SIGIR 2024
Washington, D.C.

Large Language Models for Recommendation: Past, Present, and Future

Lecture Tutorial For SIGIR 2024

Organizers: Keqin Bao, Jizhi Zhang, Xinyu Lin,
Yang Zhang, Wenjie Wang, Fuli Feng

Outline

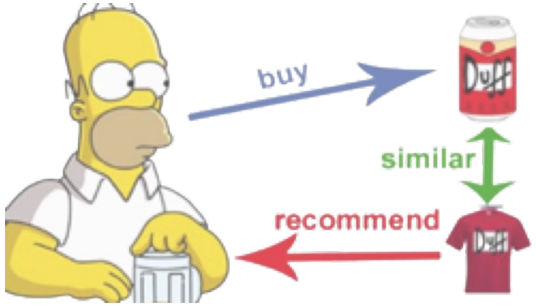
- Introduction (Wenjie Wang)
- Background: LM & LM4Rec (Yang Zhang)
- Development of LLMs (Keqin Bao)
- Progress of LLM4Rec
  - LLM4Rec (Keqin Bao & Wenjie Wang)
  - QA & Coffee Break
  - Trustworthy LLM4Rec (Jizhi Zhang)
- Open Problems (Xinyu Lin)
- Future Direction & Conclusions (Wenjie Wang)

Background of RecSys

❑ Information explosion era

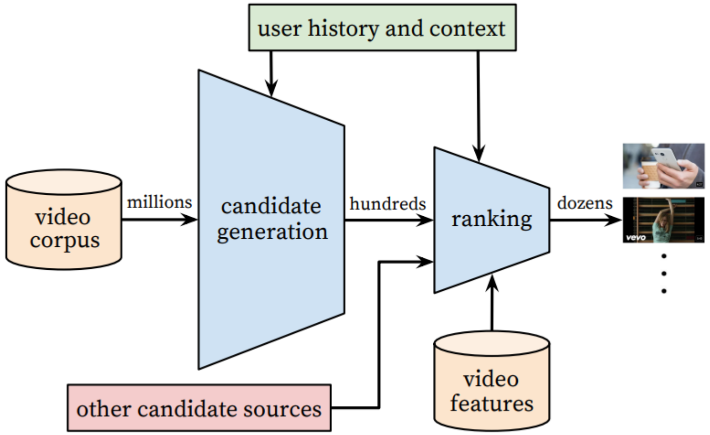
- E-commerce: 12 million items in Amazon.
- Social networks: 2.8 billion users in Facebook.
- Content sharing platforms: 720,000 hours videos uploaded to Youtube per day; 35 million videos posted on TikTok daily

❑ Recommender system



Information seeking via **user history feedback**

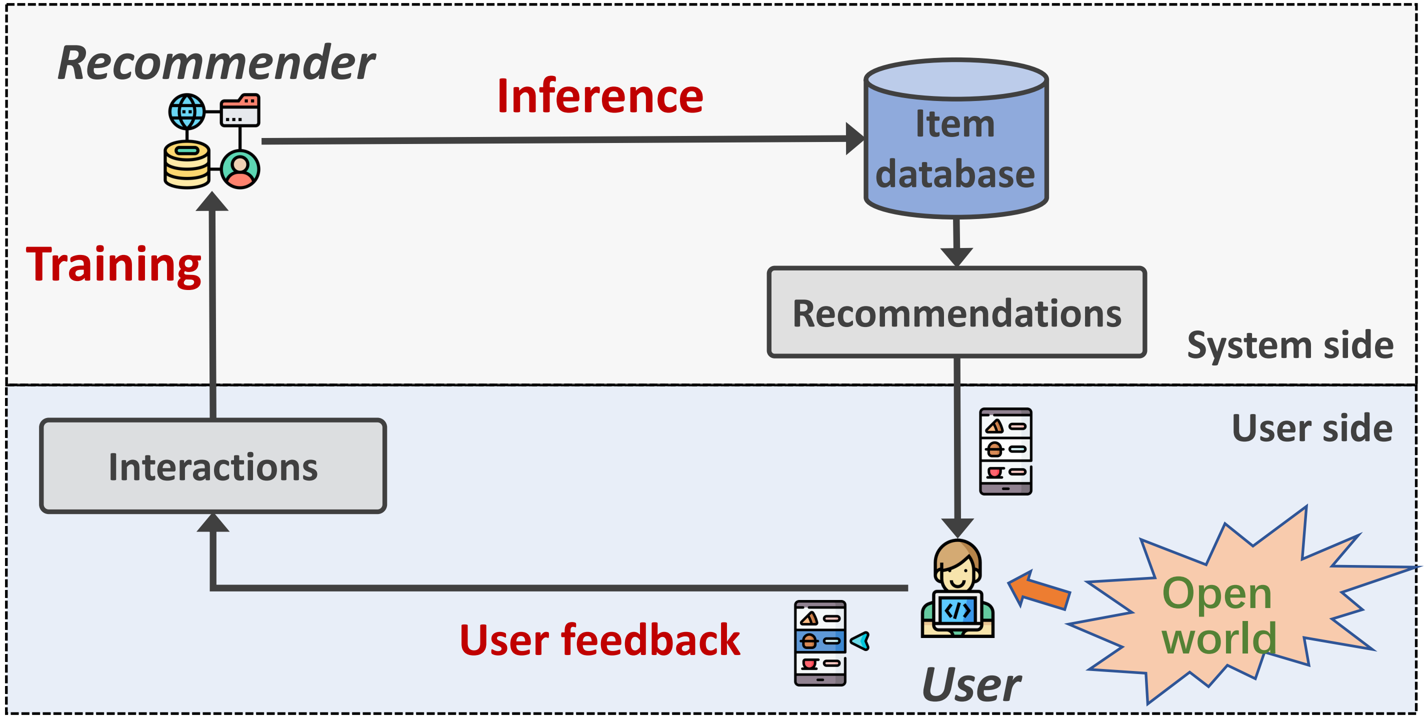
Recommendation



Images from: *Deep neural networks for youtube recommendations*

Background of RecSys

❑ Workflow of Recommender System

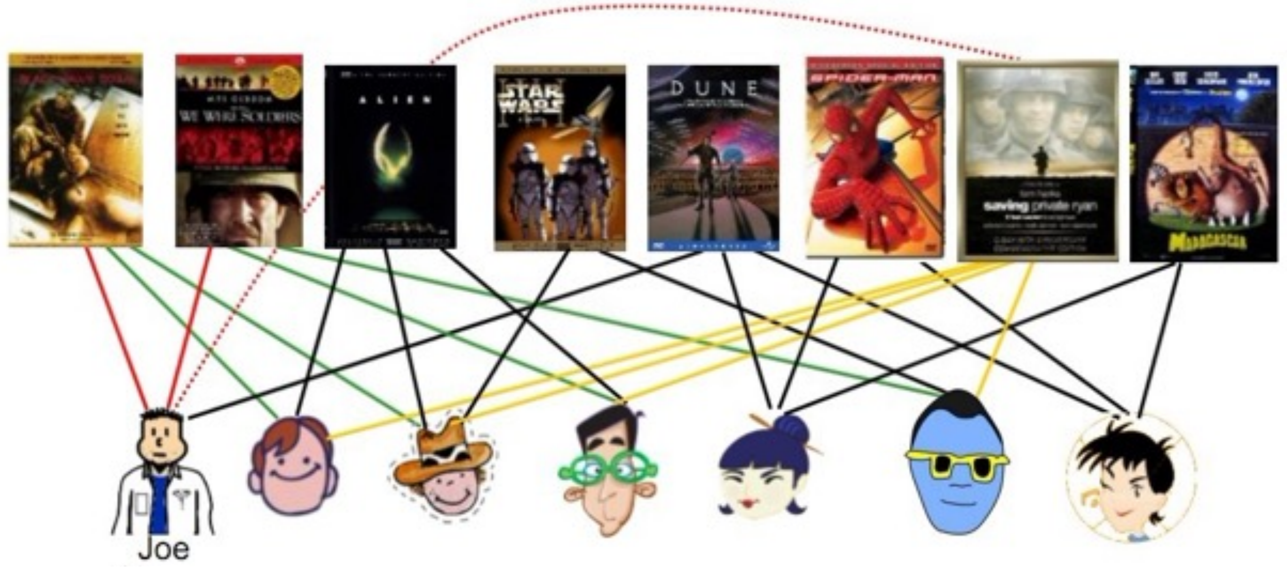


Background of RecSys

❑ Core idea of personalized recommendation

- **Collaborative filtering (CF):**

  Making automatic predictions (filtering) about the interests of a user by collecting preferences from many users (collaborating).



item

user	1	2	3	4	
1	5	?	?	?	…
2	3	4	?	?	…
3	?	1	2	4	…
	…	…	…	…	…

Interaction Matrix

Memory-based CF

- User CF
- Item CF

Model-based CF

- MF
- FISM

Neural CF

GNN-based CF

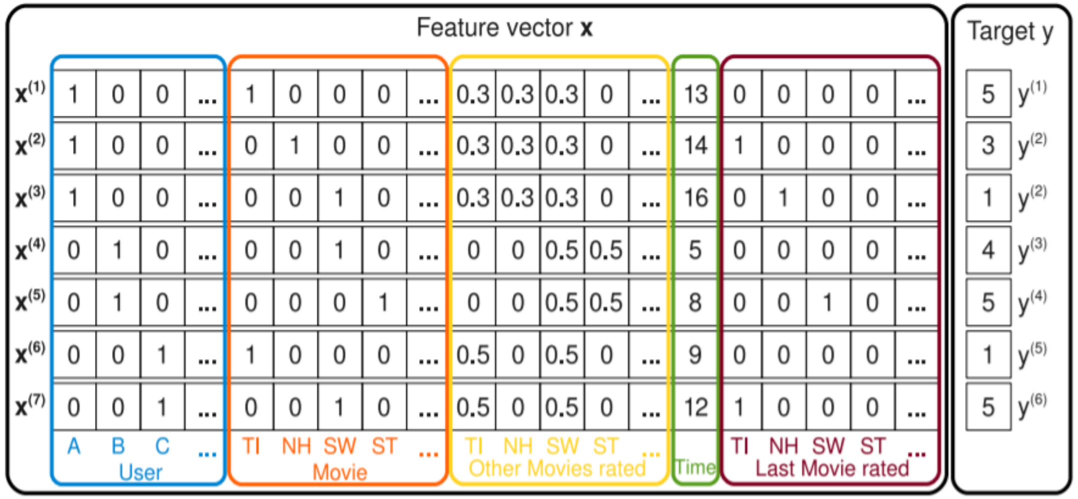
…

Images from: Neural Collaborative Filtering,

Background of RecSys

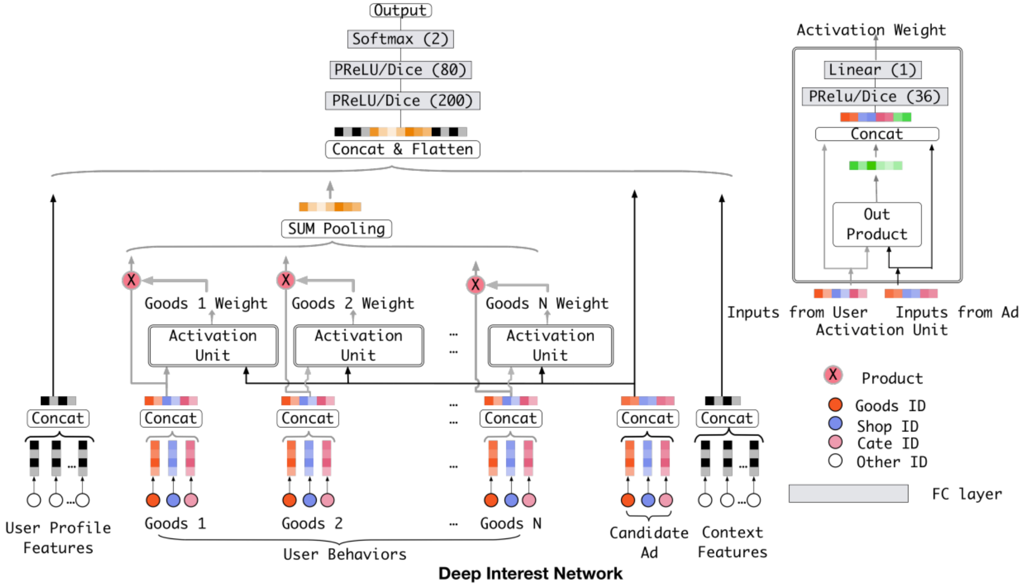
❑ Core idea of personalized recommendation

