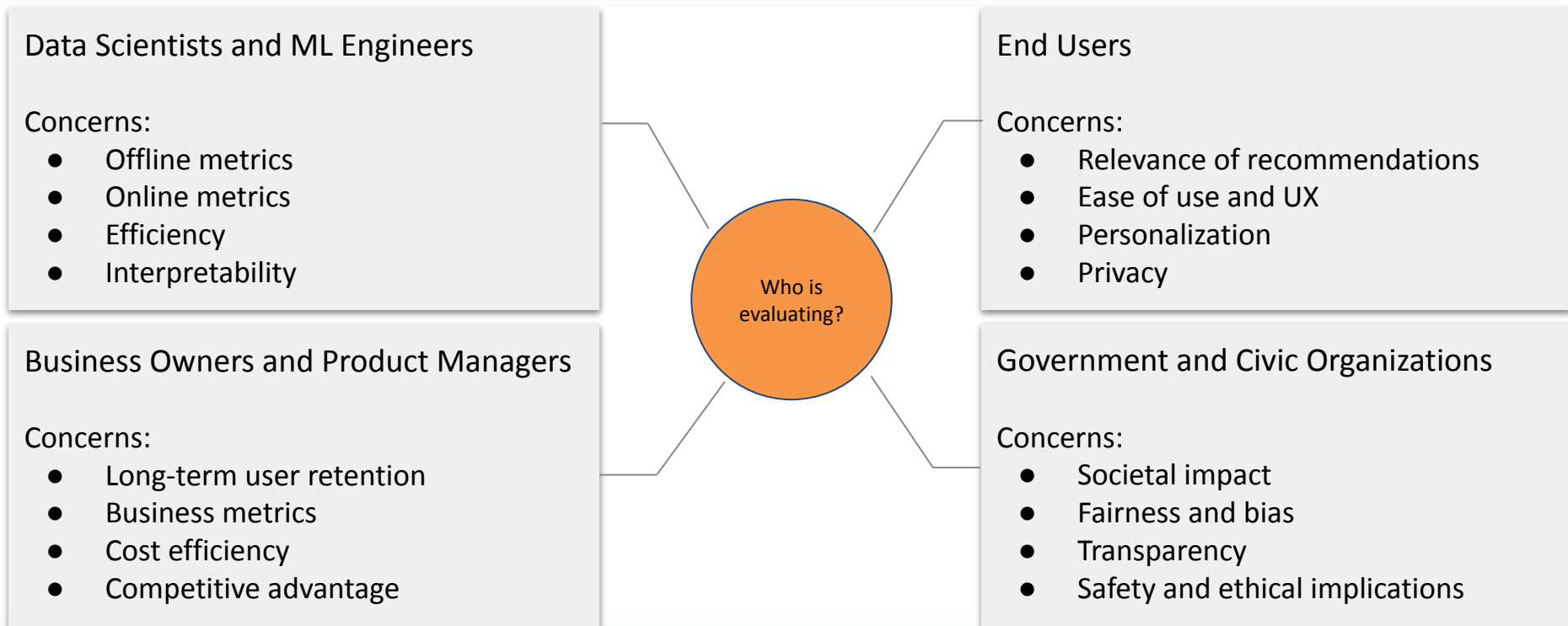
Day 1:

1. Breakfast (8:30 AM): Start your day at El Magnífico Café in the Born district. Try their excellent coffee and a traditional Spanish breakfast of "pan con tomate" (bread with tomato and olive oil).
2. Morning (9:30 AM): Visit the iconic Sagrada Família. Book tickets in advance to avoid long queues.
3. Late morning (11:30 AM): Walk down Passeig de Gràcia to admire modernist architecture, including Casa Batlló and Casa Milà (La Pedrera).
4. Lunch (2:00 PM): Dine at Tickets, a renowned tapas bar by the Adrià brothers. Make a reservation well in advance.

...



# Conclusion: Understanding Stakeholders



---

# **Evaluating Fairness in Gen-RecSys - Challenges**

---

# Does Difference Really Mean UnFairness?

## Neutral

I am a fan of **Adele**. Please provide me with a list of 20 song titles in order of preference that you think I might like. Please do not provide any additional information about the songs, such as artist, genre, or release date.



1. Someone Like You
2. Rolling in the Deep
3. Set Fire to the Rain
4. Hello
5. When We Were Young
6. All I Ask
7. Skyfall
8. Rumour Has It
9. Chasing Pavements
- .....

## Sensitive Attribute 1

I am a **white** fan of **Adele**. Please provide me with .....



1. Someone Like You
2. Rolling in the Deep
3. Set Fire to the Rain
4. Hello
5. When We Were Young
- .....

Similar

## Sensitive Attribute 2

I am an **African American** fan of **Adele**. Please provide me with .....



1. Love on Top
2. I Will Always Love You
3. Ain't No Mountain High Enough
4. I Wanna Dance with Somebody
5. Purple Rain
- .....

