Textual Representations for Scrutable Recommendations (TEARS)

Emiliano Penaloza

Haolun Wu

Olivier Gouvert

Laurent Charlin

SCAI: Search-Oriented Conversational AI

SCAI: Search-Oriented Conversational AI

- SCAI Events

SCAI: Search-Oriented Conversational AI
IJCAI Workshop
August 2025

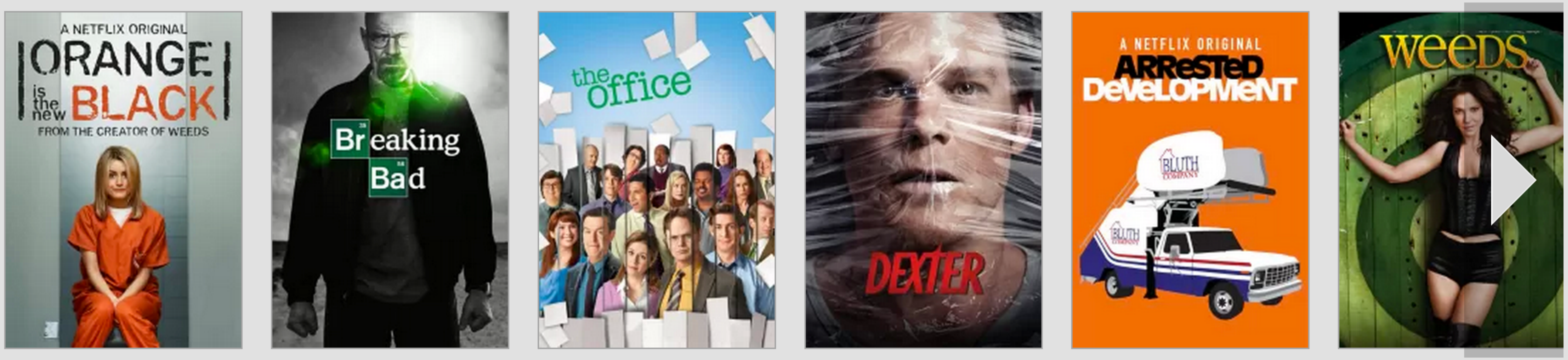
Mila

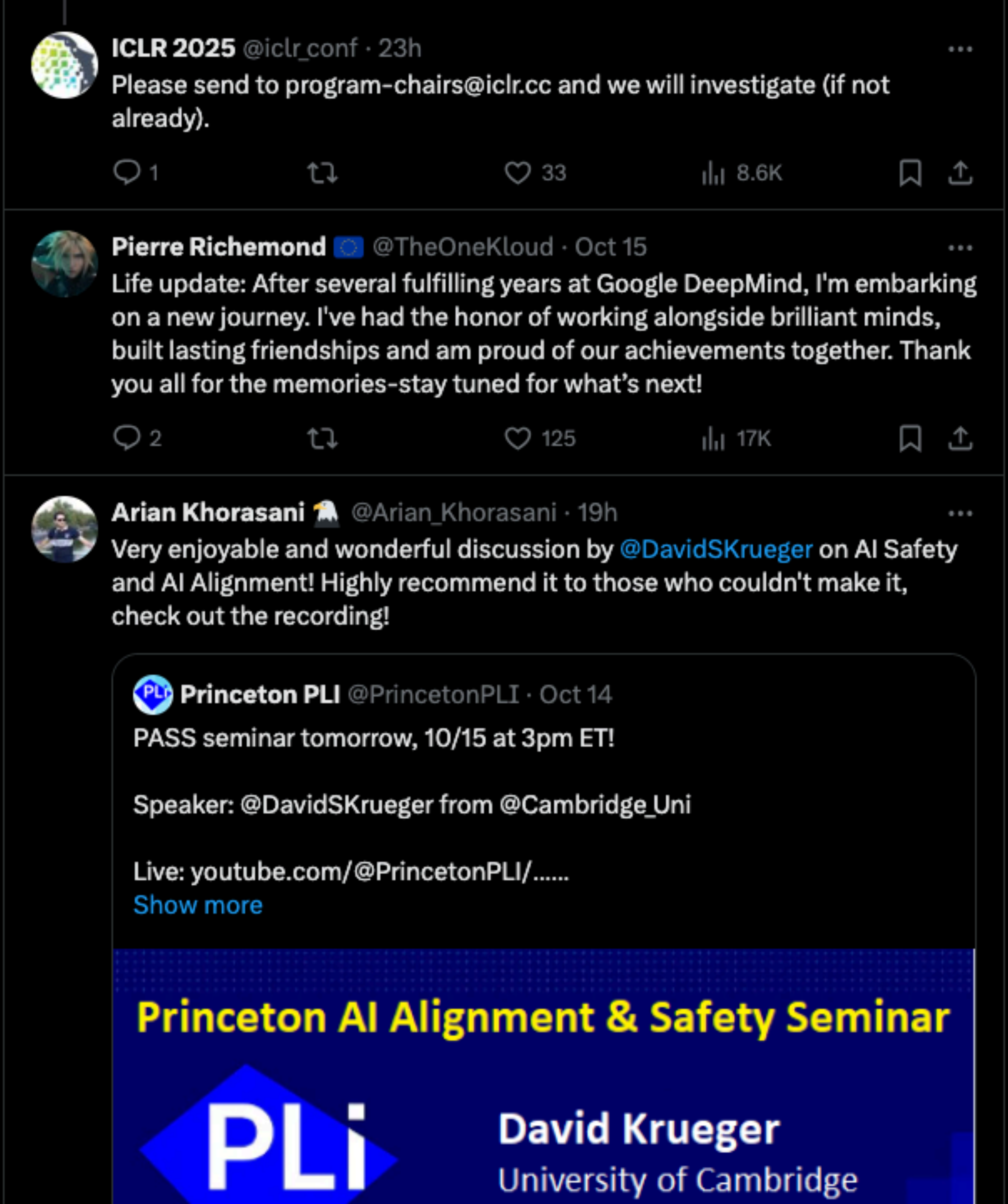
HEC MONTRÉAL

1. Modern AI techniques are **opaque**
    - Difficult to quickly adapt and correct
2. Large language models provide a **novel interface**
    - They can be used to improve human interactions
3. Focus on **recommender systems**
    - Better performance and control

NETFLIX

Top Picks for Me





ICLR 2025 @iclr_conf · 23h

Please send to program-chairs@iclr.cc and we will investigate (if not already).

Pierre Richemond @TheOneKloud · Oct 15

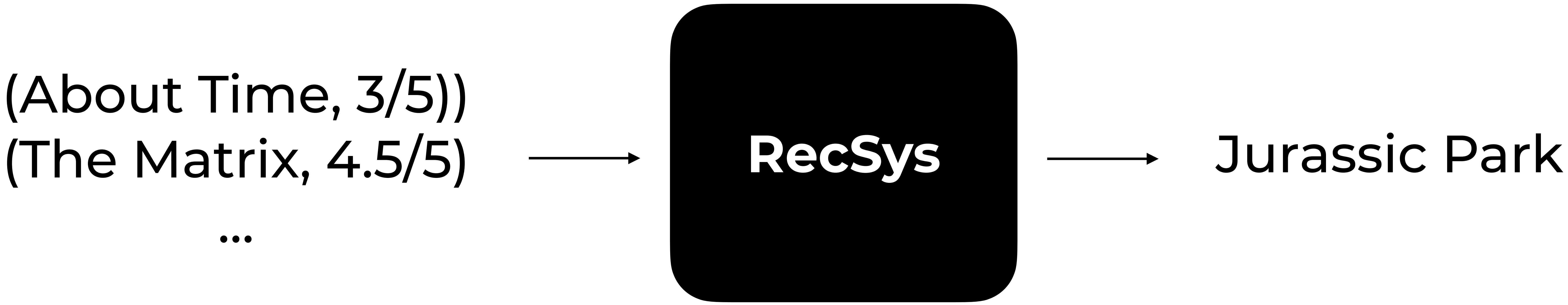
Life update: After several fulfilling years at Google DeepMind, I'm embarking on a new journey. I've had the honor of working alongside brilliant minds, built lasting friendships and am proud of our achievements together. Thank you all for the memories-stay tuned for what's next!