- Collaborative filtering (CF): collaborative information
- Content/context-aware models (CTR models): user/item side information + context
  - Click-Through Rate (CTR) prediction



	Feature vector **x**																						Target y	
	User				Movie					Other Movies rated					Time	Last Movie rated							
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...				
$x^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...			5	$y^{(1)}$
$x^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...			3	$y^{(2)}$
$x^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...			1	$y^{(2)}$
$x^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...			4	$y^{(3)}$
$x^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...			5	$y^{(4)}$
$x^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...			1	$y^{(5)}$
$x^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...			5	$y^{(6)}$

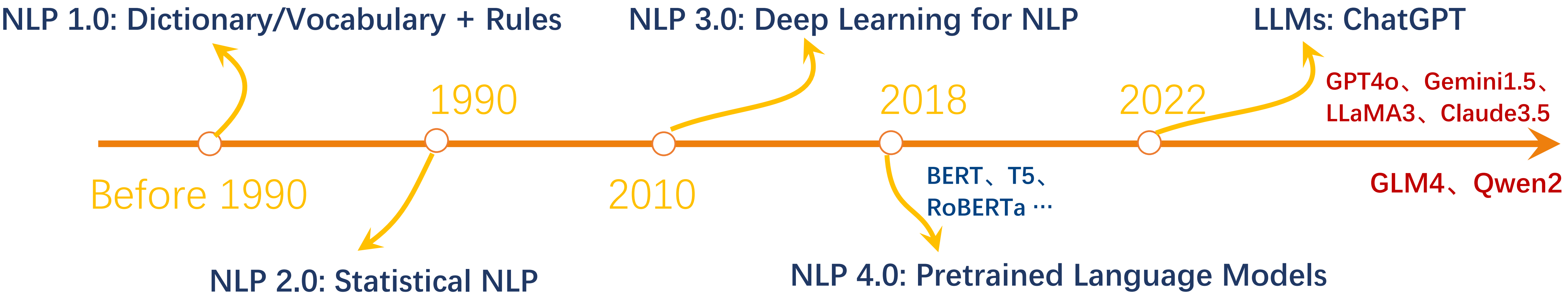
Factorization machines: FM, NFM, DeepFM



Neural network: DIN, AutoInt

Images from: Deep Interest Network for Click-Through Rate Prediction

The development of LMs



NLP 1.0: Dictionary/Vocabulary + Rules

NLP 3.0: Deep Learning for NLP

LLMs: ChatGPT

1990

2018

2022

GPT4o、Gemini1.5、
LLaMA3、Claude3.5

Before 1990

2010

BERT、T5、
RoBERTa ···

GLM4、Qwen2

NLP 2.0: Statistical NLP

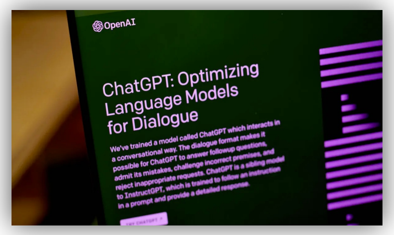
NLP 4.0: Pretrained Language Models

Large Language Model：billions of parameters, emergent capabilities

- Rich knowledge & Language Capabilities
- Instruction following
- In-context learning
- Chain-of-thought
- Planning
- ...

The development of LMs

- □ LM for recommendation (LM4Rec)
    - ❑ Model structure: BERT4Rec, SASRec, Recformer ...
    - ❑ LM as item encoder: UniSRec, VQRec, MoRec ...
    - ❑ Task formulation and learning paradigm: P5, VIP5, M6-Rec, Prompt4NR ...
- □ LLMs such as ChatGPT and GPT4 have influenced many fields in CS and IT industry.



OpenAI

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT

→

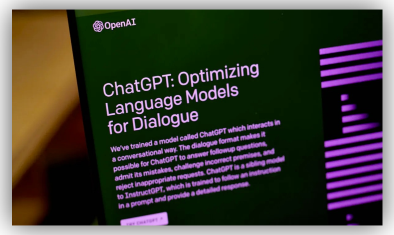
New Bing

Recommender System + LLMs?

➔ LLM4Rec

The development of LMs

- □ **LLMs such as ChatGPT and GPT4 have influenced many fields in CS and IT industry**
  - □ They have eliminated a wide range of research in basic NLP and conversational system, etc.



OpenAI

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT

+5 more
don depends on your preferences a
g (March to May) and the autumn
ant seasons, with mild temperatures
s that the summer (June to August). 😊
When is the best ti

New Bing

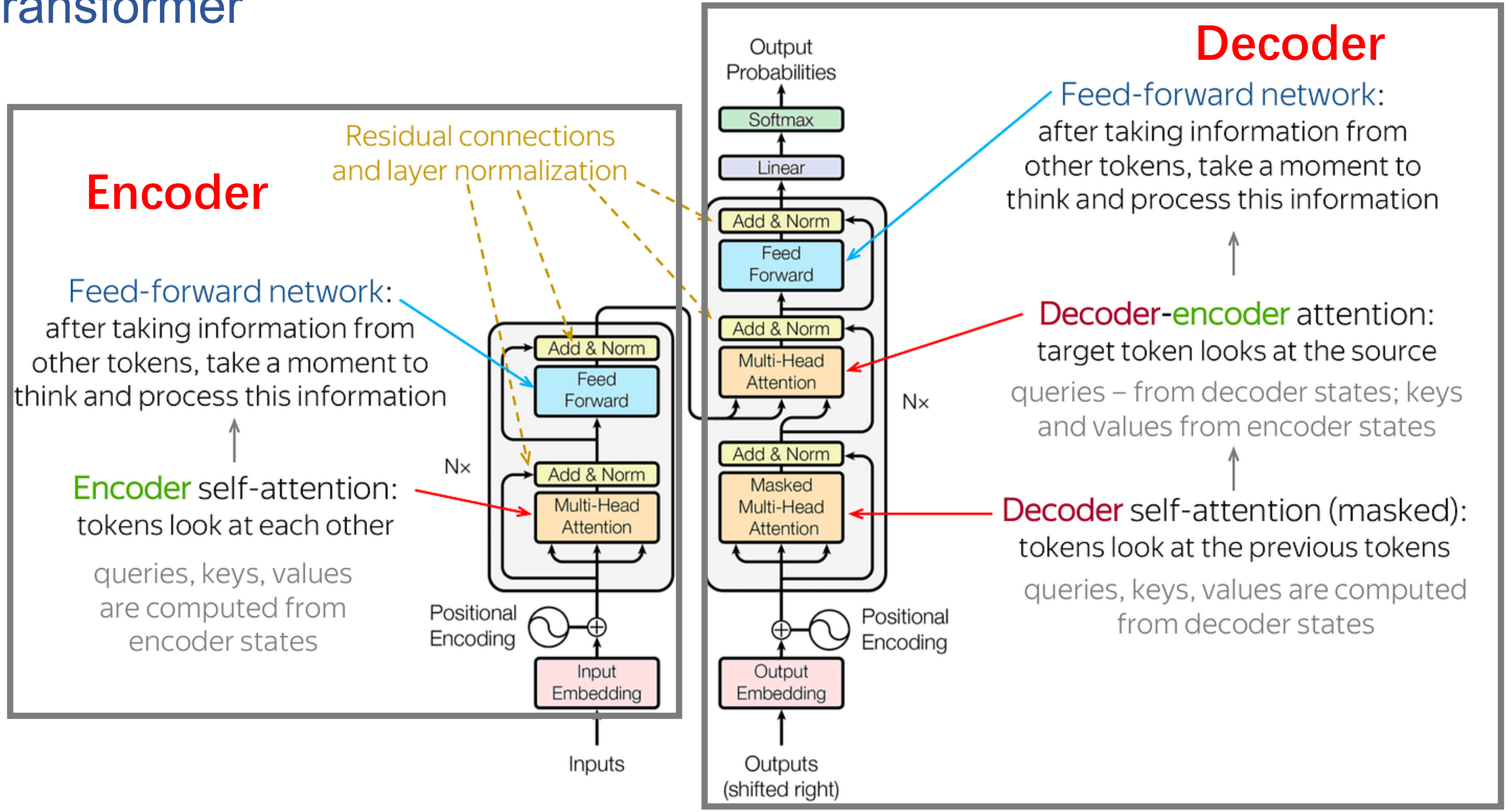
Recommender System + LLMs?

Outline

- Introduction
- **Background: LM & LM4Rec**
- Development of LLMs
- Progress of LLM4Rec
- Open Problems
- Future Direction & Conclusions

Development of LMs

❑ Transformer



Encoder

Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

Decoder

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

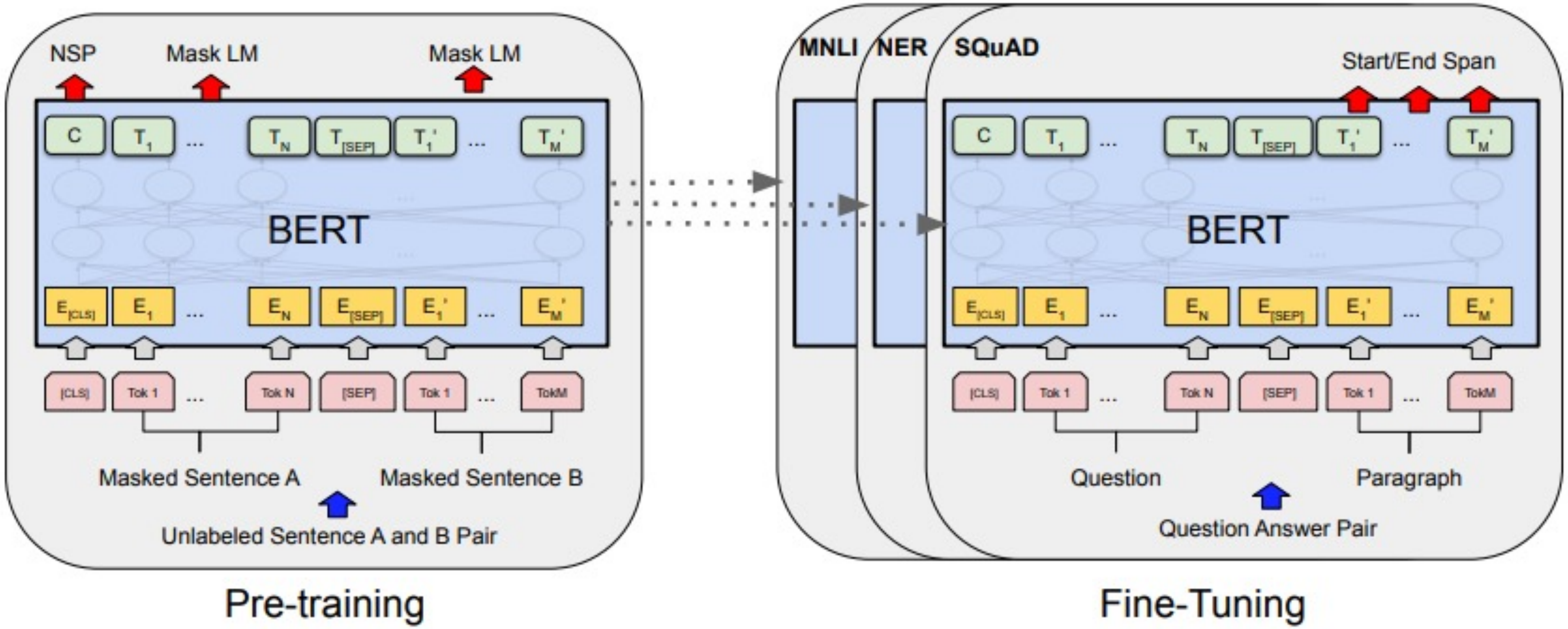
Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Ashish Vaswani et al. " Attention is All You Need“. NIPS 2017.