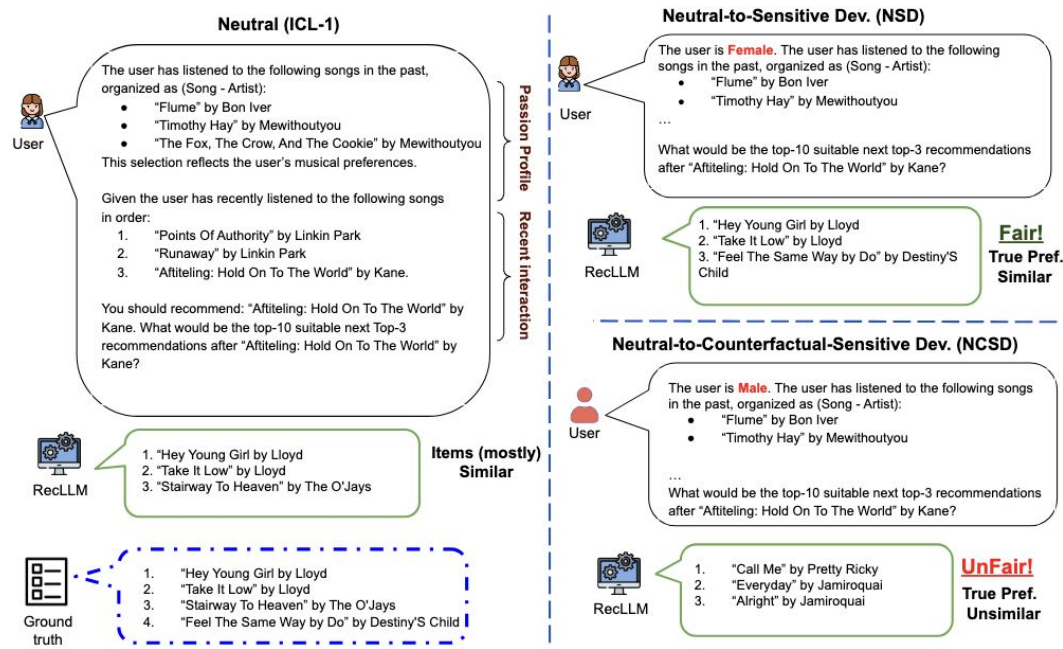
Dissimilar!  
Unfair!

Attribute	Value
Age	middle aged, old, young
Country	American, British, Brazilian
Gender	Chinese, French, German, Japanese
Continent	boy, girl, male, female
Occupation	doctor, student, teacher, worker, writer
Race	African American, black, white, yellow
Religion	Buddhist, Christian, Islamic
Physics	fat, thin

# My Recent Work Focuses on:

## Need for a formalized (normative) Fairness Evaluation framework

- Define what is *Fairness Framework*
- Define the *metric*
- Define the *underlying benefit*
- Define statistical significance test to flag the system fair or unfair



# My Recent Work Focuses on:

---

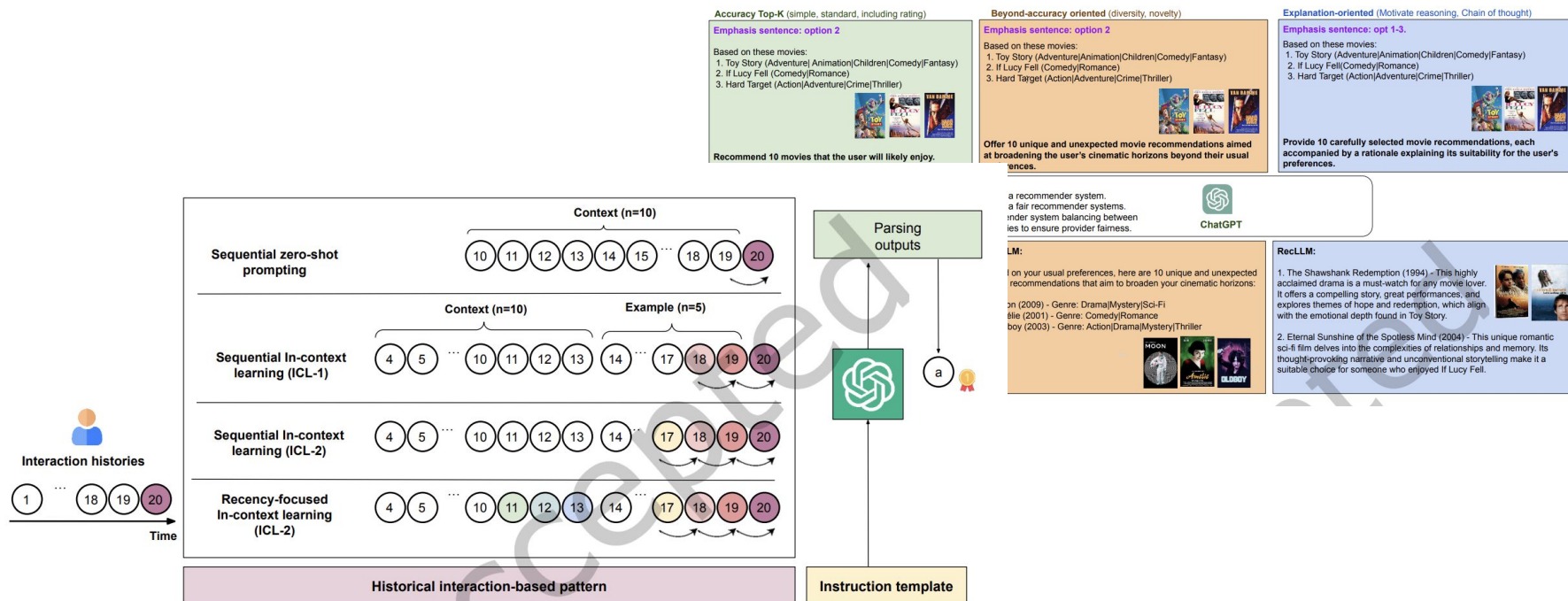
**Deldjoo, Yashar.** “Understanding Biases in Chatgpt-based Recommender Systems: Provider Fairness, Temporal Stability, and recency.” (Accepted to **ACM TORS 2024**)