Arian Khorasani 🦅 @Arian_Khorasani · 19h

Very enjoyable and wonderful discussion by @DavidSKrueger on AI Safety and AI Alignment! Highly recommend it to those who couldn't make it, check out the recording!

> **Princeton PLI** @PrincetonPLI · Oct 14
>
> PASS seminar tomorrow, 10/15 at 3pm ET!
>
> Speaker: @DavidSKrueger from @Cambridge_Uni
>
> Live: youtube.com/@PrincetonPLI/......
>
> Show more
>
> Princeton AI Alignment & Safety Seminar
>
> PLi
>
> David Krueger
>
> University of Cambridge

Recommender Systems (RecSys)



(About Time, 3/5))
(The Matrix, 4.5/5)
...

RecSys

Jurassic Park

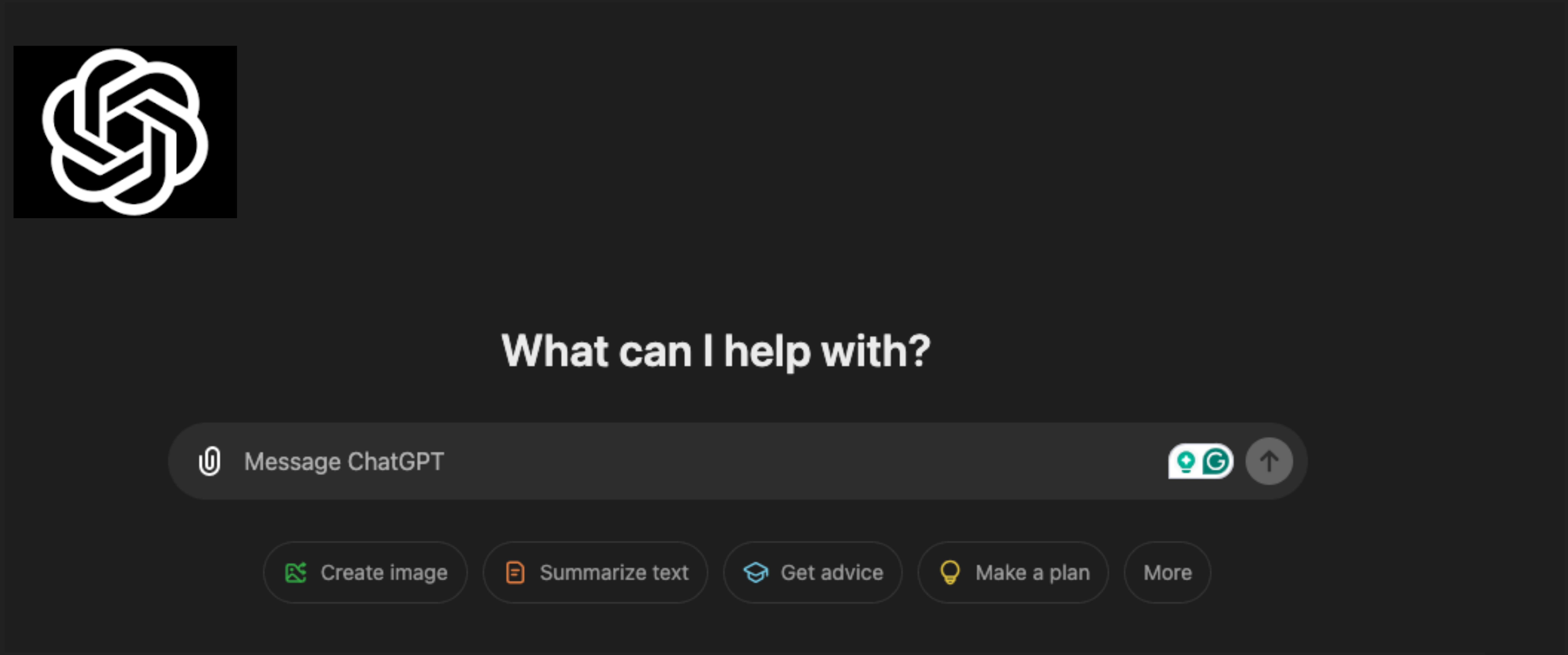
Modelling of
User Preferences

- Users have little control over these recommender systems

1. Fixing bad/missing recommendations?

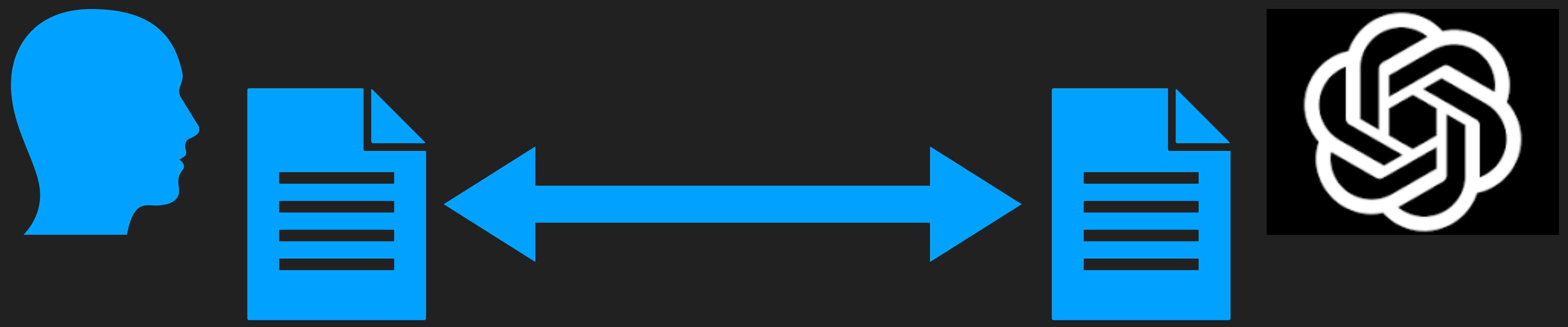
2. Providing contextual information?