Development of LMs

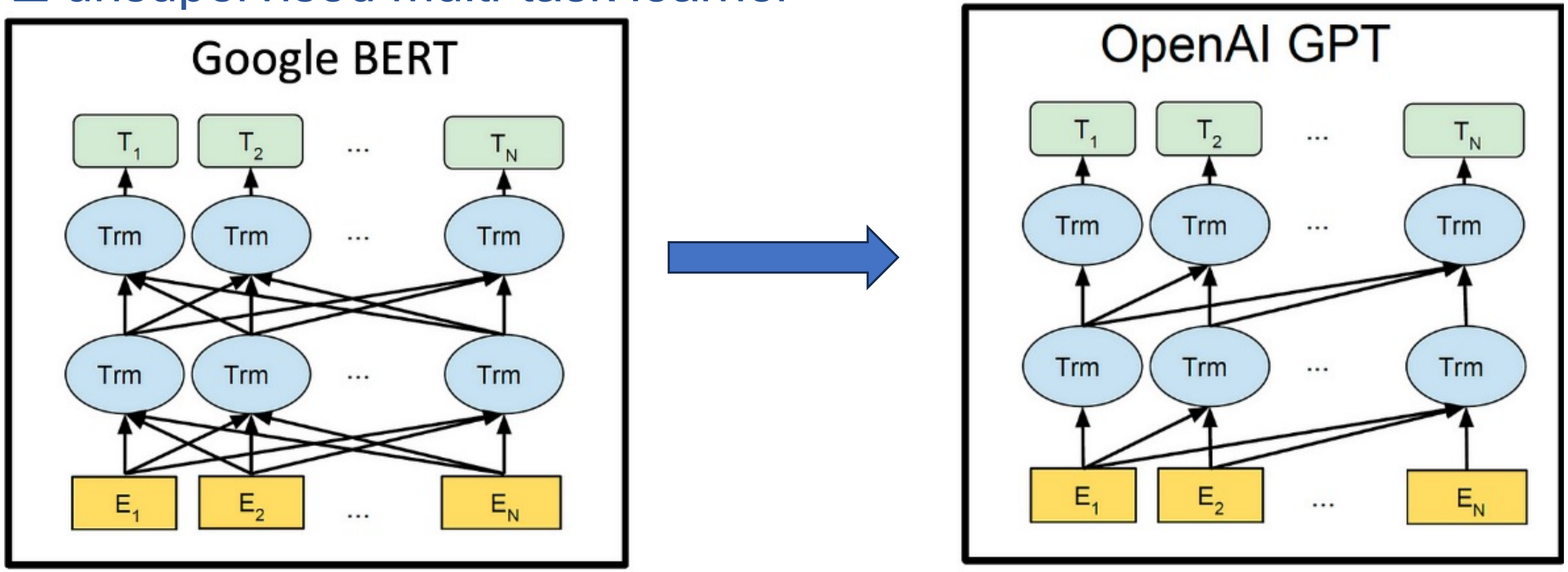
- ❑ Bert: pre-training of deep bidirectional transformers
  - ❑ Mask Language Modeling，bi-direction
  - ❑ Encoder （advantage）--> understanding



Development of LMs

- ❑ GPT2: generative pre-trained transformer
  - ❑ Causal language modeling
  - ❑ Decoder （advantage） --> Generation
  - ❑ unsupervised multi-task learner

$$p(x) = \prod_{i=1}^{n} p(s_n|s_1, ..., s_{n-1})$$



Alec Radford et al. Language Models are Unsupervised Multitask Learners. 2018.

Benefit of LMs

❑ How can recommender systems benefit from LMs

- **Model architecture:** Transformer、Self-attention

- **Task formulation**

  Use language to formulate the recommendation task

- **Representation:**

  Textual feature,

  item representation,

  knowledge representation

- **Learning paradigm:** Pretrain-finetune, Prompt learning

Overview of LM4rec

- LMs for recommendation
  - ❑ **Utilizing LMs' model structure for recommendation.**
    - ❑ ID-based: **BERT4Rec**, SASRec ...
    - ❑ Text-based: **Recformer** ...
  - ❑ **LM as item encoder. UniSRec, VQRec, MoRec ...**
  - ❑ **Recommendation as natural language processing.**
    - ❑ ID-based: **P5,** VIP5 ...
    - ❑ Text-based: **M6-Rec**, Prompt4NR ...

Utilizing LM Model Structure

❑ Bert4Rec: ID-based

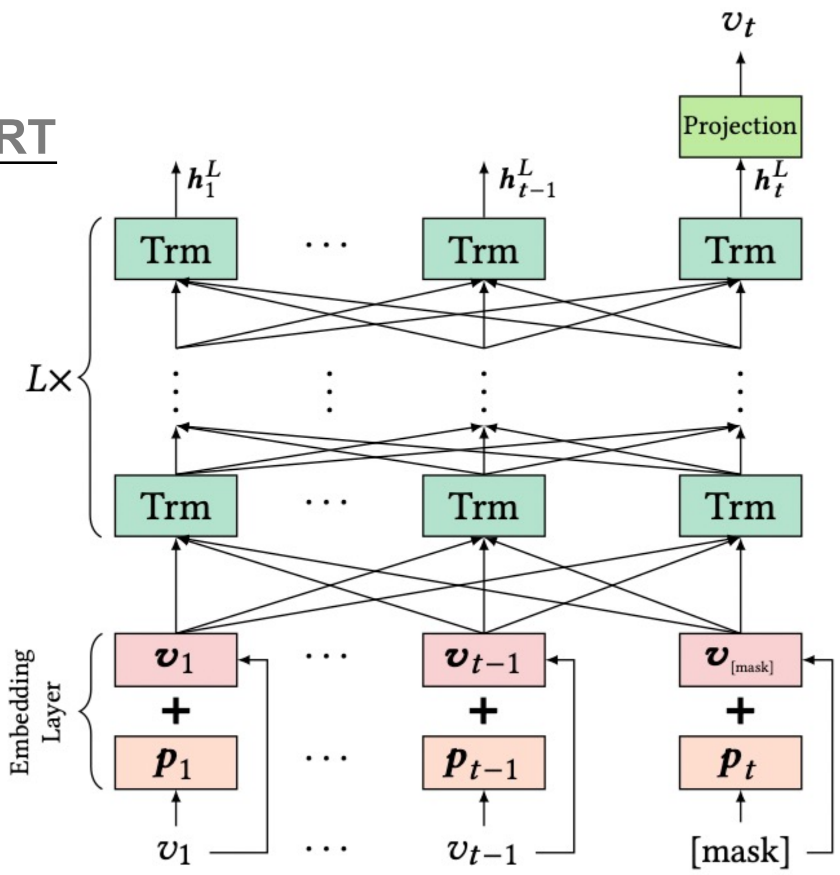
Natural Language:

- Token sequence
- Inter-token correlations

RecSys:

- ID sequence
- Inter-item correlations

BERT



(b) BERT4Rec model architecture.

Next-item prediction

Self-attention

Item ID embedding

Position embedding

Training recommender by masked item prediction as BERT.

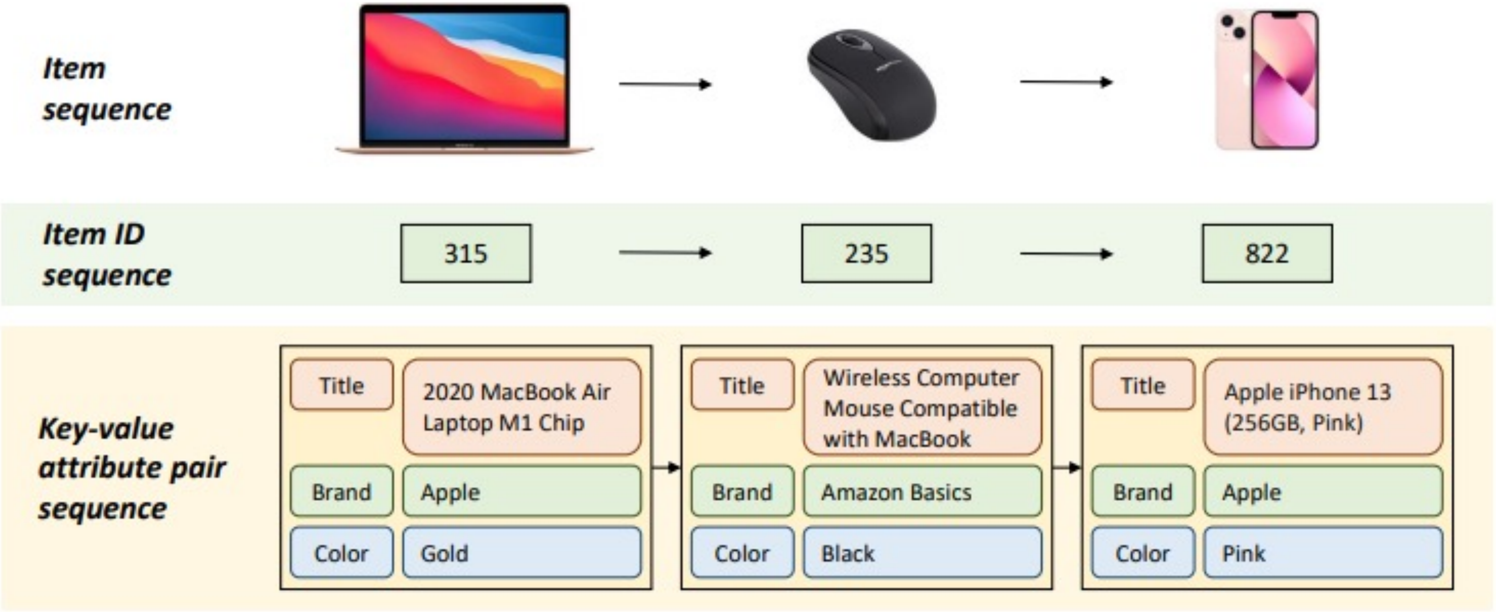
Sun, Fei, et al. "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer." CIKM. 2019.

Utilizing LM Model Structure

❑ Recformer: text-based

❑ Text is all you need (NO item ID)

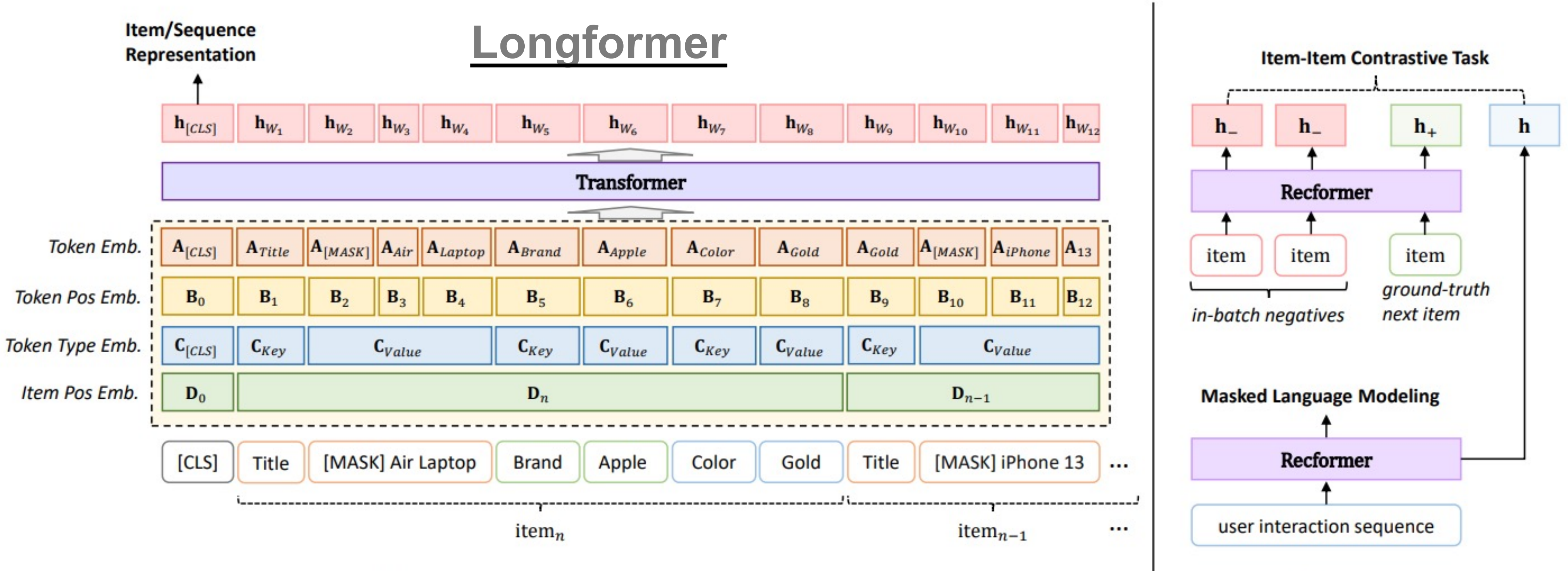
- Only use texts to represent items.
- Low resource, better cold-start recommendation.