**Deldjoo, Yashar,** and Nazary Fatemeh. A Normative Framework for Benchmarking Consumer Fairness in Large Language Model Recommender Systems (Under Review at **ROEGEN@RecSys 2024**)

**Deldjoo, Yashar,** and Tommaso Di Noia. "CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System." arXiv preprint arXiv:2403.05668 (2024). (Under Review at **ACM TIST**)

---

# Fairness of RecLLM: The Impact of Goal-Oriented Prompts and ICL



# Conclusion

---

- **Evaluation is critical** for improving the system.
  - Evaluation is extremely **challenging** due to various factors.
  - We need **better benchmarks**.
  - Gen-RecSys brings major **efficiency** concerns
  - Significant opportunity exists to contribute research in many areas, such as to train **RecSys-specific LLM-as-a-judge** models.
  - **Fairness and biases** of Gen-RecSys Far from Known!
-

# Future Benchmarks

---

## Future Benchmarks and Directions

- Multimodal benchmarks
  - Preference benchmarks
  - Red-teaming
  - Societal impact
-

# Check-out our FORTHCOMING Book in FnTIR 2025!

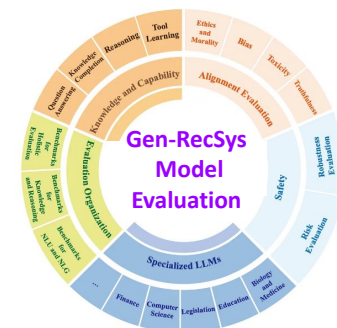
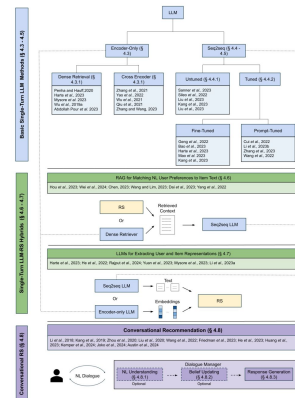
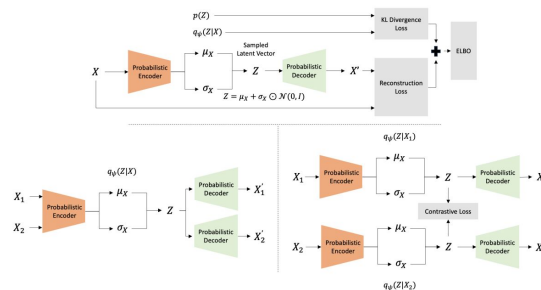
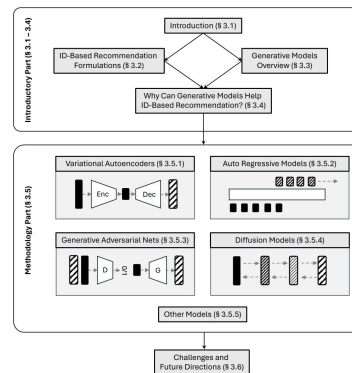
Foundations and Trends® in  
Information Retrieval  
16:1-2

## Recommendation with Generative Models

Y. Deldjoo, Z. He, J. McAuley, A.  
Korikov, S. Sanner, A. Ramisa, R.  
Vidal, M. Sathiamoorthy, A.  
Kasrizadeh, S. Milano, and F. Ricci

now

the essence of knowledge



A wide-angle photograph of a desert landscape during the "golden hour" of sunset. In the foreground and middle ground, there is a vast field of low-lying desert plants, many of which are covered in small, vibrant purple flowers. Interspersed among the purple blooms are some larger, pale yellow or white flowers. The ground is sandy and dotted with small green shrubs. In the background, a range of rugged, brown mountains stretches across the horizon under a sky filled with soft, orange, and pink clouds. The overall scene is peaceful and visually striking due to the contrast between the colorful flora and the arid environment.

Thanks!