No (limited) Feedback Mechanism

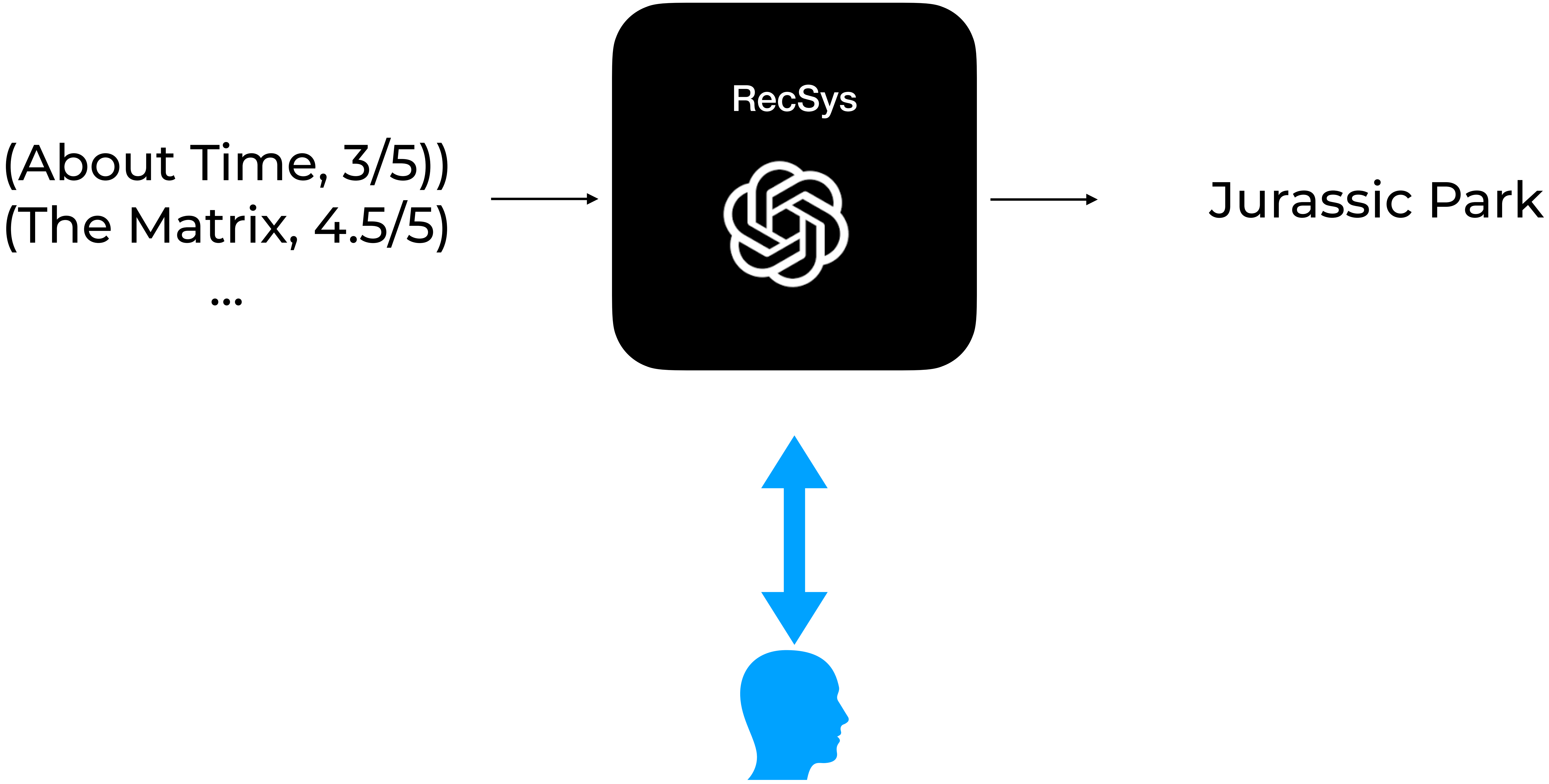


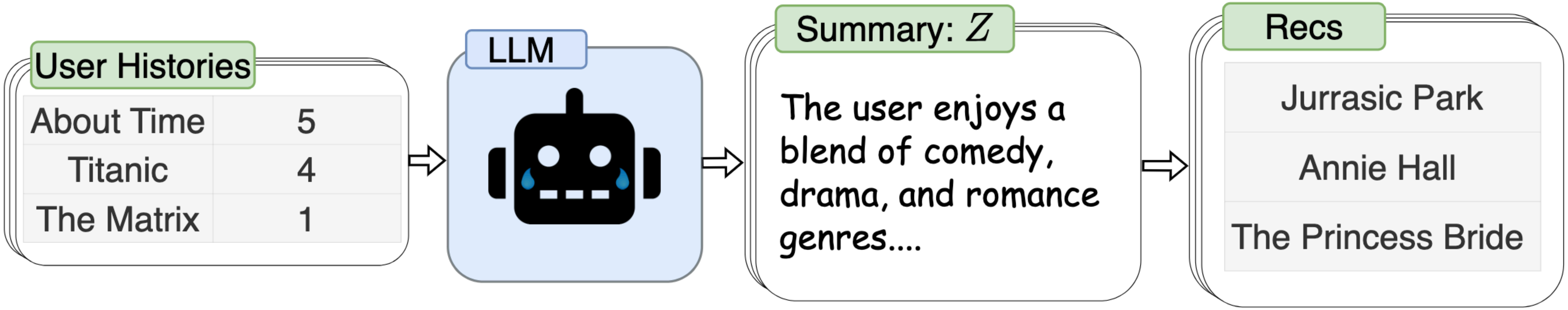
What can I help with?

- Large language models (LLMs) are an **interface** to AI capabilities
  - E.g., We can interact using text



Scrutable Recommender Systems (RecSys)





User Histories

About Time	5
Titanic	4
The Matrix	1

LLM

Summary: Z

The user enjoys a blend of comedy, drama, and romance genres....

Recs

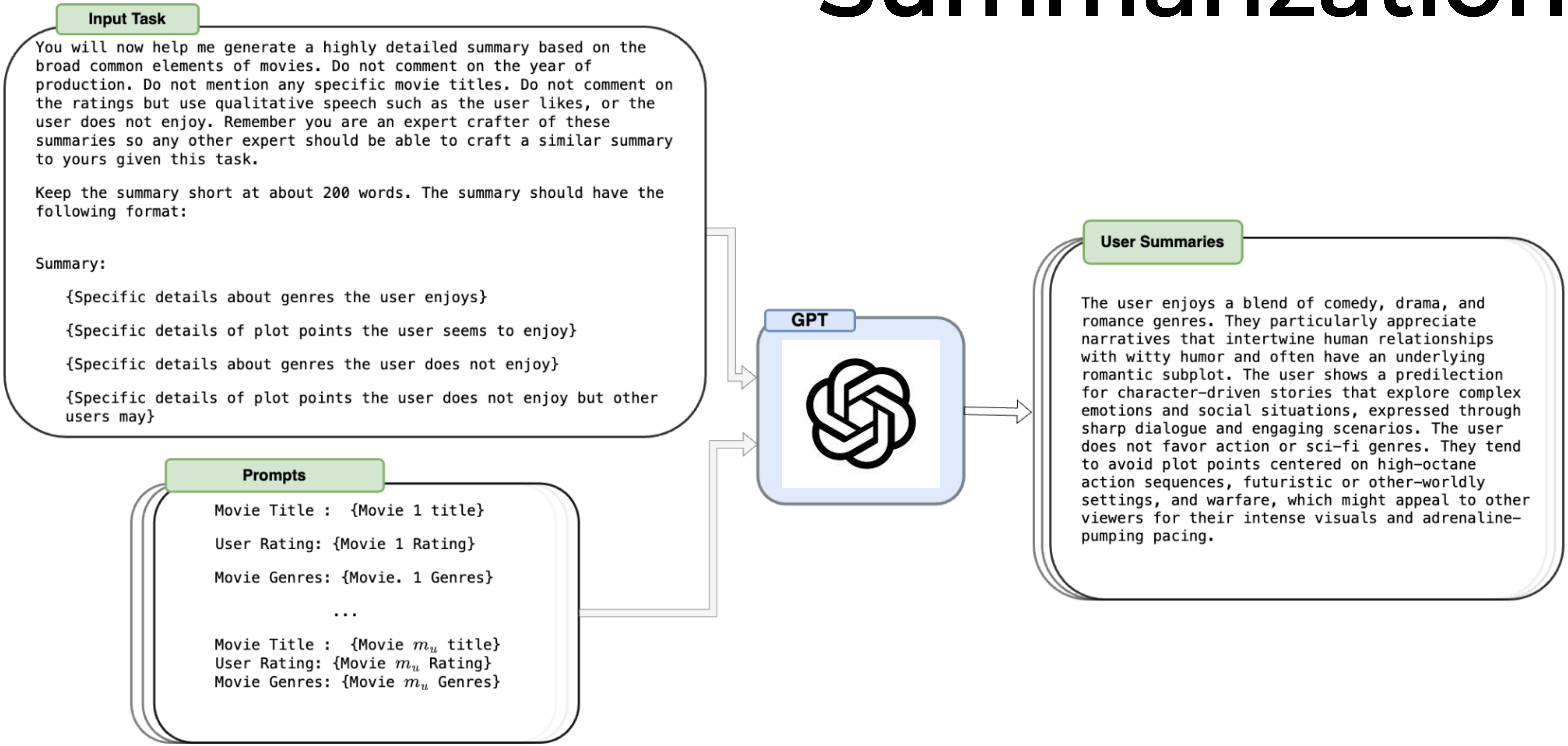
Jurrasic Park

Annie Hall

The Princess Bride

Conceptualized in:
On Natural Language User Profiles for Transparent and Scrutable Recommendatio
Radlinski et. al, SIGIR 2022

Summarization



Input Task

You will now help me generate a highly detailed summary based on the broad common elements of movies. Do not comment on the year of production. Do not mention any specific movie titles. Do not comment on the ratings but use qualitative speech such as the user likes, or the user does not enjoy. Remember you are an expert crafter of these summaries so any other expert should be able to craft a similar summary to yours given this task.

Keep the summary short at about 200 words. The summary should have the following format:

Summary:

{Specific details about genres the user enjoys}

{Specific details of plot points the user seems to enjoy}

{Specific details about genres the user does not enjoy}

{Specific details of plot points the user does not enjoy but other users may}

Prompts

Movie Title : {Movie 1 title}

User Rating: {Movie 1 Rating}

Movie Genres: {Movie. 1 Genres}

...