Item sequence
Item ID sequence
315
235
822
Key-value attribute pair sequence
Title
2020 MacBook Air Laptop M1 Chip
Brand
Apple
Color
Gold
Title
Wireless Computer Mouse Compatible with MacBook
Brand
Amazon Basics
Color
Black
Title
Apple iPhone 13 (256GB, Pink)
Brand
Apple
Color
Pink

Li Jiacheng et al. "Text Is All You Need: Learning Language Representations for Sequential Recommendation“ KDD 2023.

Utilizing LM Model Structure

- ❑ **Recformer: text-based**
  - ❑ **Text is all you need (NO item ID)**


Longformer
Item/Sequence Representation
$\mathbf{h}_{[CLS]}$ $\mathbf{h}_{W_1}$ $\mathbf{h}_{W_2}$ $\mathbf{h}_{W_3}$ $\mathbf{h}_{W_4}$ $\mathbf{h}_{W_5}$ $\mathbf{h}_{W_6}$ $\mathbf{h}_{W_7}$ $\mathbf{h}_{W_8}$ $\mathbf{h}_{W_9}$ $\mathbf{h}_{W_{10}}$ $\mathbf{h}_{W_{11}}$ $\mathbf{h}_{W_{12}}$
Transformer
Token Emb.
$\mathbf{A}_{[CLS]}$ $\mathbf{A}_{Title}$ $\mathbf{A}_{[MASK]}$ $\mathbf{A}_{Air}$ $\mathbf{A}_{Laptop}$ $\mathbf{A}_{Brand}$ $\mathbf{A}_{Apple}$ $\mathbf{A}_{Color}$ $\mathbf{A}_{Gold}$ $\mathbf{A}_{Gold}$ $\mathbf{A}_{[MASK]}$ $\mathbf{A}_{iPhone}$ $\mathbf{A}_{13}$
Token Pos Emb.
$\mathbf{B}_0$ $\mathbf{B}_1$ $\mathbf{B}_2$ $\mathbf{B}_3$ $\mathbf{B}_4$ $\mathbf{B}_5$ $\mathbf{B}_6$ $\mathbf{B}_7$ $\mathbf{B}_8$ $\mathbf{B}_9$ $\mathbf{B}_{10}$ $\mathbf{B}_{11}$ $\mathbf{B}_{12}$
Token Type Emb.
$\mathbf{C}_{[CLS]}$ $\mathbf{C}_{Key}$ $\mathbf{C}_{Value}$ $\mathbf{C}_{Key}$ $\mathbf{C}_{Value}$ $\mathbf{C}_{Key}$ $\mathbf{C}_{Value}$ $\mathbf{C}_{Key}$ $\mathbf{C}_{Value}$
Item Pos Emb.
$\mathbf{D}_0$ $\mathbf{D}_n$ $\mathbf{D}_{n-1}$
[CLS] Title [MASK] Air Laptop Brand Apple Color Gold Title [MASK] iPhone 13 ...
item_n item_{n-1} ...
Item-Item Contrastive Task
$\mathbf{h}_-$ $\mathbf{h}_-$ $\mathbf{h}_+$ $\mathbf{h}$
Recformer
item item item
in-batch negatives
ground-truth next item
Masked Language Modeling
Recformer
user interaction sequence

(a) Recformer Model Structure

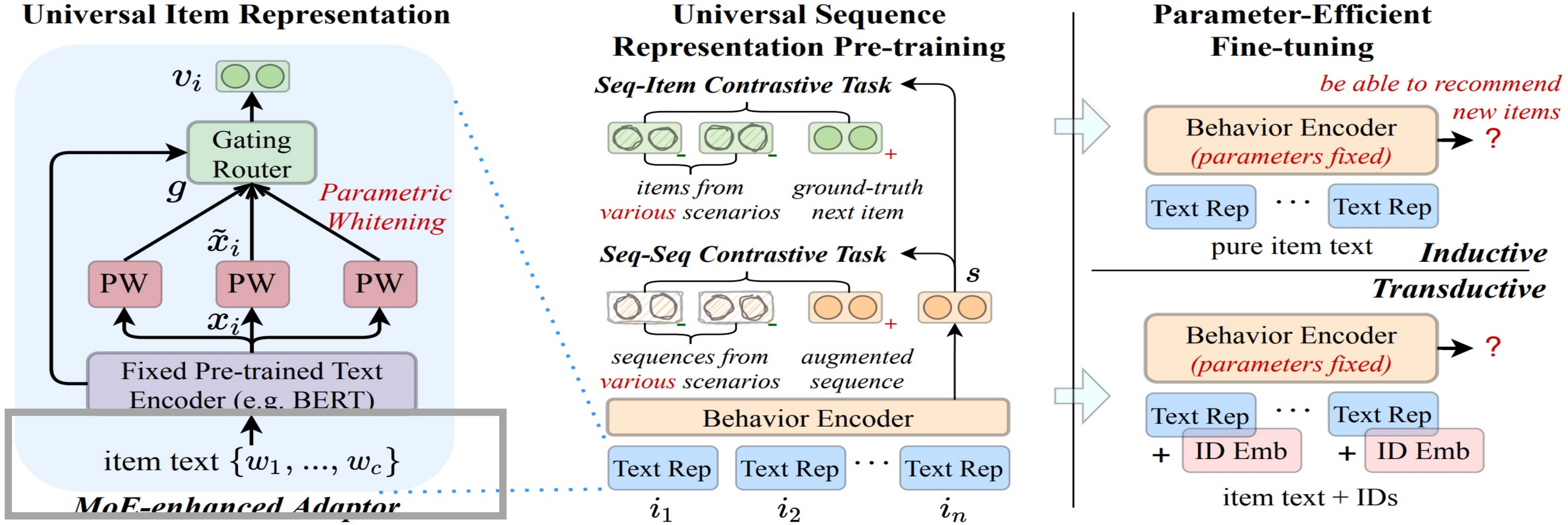
(b) Pretraining

Li Jiacheng et al. "Text Is All You Need: Learning Language Representations for Sequential Recommendation“ KDD 2023.

LM as Text Encoder

❑ UniSRec

- ❑ Enhance the recommendatoin model by using LMs to encode the natural language representation of items.



Universal Item Representation
v_i
Gating Router
g
Parametric Whitening
$\tilde{x}_i$
PW
PW
PW
x_i
Fixed Pre-trained Text Encoder (e.g. BERT)
item text $\{w_1, \dots, w_c\}$
MoE-enhanced Adaptor
Universal Sequence Representation Pre-training
Seq-Item Contrastive Task
items from various scenarios
ground-truth next item
Seq-Seq Contrastive Task
s
sequences from various scenarios
augmented sequence
Behavior Encoder
Text Rep
Text Rep
Text Rep
i_1
i_2
i_n
Parameter-Efficient Fine-tuning
be able to recommend new items
Behavior Encoder (parameters fixed)
?
Text Rep
Text Rep
pure item text
Inductive
Transductive
Behavior Encoder (parameters fixed)
?
Text Rep
Text Rep
+ ID Emb
+ ID Emb
item text + IDs

Figure 1: The overall framework of the proposed universal sequence representation learning approach (UniSRec).

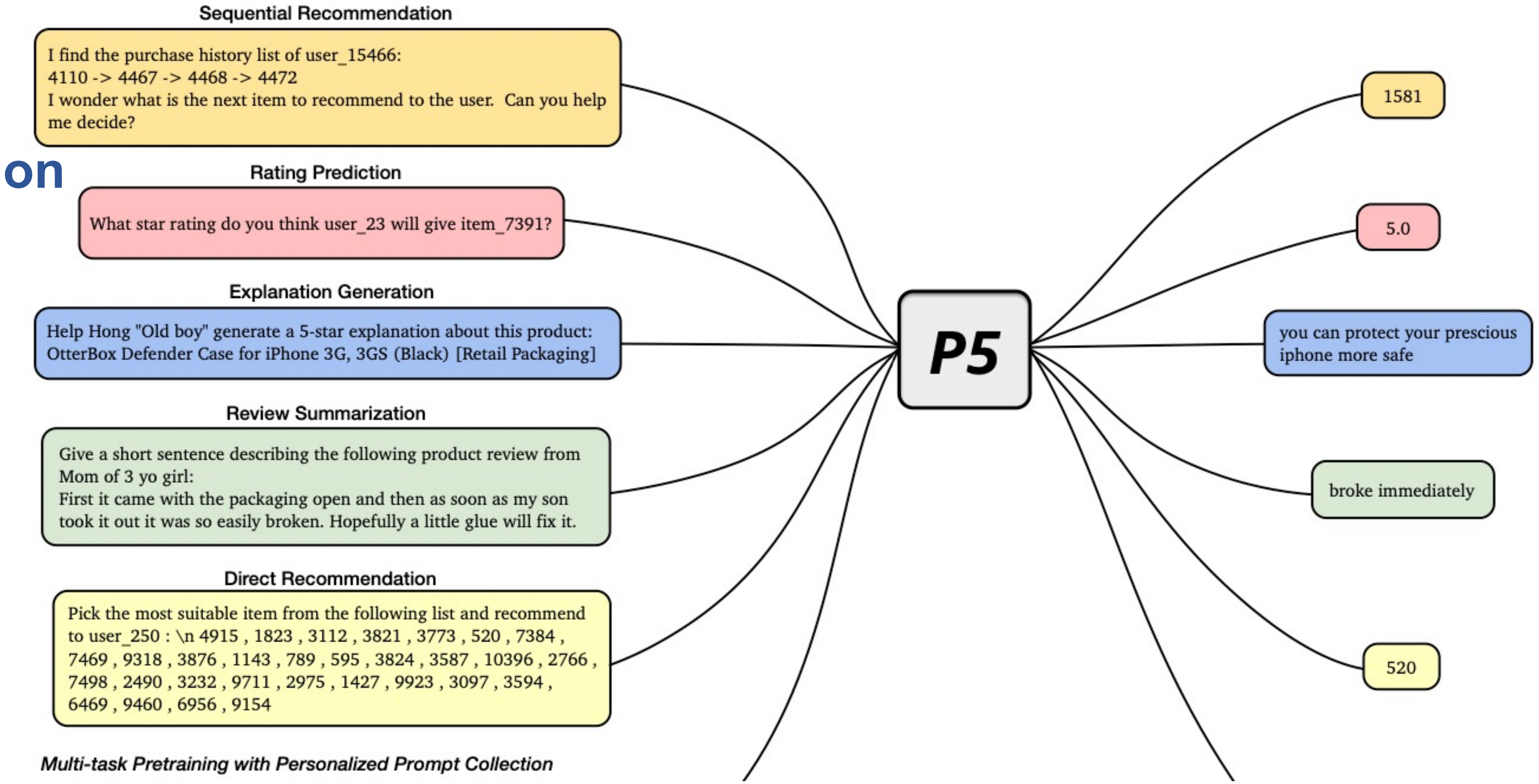
Yupeng Hou, et al. "Towards Universal Sequence Representaion Learning for Recommendation Systems." SIGKDD. 2022.

Recommendation as NLP

❑ P5: use natural language to describe different rec. tasks.

❑ **Multi-task prompts**

- ➢ Sequential recommendation
- ➢ Rating prediction
- ➢ Explain generation
- ➢ Review summarization
- ➢ Direct recommendation



Sequential Recommendation

I find the purchase history list of user_15466:
4110 -> 4467 -> 4468 -> 4472
I wonder what is the next item to recommend to the user. Can you help me decide?

Rating Prediction

What star rating do you think user_23 will give item_7391?

Explanation Generation

Help Hong "Old boy" generate a 5-star explanation about this product: OtterBox Defender Case for iPhone 3G, 3GS (Black) [Retail Packaging]

Review Summarization

Give a short sentence describing the following product review from Mom of 3 yo girl:
First it came with the packaging open and then as soon as my son took it out it was so easily broken. Hopefully a little glue will fix it.