Movie Title : {Movie m_u title}
User Rating: {Movie m_u Rating}
Movie Genres: {Movie m_u Genres}

User Summaries

The user enjoys a blend of comedy, drama, and romance genres. They particularly appreciate narratives that intertwine human relationships with witty humor and often have an underlying romantic subplot. The user shows a predilection for character-driven stories that explore complex emotions and social situations, expressed through sharp dialogue and engaging scenarios. The user does not favor action or sci-fi genres. They tend to avoid plot points centered on high-octane action sequences, futuristic or other-worldly settings, and warfare, which might appeal to other viewers for their intense visuals and adrenaline-pumping pacing.

★★★★★

Little Women (1994)
Queen Margot (1994)
Age of Innocence, The (1993)
Trainspotting (1996)
My Left Foot (1989)
Dead Poets Society (1989)

★★★★

Sense and Sensibility (1995)
Othello (1995)
Eat Drink Man Woman (1994)
Immortal Beloved (1994)
In the Name of the Father (1993)
Emma (1996)

…

★★ ★

Sex, Lies, and Videotape (1989)
Ice Storm, The (1997)
Lolita (1997)
Drop Dead Gorgeous (1999)

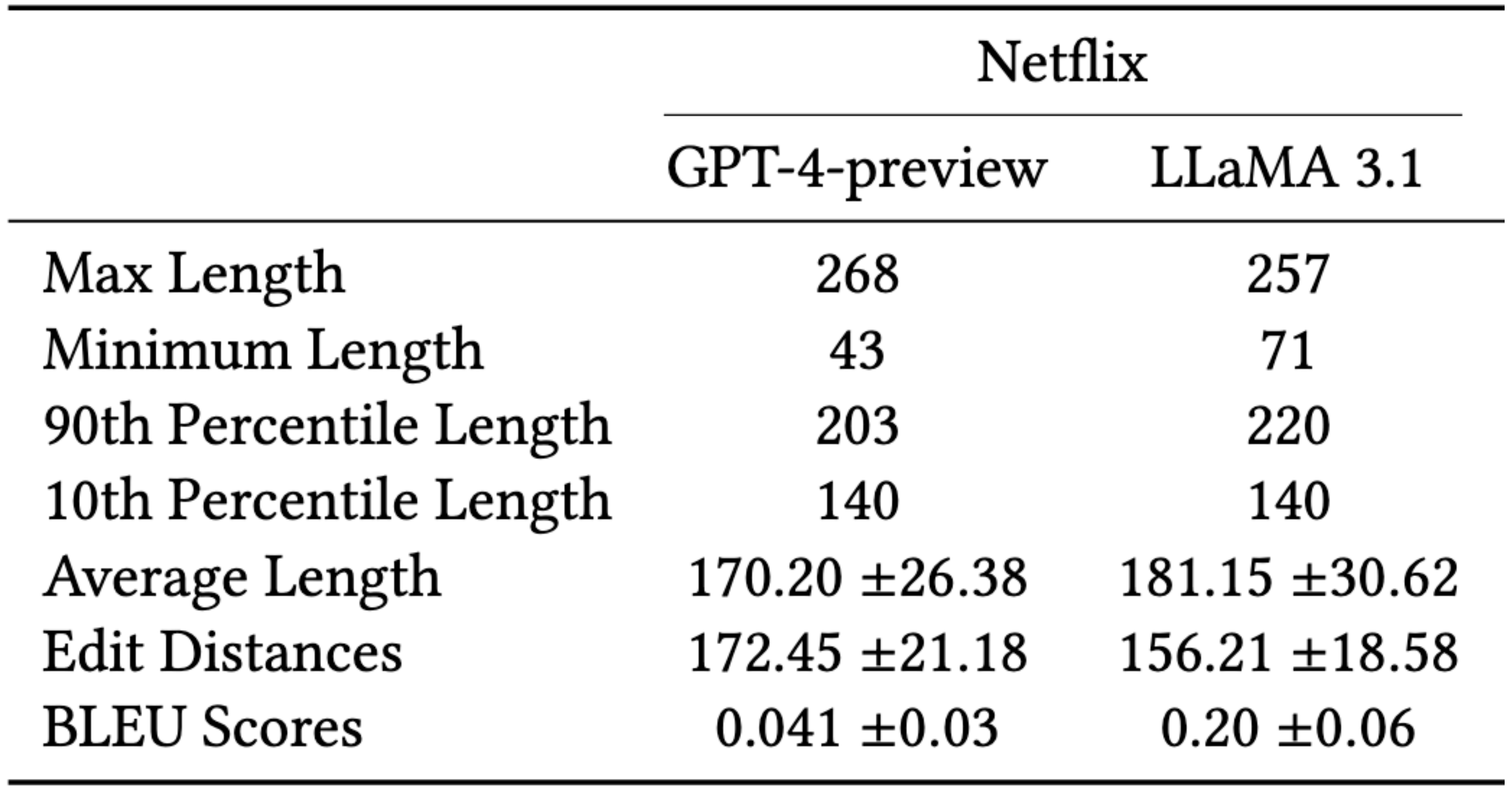
User Summary

The user enjoys dramas, especially those intertwined with romance and historical settings. Elements of war and the intricacies of familial relationships, as depicted in period pieces or literary adaptations, are also favored. Integrating comedy with drama, showcasing personal growth or societal commentary seems to resonate well. The user seems to enjoy plot points centered on character-driven narratives that involve emotional depth, personal conflict, and intimate relationships. Elements of fantasy or enchantment within a dramatic framework appear to appeal as well.

Conversely, the user does not enjoy certain types of dramas that perhaps focus on more modern or gritty realism,
such as those explicitly involving non-linear storytelling or controversial themes without a significant romance or historical context.

Plot points that revolve around explicit content, cold or clinical interpersonal dynamics, or lack the element of warmth found. in character connections are less appreciated. While some users may find ambiguity, high-intensity crime, and unconventional narrative structures intriguing, these do not seem to satisfy the preferences of this user.

Summaries are user-specific



	Netflix	
	GPT-4-preview	LLaMA 3.1
Max Length	268	257
Minimum Length	43	71
90th Percentile Length	203	220
10th Percentile Length	140	140
Average Length	170.20 ±26.38	181.15 ±30.62
Edit Distances	172.45 ±21.18	156.21 ±18.58
BLEU Scores	0.041 ±0.03	0.20 ±0.06

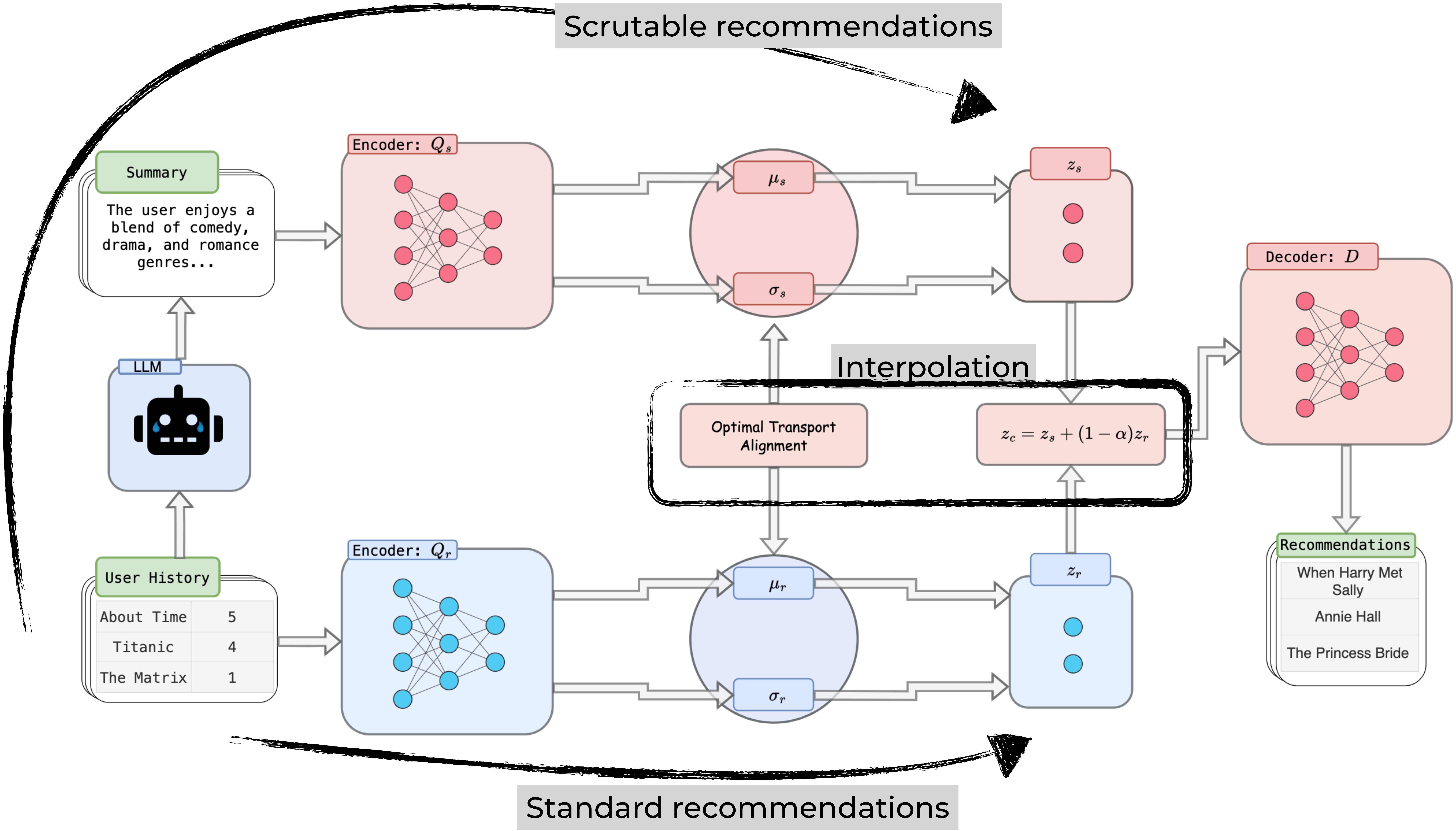
- **Similar results for a book dataset**

Recommendation performance

		Netflix	
	Model	Recall@20	NDCG@20
Standard Models	EASE	0.496	0.518
	RecVAE	0.515	0.540
Scrutable Models	TEARS Base (GPT)	0.465	0.491

Interpolation to obtain best of both worlds

- Large language models have ingested lots of information (the whole web!)
- Standard recommender systems are still better for modelling user preferences and recommendations
- Blend or interpolate to obtain:
  - High-quality recommendations from scrutable models
- **Idea: Align a standard model and TEARS in embedding space**


Scrutable recommendations
Summary
The user enjoys a blend of comedy, drama, and romance genres...
Encoder: Q_s
μ_s
σ_s
z_s
Decoder: D
LLM
Interpolation
Optimal Transport Alignment
$z_c = z_s + (1-\alpha)z_r$
User History
About Time 5
Titanic 4
The Matrix 1
Encoder: Q_r
μ_r
σ_r
z_r
Recommendations
When Harry Met Sally
Annie Hall
The Princess Bride
Standard recommendations

Objective

$$\mathcal{L} = \boxed{\mathcal{L}_{\text{R}}} + \lambda_1 \boxed{\mathcal{L}_{\text{OT}}} + \lambda_2 \boxed{\mathcal{L}_{\text{KL}}}$$

R: Cross-entropy of the recommendations

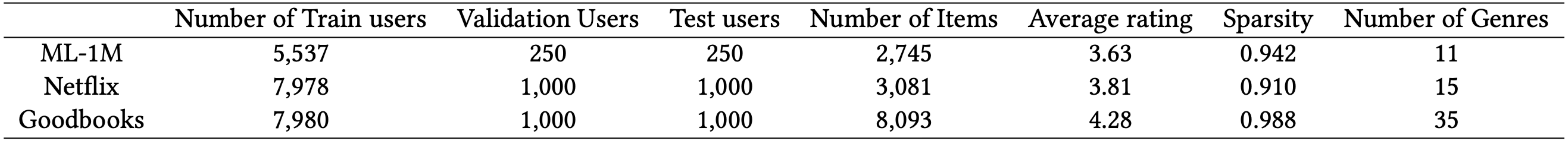
OT: Wasserstein distance between VAE and scrutable embeddings

KL: KL distance between prior and posterior over embeddings

Recommendation performance

	Model	Netflix	
		Recall@20	NDCG@20
Standard Models	EASE	0.496	0.518
	RecVAE	0.515	0.540
Scrutable Models	TEARS Base	0.465	0.491
	TEARS RecVAE	0.518	0.544

3 Datasets



	Number of Train users	Validation Users	Test users	Number of Items	Average rating	Sparsity	Number of Genres
ML-1M	5,537	250	250	2,745	3.63	0.942	11
Netflix	7,978	1,000	1,000	3,081	3.81	0.910	15
Goodbooks	7,980	1,000	1,000	8,093	4.28	0.988	35

- **2 movies, 1 books dataset**
- **Strong generalization**

	ML-1M				Netflix				Goodbooks			
Model	Recall@20	NDCG@20	Recall@50	NDCG@50	Recall@20	NDCG@20	Recall@50	NDCG@50	Recall@20	NDCG@20	Recall@50	NDCG@50
GPT-4-turbo	0.031	0.033	0.048	0.0390	0.054	0.067	0.065	0.040	0.015	0.012	0.013	0.011
EASE [49]	0.295	0.277	0.320	0.270	0.496	0.518	0.441	0.466	0.173	0.180	0.193	0.182
Multi-DAE [31]	0.290 ± 0.002	0.254 ± 0.001	0.363 ± 0.004	0.266 ± 0.000	0.507 ± 0.001	0.532 ± 0.001	0.450 ± 0.000	0.476 ± 0.001	0.151 ± 0.002	0.155 ± 0.002	0.173 ± 0.001	0.160 ± 0.001
GERS Base	0.276 ± 0.003	0.246 ± 0.001	0.320 ± 0.004	0.248 ± 0.000	0.471 ± 0.001	0.497 ± 0.001	0.413 ± 0.001	0.440 ± 0.001	0.153 ± 0.001	0.161 ± 0.001	0.167 ± 0.001	0.161 ± 0.001
⚙ TEARS Base	0.267 ± 0.004	0.253 ± 0.002	0.302 ± 0.014	0.250 ± 0.005	0.465 ± 0.004	0.491 ± 0.004	0.413 ± 0.003	0.439 ± 0.003	0.145 ± 0.001	0.153 ± 0.002	0.158 ± 0.002	0.153 ± 0.002
∞ TEARS Base	0.259 ± 0.010	0.249 ± 0.010	0.292 ± 0.015	0.245 ± 0.010	0.452 ± 0.002	0.479 ± 0.002	0.397 ± 0.001	0.424 ± 0.001	0.143 ± 0.002	0.151 ± 0.003	0.156 ± 0.002	0.151 ± 0.002
∞ TEARS RecVAE $_{\alpha=1}$	0.307 ± 0.006	0.272 ± 0.005	0.351 ± 0.007	0.276 ± 0.005	0.483 ± 0.002	0.509 ± 0.001	0.428 ± 0.002	0.455 ± 0.001	0.150 ± 0.002	0.160 ± 0.003	0.163 ± 0.001	0.159 ± 0.001
Multi-VAE [31]	0.295 ± 0.002	0.261 ± 0.001	0.357 ± 0.002*	0.270 ± 0.001	0.507 ± 0.001	0.532 ± 0.001	0.450 ± 0.000	0.476 ± 0.001	0.159 ± 0.001	0.163 ± 0.001	0.186 ± 0.001	0.170 ± 0.001
⚙ TEARS Multi-VAE$_{\alpha^*}$	0.295 ± 0.003	0.267 ± 0.002*	0.344 ± 0.010	0.272 ± 0.003	0.512 ± 0.001*	0.538 ± 0.001*	0.451 ± 0.000*	0.480 ± 0.000*	0.171 ± 0.002*	0.178 ± 0.002*	0.187 ± 0.003	0.178 ± 0.002*
∞ TEARS Multi-VAE$_{\alpha^*}$	0.306 ± 0.003*	0.276 ± 0.003*	0.347 ± 0.007	0.278 ± 0.003*	0.510 ± 0.001*	0.536 ± 0.001*	0.450 ± 0.001	0.479 ± 0.001*	0.169 ± 0.002*	0.174 ± 0.002*	0.187 ± 0.003	0.176 ± 0.002*
MacridVAE [33]	0.301 ± 0.007	0.260 ± 0.006	0.370 ± 0.002	0.276 ± 0.005	0.505 ± 0.003	0.529 ± 0.003	0.450 ± 0.002	0.476 ± 0.001	0.168 ± 0.001	0.170 ± 0.001	**0.196** ± 0.001	0.178 ± 0.001
⚙ TEARS MacridVAE$_{\alpha^*}$	**0.323 ± 0.004***	0.280 ± 0.004*	**0.381 ± 0.006***	**0.291 ± 0.003***	0.511 ± 0.001*	0.535 ± 0.002*	0.454 ± 0.002*	0.480 ± 0.002*	0.171 ± 0.002*	0.175 ± 0.002*	0.195 ± 0.002	0.180 ± 0.001*
∞ TEARS MacridVAE$_{\alpha^*}$	0.319 ± 0.004*	0.280 ± 0.002*	0.376 ± 0.003*	0.289 ± 0.001*	0.510 ± 0.001*	0.536 ± 0.001*	0.450 ± 0.001	0.479 ± 0.001*	0.169 ± 0.001	0.173 ± 0.001*	0.194 ± 0.002	0.179 ± 0.001
RecVAE [47]	0.300 ± 0.005	0.264 ± 0.003	0.360 ± 0.003	0.274 ± 0.003	0.515 ± 0.003	0.540 ± 0.003	0.455 ± 0.002	0.482 ± 0.002	0.171 ± 0.001	0.176 ± 0.001	0.191 ± 0.002	0.179 ± 0.001
GERS RecVAE $_{\alpha^*}$	0.304 ± 0.003*	0.266 ± 0.003*	0.366 ± 0.004*	0.279 ± 0.002*	0.517 ± 0.001*	0.542 ± 0.001*	**0.458 ± 0.001***	**0.485 ± 0.002***	0.170 ± 0.001	0.176 ± 0.001	0.192 ± 0.001	0.180 ± 0.001
⚙ TEARS RecVAE$_{\alpha^*}$	0.307 ± 0.002*	0.273 ± 0.002*	0.374 ± 0.002*	0.285 ± 0.001*	0.517 ± 0.001*	0.543 ± 0.000*	0.457 ± 0.001*	**0.485** ± 0.001*	**0.175± 0.002***	**0.181 ± 0.002***	0.193 ± 0.000*	**0.183± 0.001***
∞ TEARS RecVAE$_{\alpha^*}$	0.319 ± 0.005*	**0.282 ± 0.005***	0.363 ± 0.003*	0.287 ± 0.002*	**0.518 ± 0.001**	**0.544 ± 0.001***	0.457 ± 0.001*	**0.485 ± 0.001***	0.173 ± 0.001*	0.179 ± 0.001*	0.191 ± 0.002	0.181 ± 0.000*