Direct Recommendation

Pick the most suitable item from the following list and recommend to user_250 : \n 4915 , 1823 , 3112 , 3821 , 3773 , 520 , 7384 , 7469 , 9318 , 3876 , 1143 , 789 , 595 , 3824 , 3587 , 10396 , 2766 , 7498 , 2490 , 3232 , 9711 , 2975 , 1427 , 9923 , 3097 , 3594 , 6469 , 9460 , 6956 , 9154

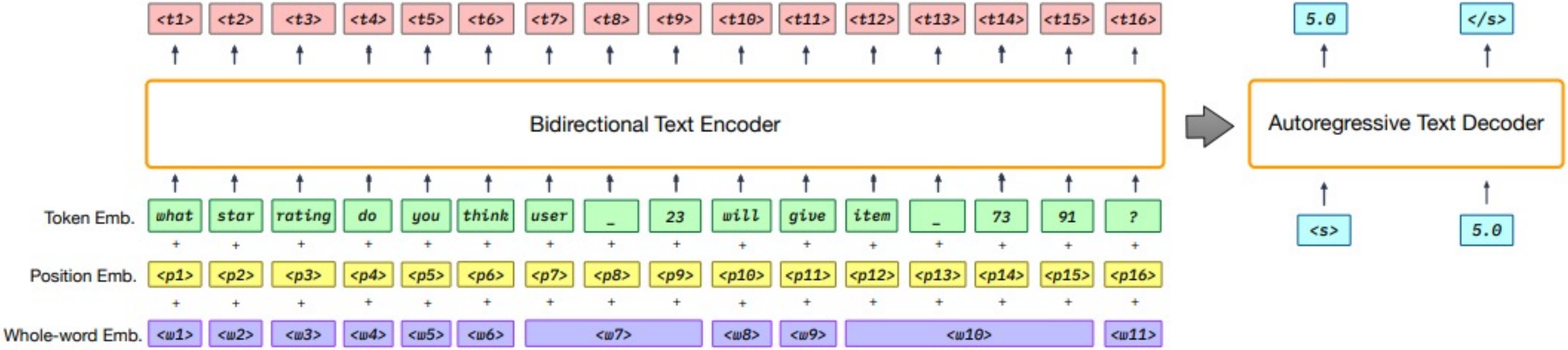
Multi-task Pretraining with Personalized Prompt Collection

Geng Shijie et al. "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)“ RecSys 2022.

Recommendation as NLP

❑ P5 Architecture:

- ➢ Autoregressive decoding
- ➢ Users and items are represented with ID information



<t1> <t2> <t3> <t4> <t5> <t6> <t7> <t8> <t9> <t10> <t11> <t12> <t13> <t14> <t15> <t16>
5.0
</s>
Bidirectional Text Encoder
Autoregressive Text Decoder
Token Emb.
what star rating do you think user _ 23 will give item _ 73 91 ?
<s>
5.0
Position Emb.
<p1> <p2> <p3> <p4> <p5> <p6> <p7> <p8> <p9> <p10> <p11> <p12> <p13> <p14> <p15> <p16>
Whole-word Emb.
<w1> <w2> <w3> <w4> <w5> <w6> <w7> <w8> <w9> <w10> <w11>

Geng Shijie et al. "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)“ RecSys 2022.

Recommendation as NLP

- **M6-Rec: represent users/item with plain texts and converting the tasks to either language understanding or generation**
  - **Understanding（scoring） task：CTR， CVR prediction**
  - **Generation task：personalized product design，explanation generation…**

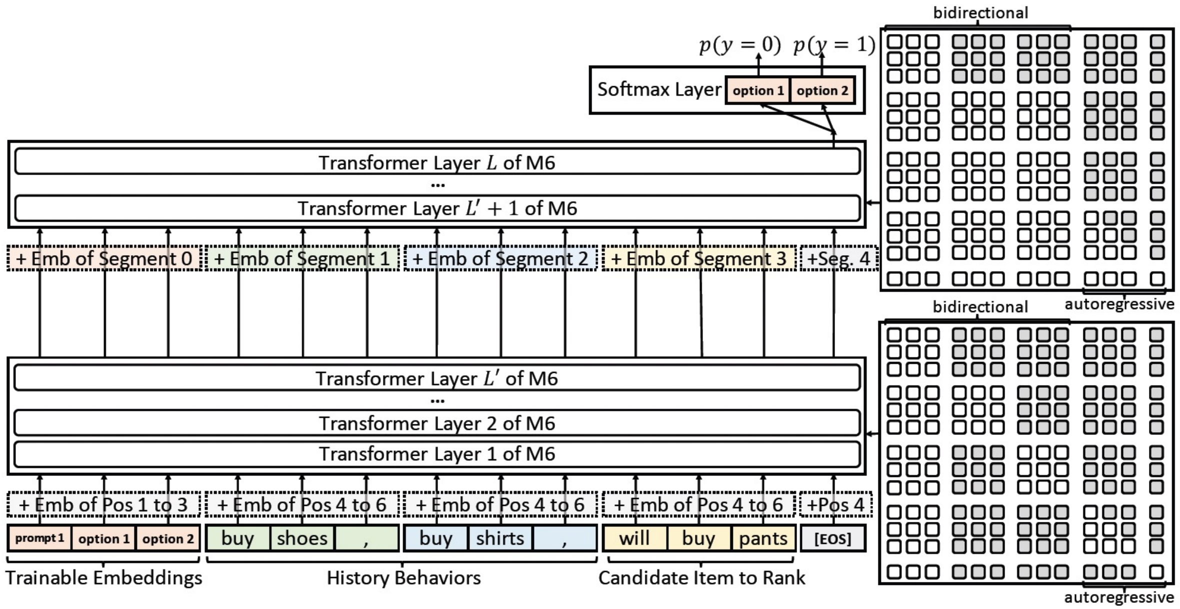
User description

[BOS′] December. Beijing, China. Cold weather. A male user in early twenties, searched "winter stuff" 23 minutes ago, clicked a product of category "jacket" named "men's lightweight warm winter hooded jacket" 19 minutes ago, clicked a product of category "sweatshirt" named "men's plus size sweatshirt stretchy pullover hoodies" 13 minutes ago, clicked … [EOS′]

[BOS] The user is now recommended a product of category "boots" named "waterproof hiking shoes mens outdoor". The product has a high population-level CTR in the past 14 days, among the top 5%. The user clicked the category 4 times in the last 2 years. [EOS]

Item description

M6 (~300M parameters)



Cui Zeyu et al. "M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems" arXiv 2022.

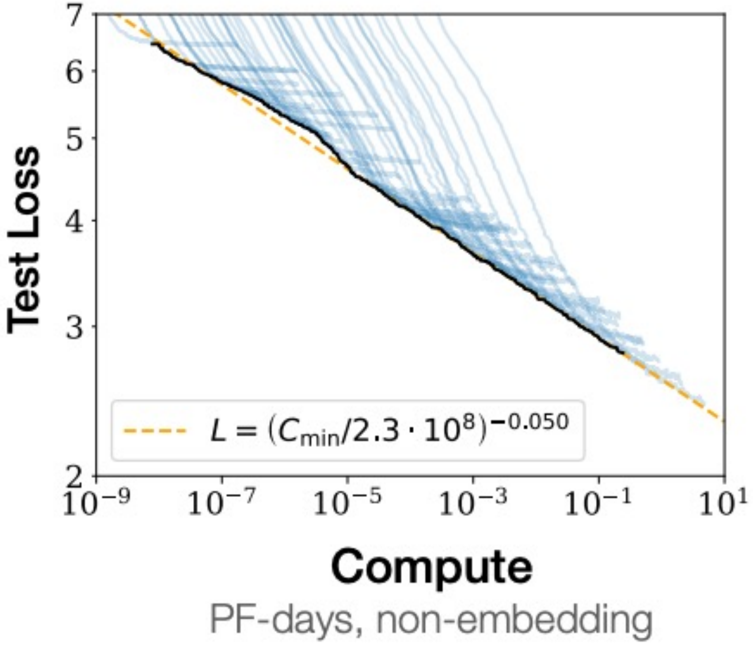
Outline

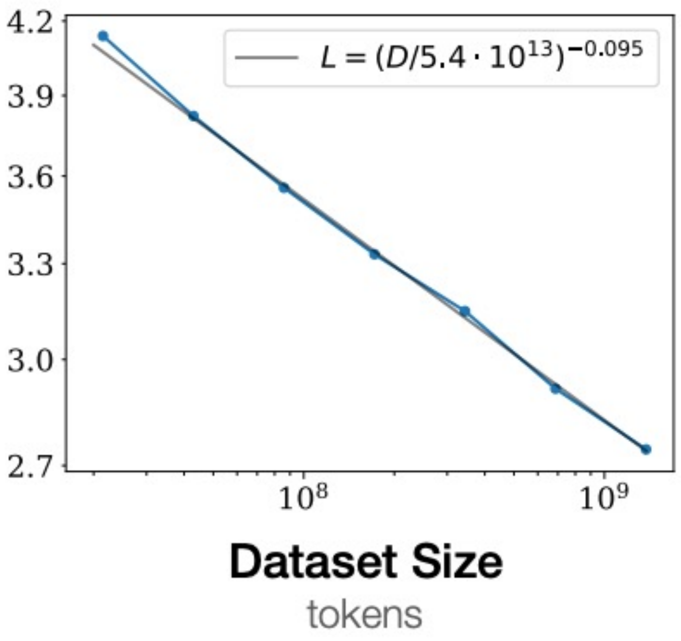
- Introduction
- Background: LM & LM4Rec
- **Development of LLMs**
- Progress of LLM4Rec
- Open Problems
- Future Direction & Conclusions

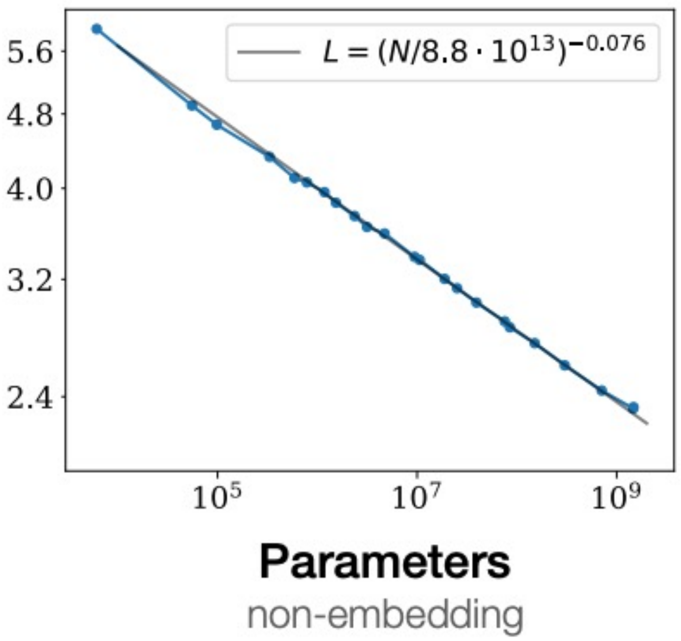
Developments of LLMs

❑ Scaling Laws

- ❑ The greater the amount of the data and the model parameters, the better the performance of the model
- ❑ Performance can be predicted