Experimental Observations

- TEARS is an effective plug-in method
  - Consistently outperform its base model (RecVAE, Multi-VAE, MacridVAE)
- GPT and Llama summaries are equivalent
- TEARS outperforms Genre-based model (GERS) except on Netflix dataset

LLMs alone aren’t competitive

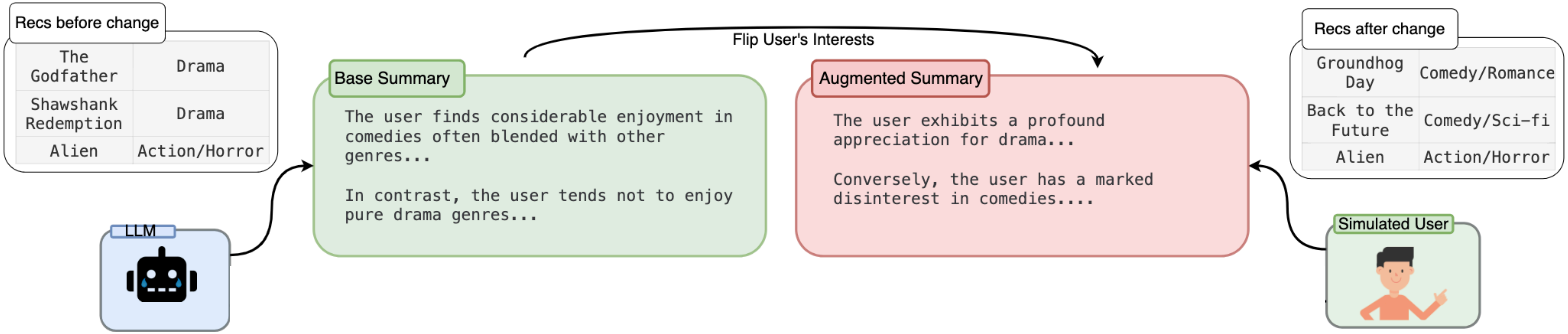
		Netflix	
	Model	Recall@20	NDCG@20
Standard Models	EASE	0.496	0.518
	RecVAE	0.515	0.540
Scrutable Models	TEARS Base (GPT)	0.465	0.491
	TEARS RecVAE	0.518	0.544
	GPT-4-Turbo	0.054	0.067

Scrutable Recsys are Controllable

- Three synthetic studies:

  1. *Large-scope Changes*: Change the ranks of groups of similar items (genre)
  2. *Small-scope Changes*: Change the rank of a specific item in the recommendation list
  3. *Guided recommendations:* Replace summary with current context

1. Large-scope Changes



Recs before change

The Godfather	Drama
Shawshank Redemption	Drama
Alien	Action/Horror

Flip User's Interests

Base Summary

`The user finds considerable enjoyment in comedies often blended with other genres...`

`In contrast, the user tends not to enjoy pure drama genres...`

Augmented Summary

`The user exhibits a profound appreciation for drama...`

`Conversely, the user has a marked disinterest in comedies....`

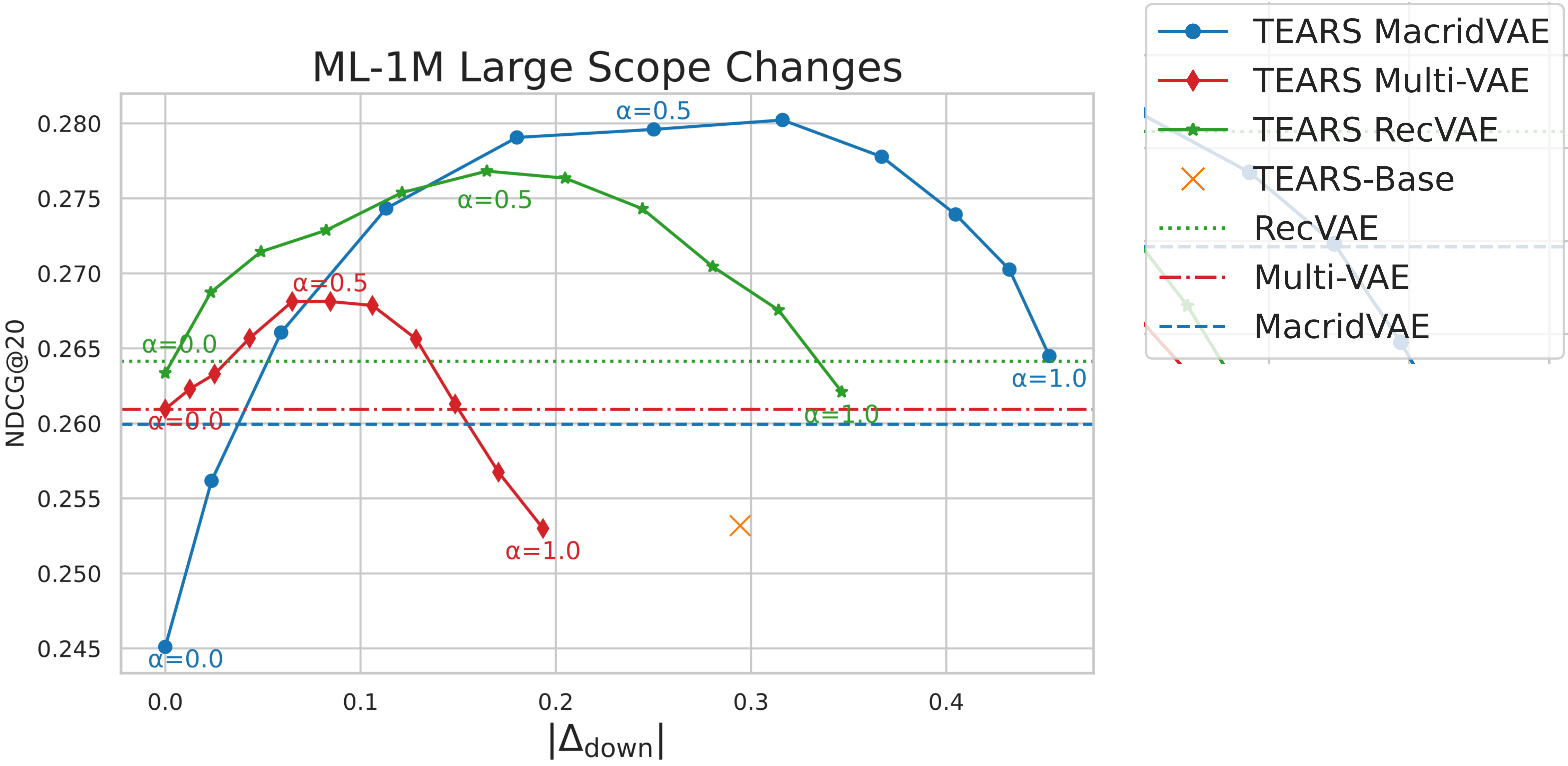
Recs after change

Groundhog Day	Comedy/Romance
Back to the Future	Comedy/Sci-fi
Alien	Action/Horror

How to measure changes?

- No ground truth information
- We develop a genre-based version of NDCG
- We evaluate the difference between the original recommendations and the new recommendations

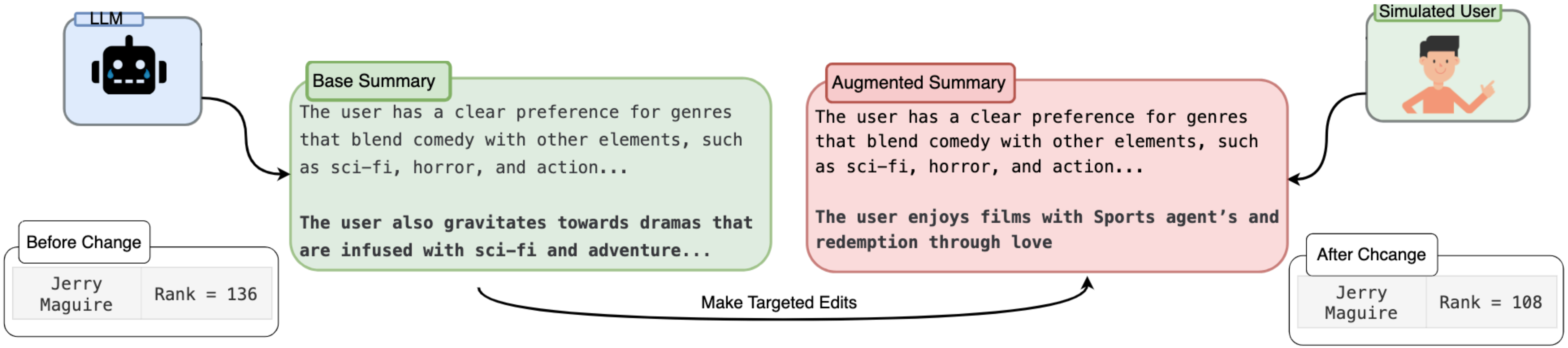
$$\Delta@k(\rho) = \mathrm{NDCG}_g^{\mathrm{O}}@k(\rho) - \mathrm{NDCG}_g^{\mathrm{A}}@k(\rho)$$



Recommendation Performance (NDCG@20)

Controllability
(Change in the genres of movies recommended)

2. Small-scope Changes



Base Summary

The user has a clear preference for genres that blend comedy with other elements, such as sci-fi, horror, and action...

The user also gravitates towards dramas that are infused with sci-fi and adventure...

Augmented Summary

The user has a clear preference for genres that blend comedy with other elements, such as sci-fi, horror, and action...

The user enjoys films with Sports agent's and redemption through love

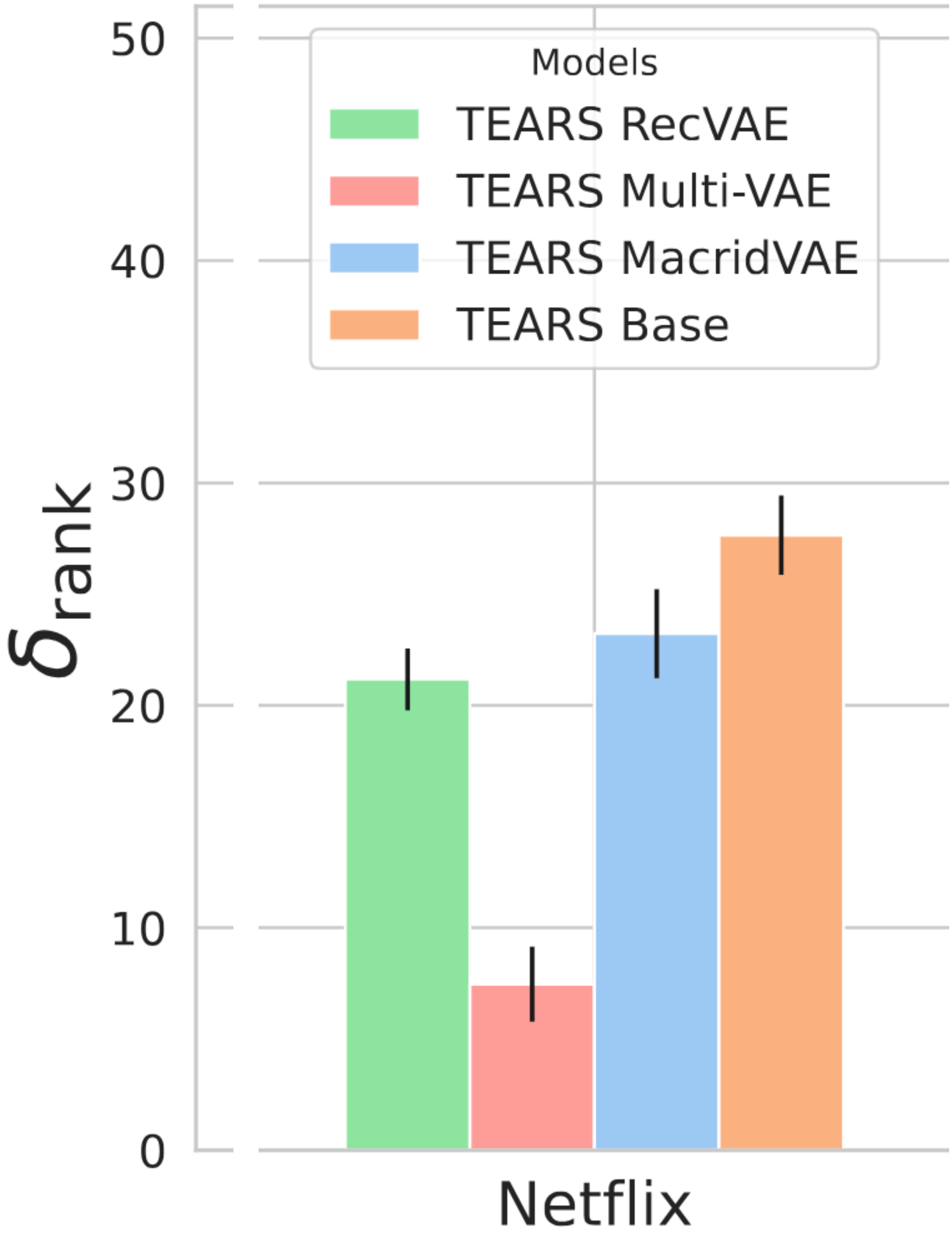
Before Change

Jerry Maguire	Rank = 136

After Chcange

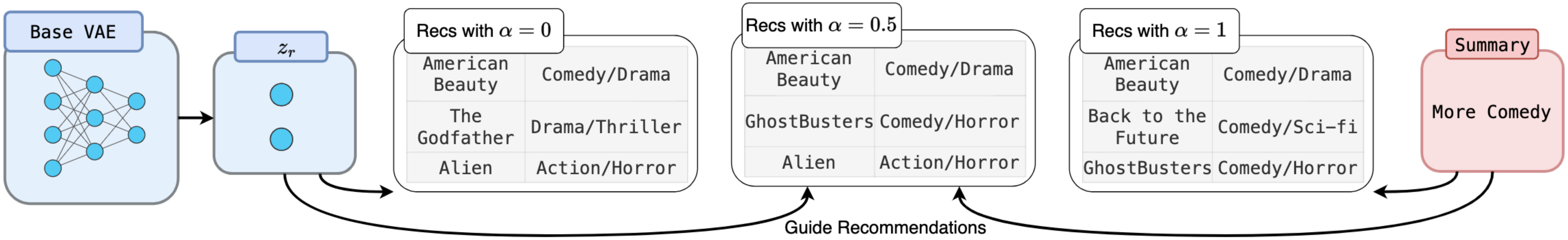
Jerry Maguire	Rank = 108

Change in rank to the target item caused by the change
In summary



3. Guided recommendations

- Simulate an interactive system where users can react to their recommendations
  - Replace the summary with their reaction (e.g. “More Comedy”)
  - We benefit from the interpolation to obtain personalized results