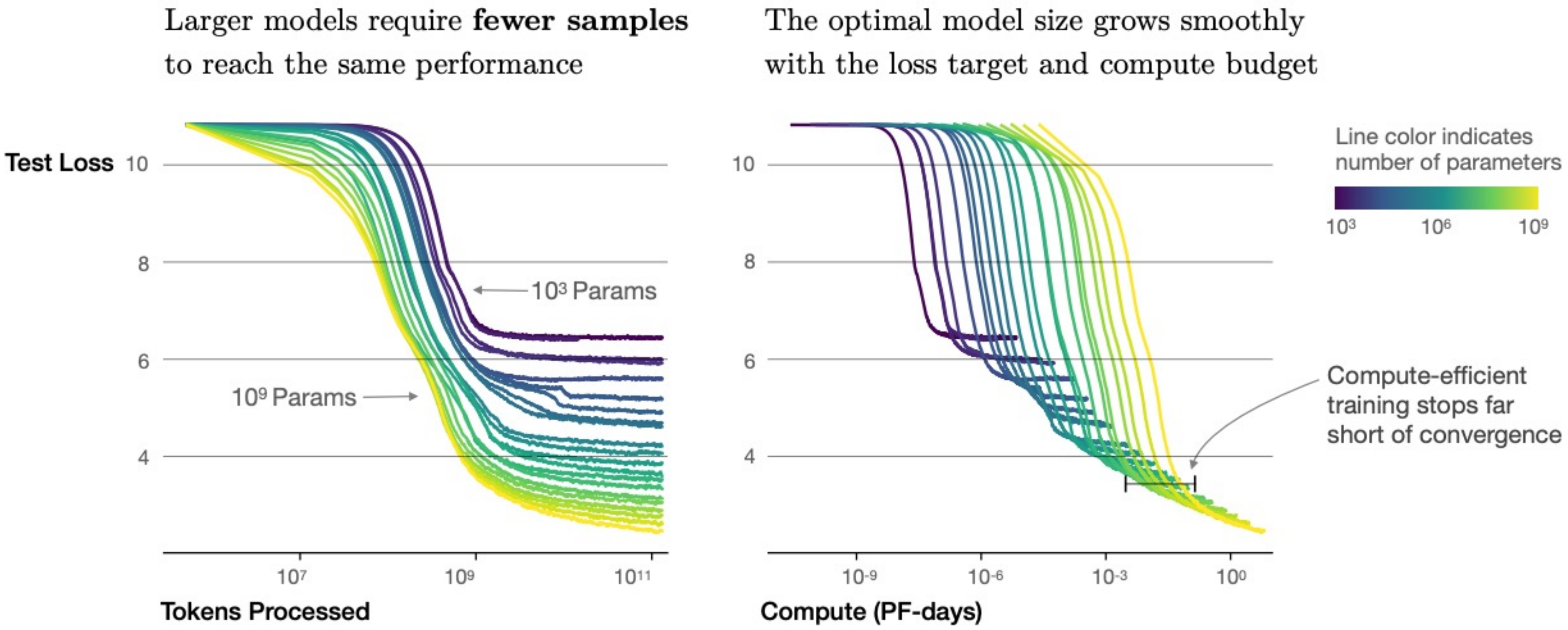




Jared Kaplan et al. Scaling Laws for Neural Language Models arxiv

Developments of LLMs

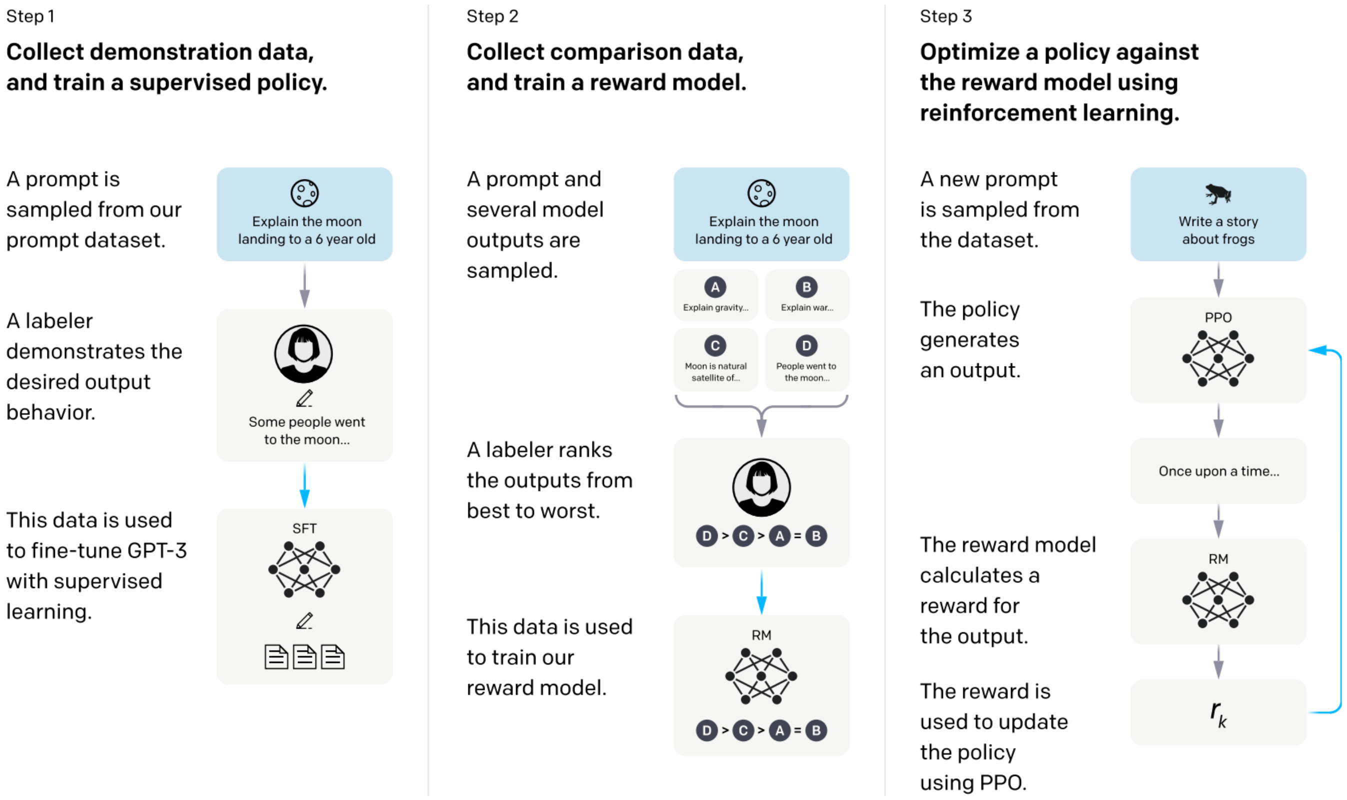
❑ Scaling Laws



Jared Kaplan et al. Scaling Laws for Neural Language Models arxiv

Developments of LLMs

❑ Align with human



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Explain the moon landing to a 6 year old
A Explain gravity...
B Explain war...
C Moon is natural satellite of...
D People went to the moon...
D > C > A = B
RM
D > C > A = B

Step 3

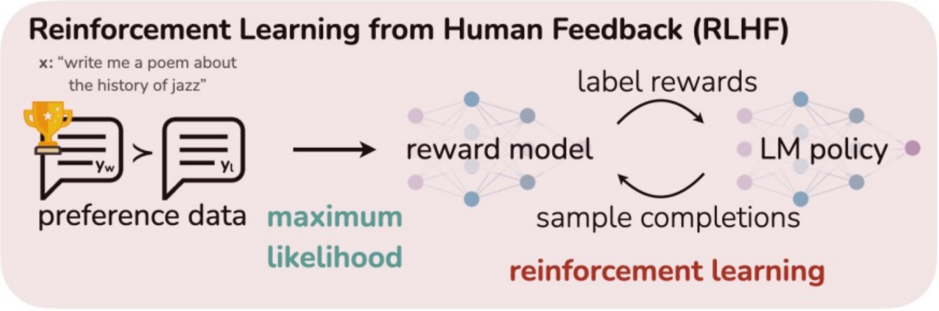
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

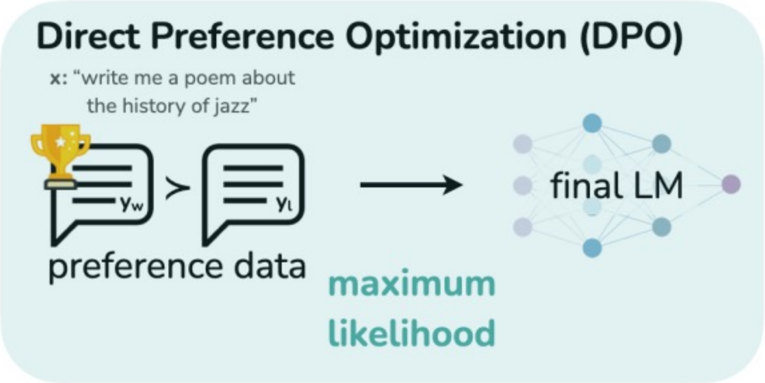
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Reinforcement Learning from Human Feedback (RLHF)

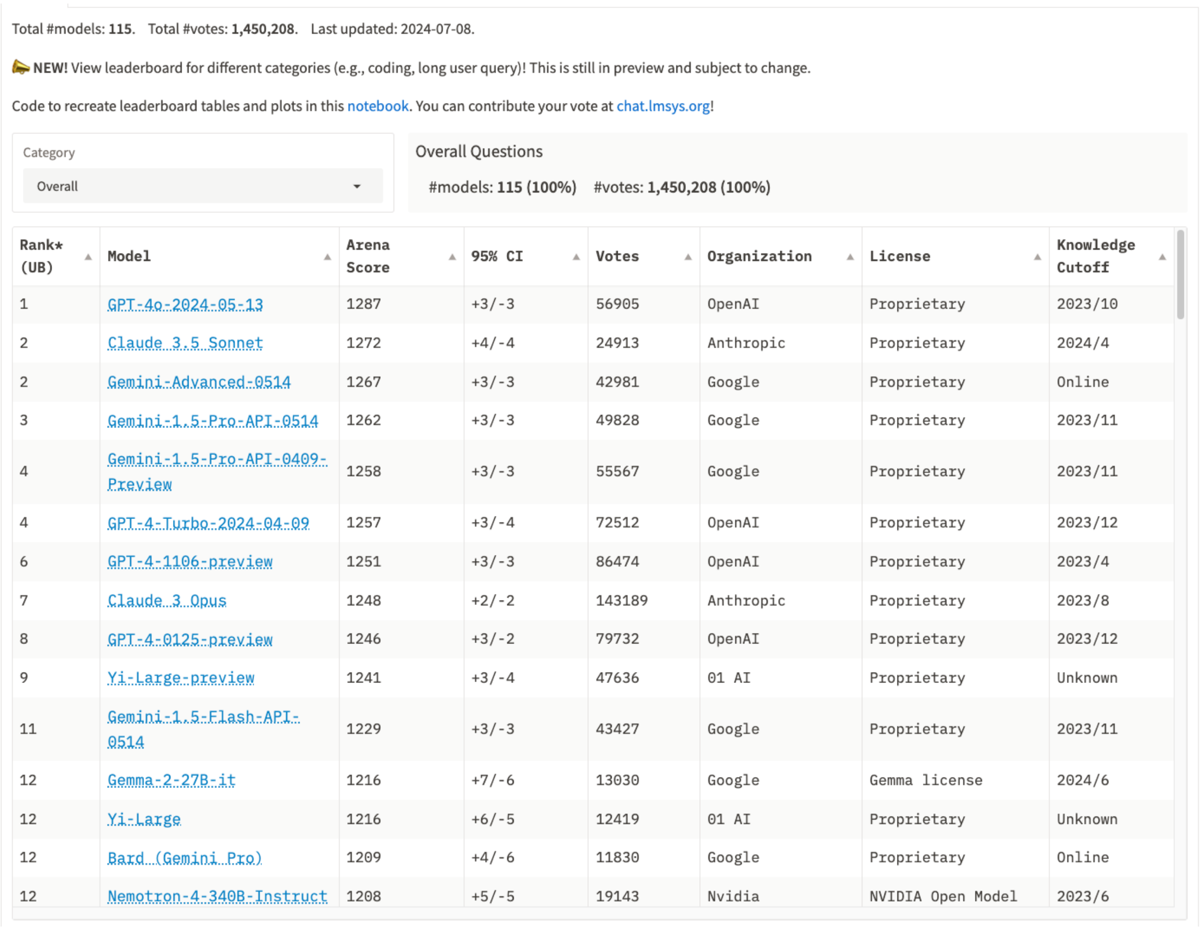


Direct Preference Optimization (DPO)

Long Ouyang et al., 2022 Training language models to follow instructions with human feedback NeurIPS 2022

Rafael Rafailov et al., 2023 Direct Preference Optimization: Your Language Model is Secretly a Reward Model NeurIPS 2023

Developments of LLMs



Total #models: **115**. Total #votes: **1,450,208**. Last updated: 2024-07-08.

📣 **NEW!** View leaderboard for different categories (e.g., coding, long user query)! This is still in preview and subject to change.