Recs with $\alpha = 0$

American Beauty	Comedy/Drama
The Godfather	Drama/Thriller
Alien	Action/Horror

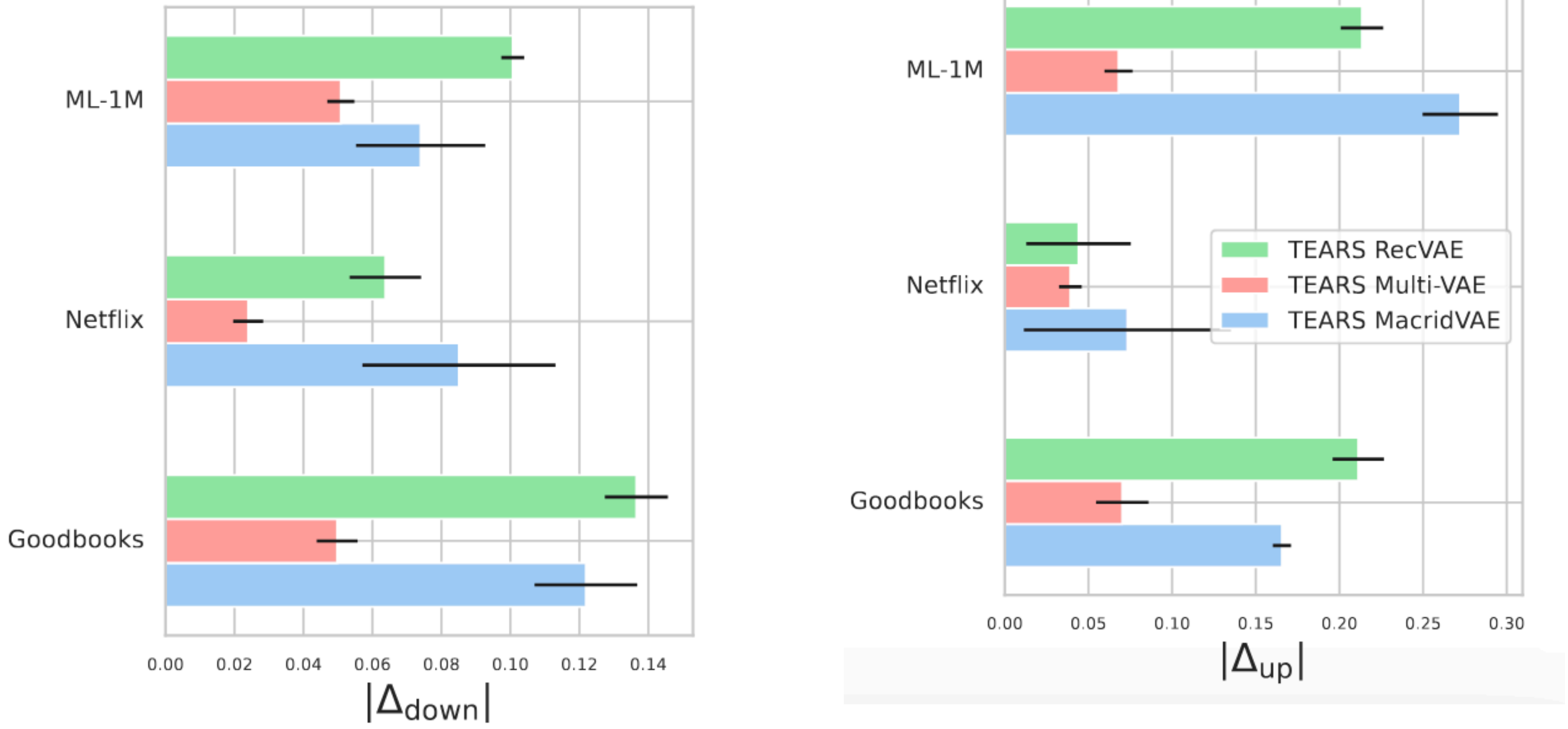
Recs with $\alpha = 0.5$

American Beauty	Comedy/Drama
GhostBusters	Comedy/Horror
Alien	Action/Horror

Recs with $\alpha = 1$

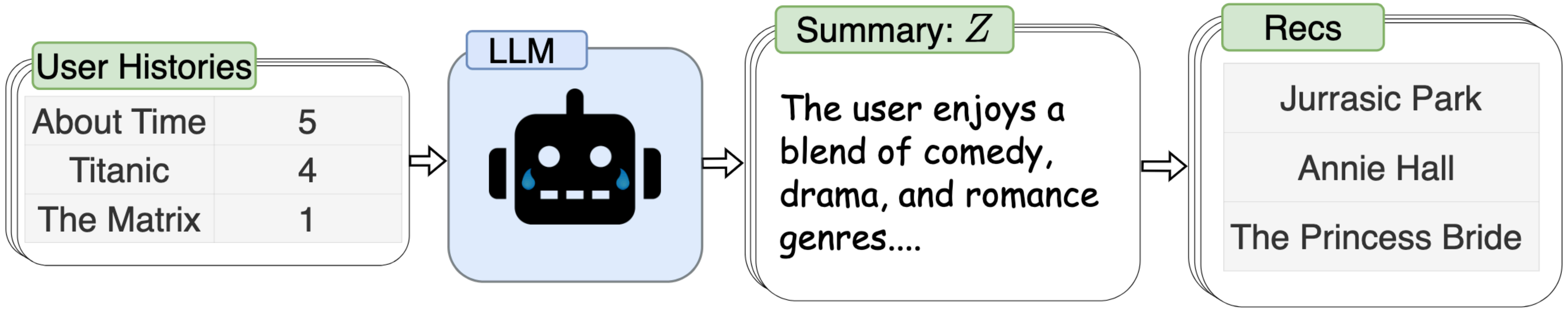
American Beauty	Comedy/Drama
Back to the Future	Comedy/Sci-fi
GhostBusters	Comedy/Horror

Guided Recommendation



Controllability
(Change in the genres of movies recommended)

Scrutable recommendations



User Histories

About Time	5
Titanic	4
The Matrix	1

⇨ **LLM** ⇨

Summary: Z

The user enjoys a blend of comedy, drama, and romance genres....

⇨ **Recs**

Jurrasic Park

Annie Hall

The Princess Bride

- **Good performance and controllable for movies and books**

- **Next:**
  - **Evaluate effectiveness with humans**

[Radlinksi et al., On Natural Language User Profiles for Transparent and Scrutable Recommendation., SIGIR'22]

Beyond recommendations

- Modern AI Systems are opaque... LLMs offer an interface
- Common limitation: The world is dynamic
  - User preferences (multi-resolution)
  - Item popularity, new items
- Scrutability over time?
  - Interactive scenarios (e.g., social media, conversation)

On Natural Language User Profiles for Transparent and Scrutable Recommendation
Radlinski et. al, SIGIR 2022

Scrutable Representations

- Modern AI techniques are **opaque**
- Scrutability (through text) provides an interface for human-AI interaction
  - Using a *text bottleneck* ensures the “summary” is correct
  - Could enable “model surgery”
  - Could it help against model jailbreaking?