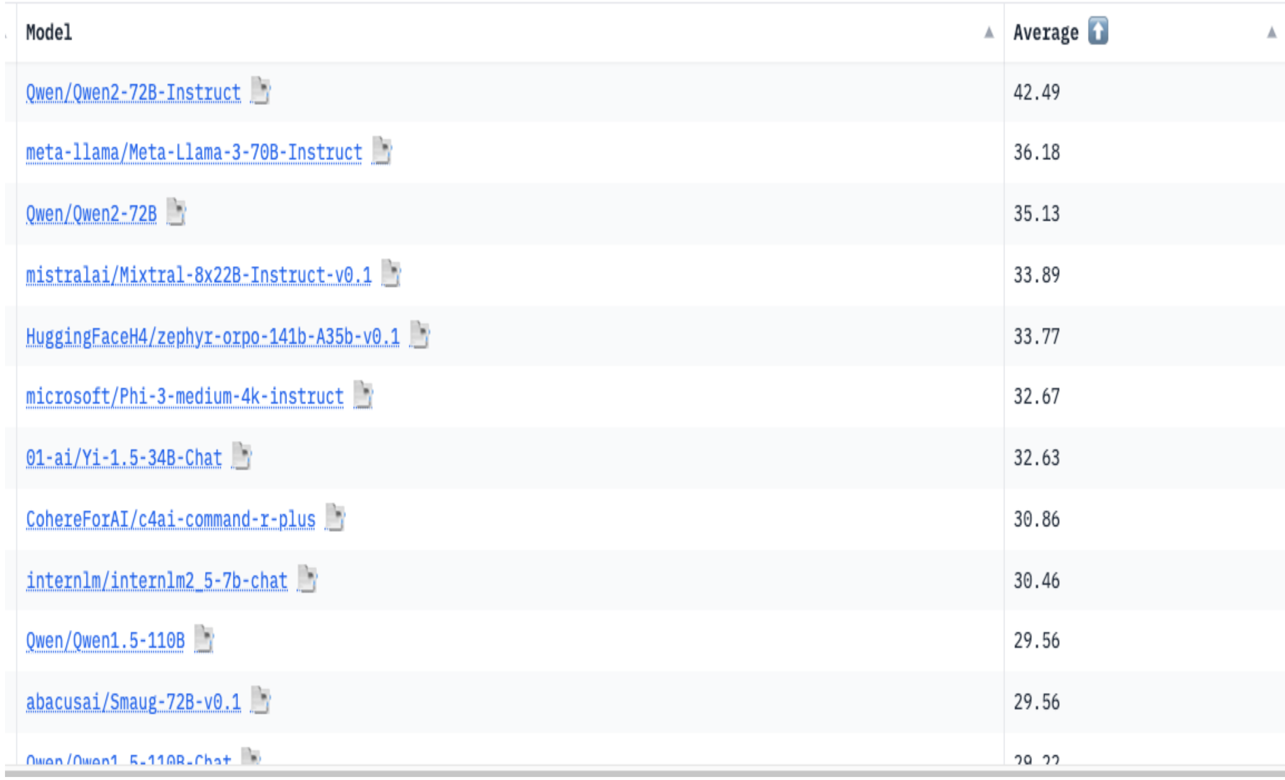
Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at chat.lmsys.org!

Category: Overall

Overall Questions

#models: **115 (100%)** #votes: **1,450,208 (100%)**

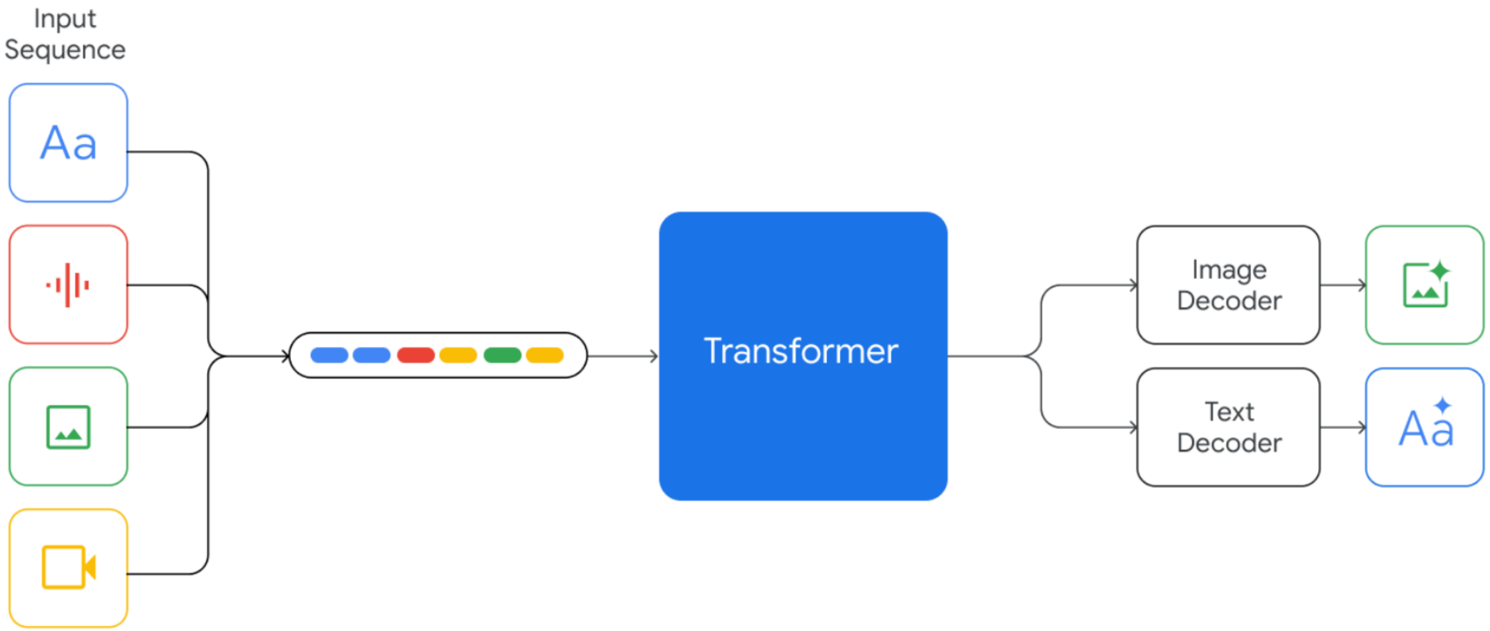
Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1287	+3/-3	56905	OpenAI	Proprietary	2023/10
2	Claude 3.5 Sonnet	1272	+4/-4	24913	Anthropic	Proprietary	2024/4
2	Gemini-Advanced-0514	1267	+3/-3	42981	Google	Proprietary	Online
3	Gemini-1.5-Pro-API-0514	1262	+3/-3	49828	Google	Proprietary	2023/11
4	Gemini-1.5-Pro-API-0409-Preview	1258	+3/-3	55567	Google	Proprietary	2023/11
4	GPT-4-Turbo-2024-04-09	1257	+3/-4	72512	OpenAI	Proprietary	2023/12
6	GPT-4-1106-preview	1251	+3/-3	86474	OpenAI	Proprietary	2023/4
7	Claude 3 Opus	1248	+2/-2	143189	Anthropic	Proprietary	2023/8
8	GPT-4-0125-preview	1246	+3/-2	79732	OpenAI	Proprietary	2023/12
9	Yi-Large-preview	1241	+3/-4	47636	01 AI	Proprietary	Unknown
11	Gemini-1.5-Flash-API-0514	1229	+3/-3	43427	Google	Proprietary	2023/11
12	Gemma-2-27B-it	1216	+7/-6	13030	Google	Gemma license	2024/6
12	Yi-Large	1216	+6/-5	12419	01 AI	Proprietary	Unknown
12	Bard (Gemini Pro)	1209	+4/-6	11830	Google	Proprietary	Online
12	Nemotron-4-340B-Instruct	1208	+5/-5	19143	Nvidia	NVIDIA Open Model	2023/6

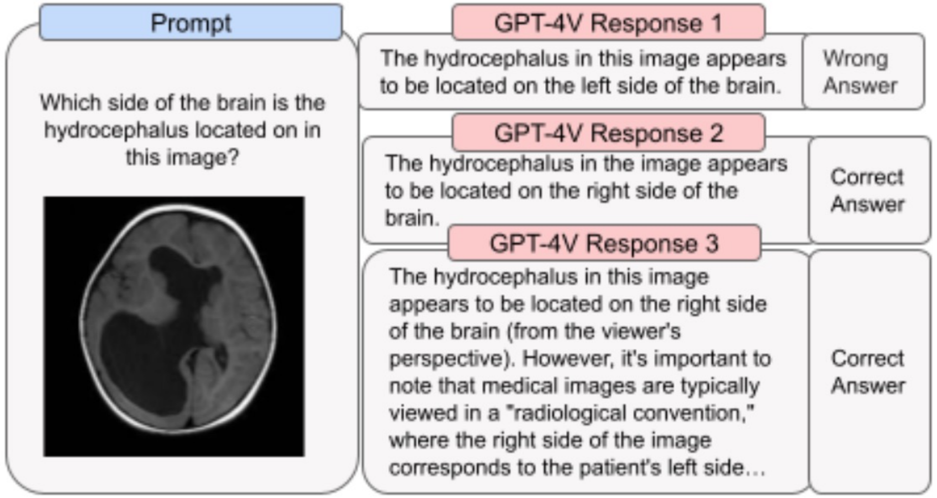


Model	Average ⬆️
Qwen/Qwen2-72B-Instruct	42.49
meta-llama/Meta-Llama-3-70B-Instruct	36.18
Qwen/Qwen2-72B	35.13
mistralai/Mixtral-8x22B-Instruct-v0.1	33.89
HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1	33.77
microsoft/Phi-3-medium-4k-instruct	32.67
01-ai/Yi-1.5-34B-Chat	32.63
CohereForAI/c4ai-command-r-plus	30.86
internlm/internlm2_5-7b-chat	30.46
Qwen/Qwen1.5-110B	29.56
abacusai/Smaug-72B-v0.1	29.56
Qwen/Qwen1.5-110B-Chat	29.22

❑ **More and more LLMs have shown powerful capabilities**

Developments of LLMs





Which side of the brain is the hydrocephalus located on in this image?

The hydrocephalus in this image appears to be located on the left side of the brain. Wrong Answer

The hydrocephalus in the image appears to be located on the right side of the brain. Correct Answer

The hydrocephalus in this image appears to be located on the right side of the brain (from the viewer's perspective). However, it's important to note that medical images are typically viewed in a "radiological convention," where the right side of the image corresponds to the patient's left side… Correct Answer

- ❑ **Multi-model to Multi-model unified model is now developing at a rapid pace.**
  - ❑ **E.g. GPT-4o Gemini Claude-3.5**

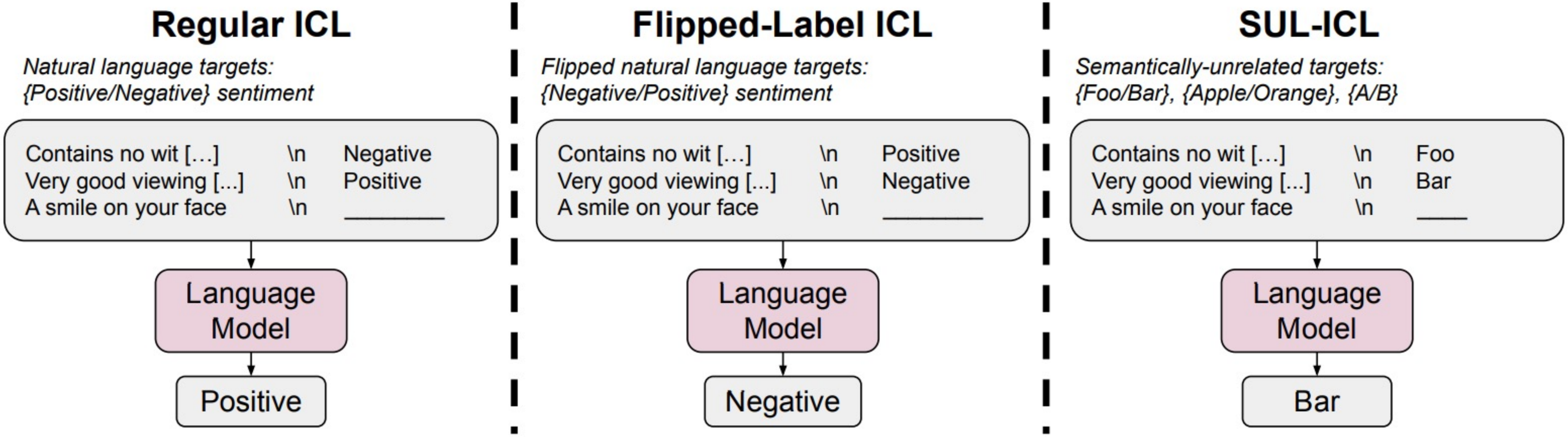
Augmented capabilities of LLMs

- ❑ Emergent abilities of LLM
  - ❑ Sufficient world knowledge
  - ❑ Chatting
  - ❑ Incontext Learning & Instruction Following
  - ❑ Reasoning & Planning
  - ❑ Tool using
  - ❑ LLM as an Agent
  - ❑ …

Augmented capabilities of LLMs

❑ In-context Learning

- ❑ Following their example to overide the semantic prior



Regular ICL

Natural language targets:
{Positive/Negative} sentiment

Contains no wit […] \n Negative
Very good viewing [...] \n Positive
A smile on your face \n ________

Language Model

Positive

Flipped-Label ICL

Flipped natural language targets:
{Negative/Positive} sentiment

Contains no wit […] \n Positive
Very good viewing [...] \n Negative
A smile on your face \n ________

Language Model

Negative

SUL-ICL

Semantically-unrelated targets:
{Foo/Bar}, {Apple/Orange}, {A/B}

Contains no wit […] \n Foo
Very good viewing [...] \n Bar
A smile on your face \n ____

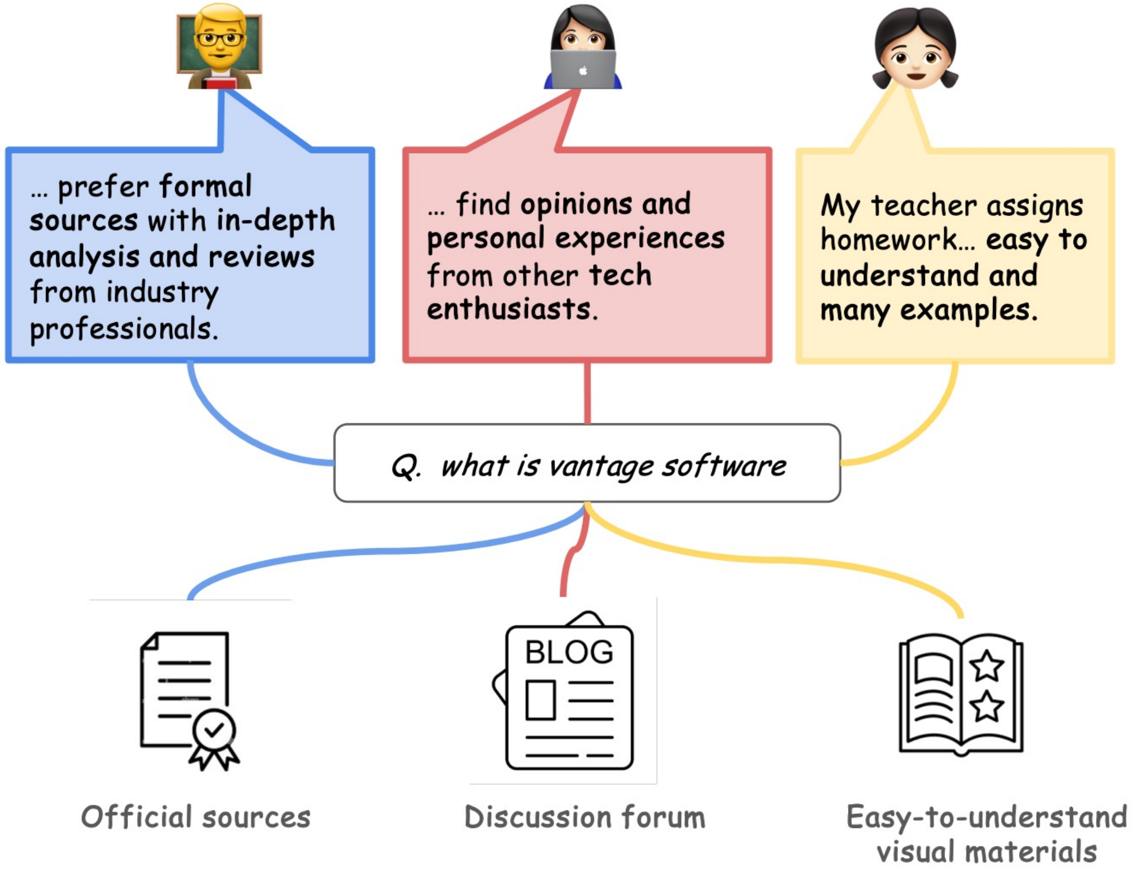
Language Model

Bar

Jerrt Wei et al. Language Models Do In-context Learning Differently

Augmented capabilities of LLMs

❑ Instruction following

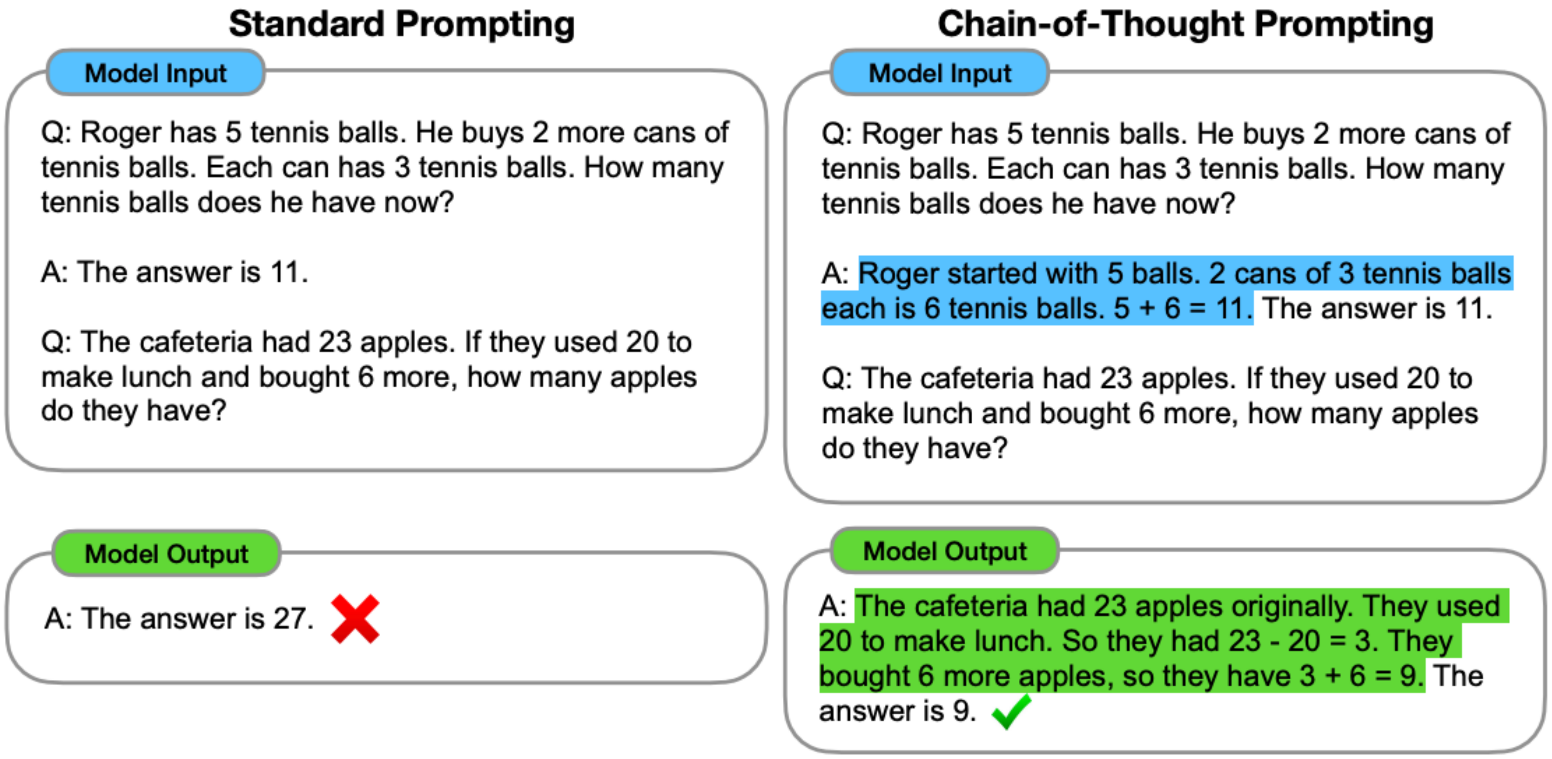

… prefer formal sources with in-depth analysis and reviews from industry professionals.
… find opinions and personal experiences from other tech enthusiasts.
My teacher assigns homework… easy to understand and many examples.
Q. what is vantage software
BLOG
Official sources
Discussion forum
Easy-to-understand visual materials

Jerrt Wei et al. Language Models Do In-context Learning Differently

Augmented capabilities of LLMs

❑ Reasoning & Planning

- ❑ LLM can decompose the problem into simple sub-problems to improve their ability



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

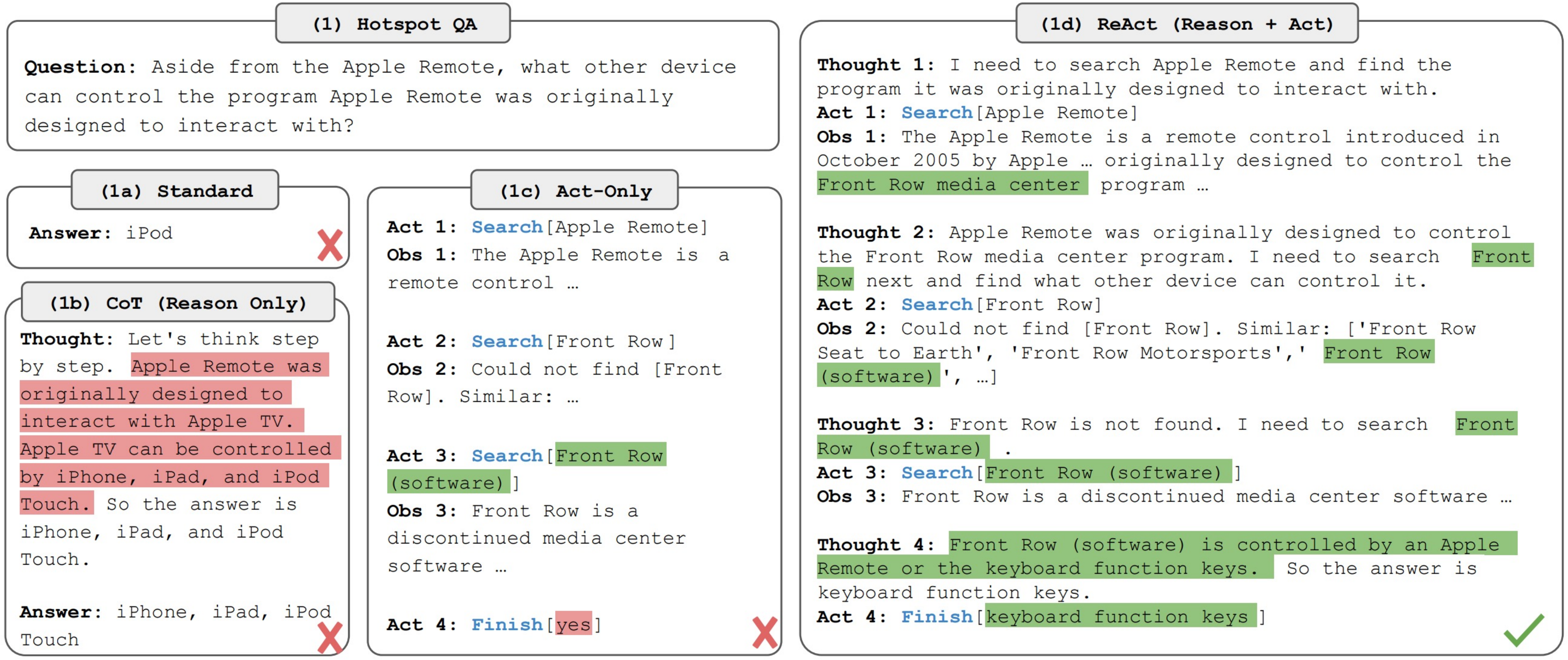
Jason Wei et al, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Augmented capabilities of LLMs

❑ Reasoning & Planning

- ❑ LLM can break down the targe task according to the environment and develop a

  .



(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod ✗

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch ✗

(1c) Act-Only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control …

Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: …

Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Act 4: Finish[yes] ✗

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

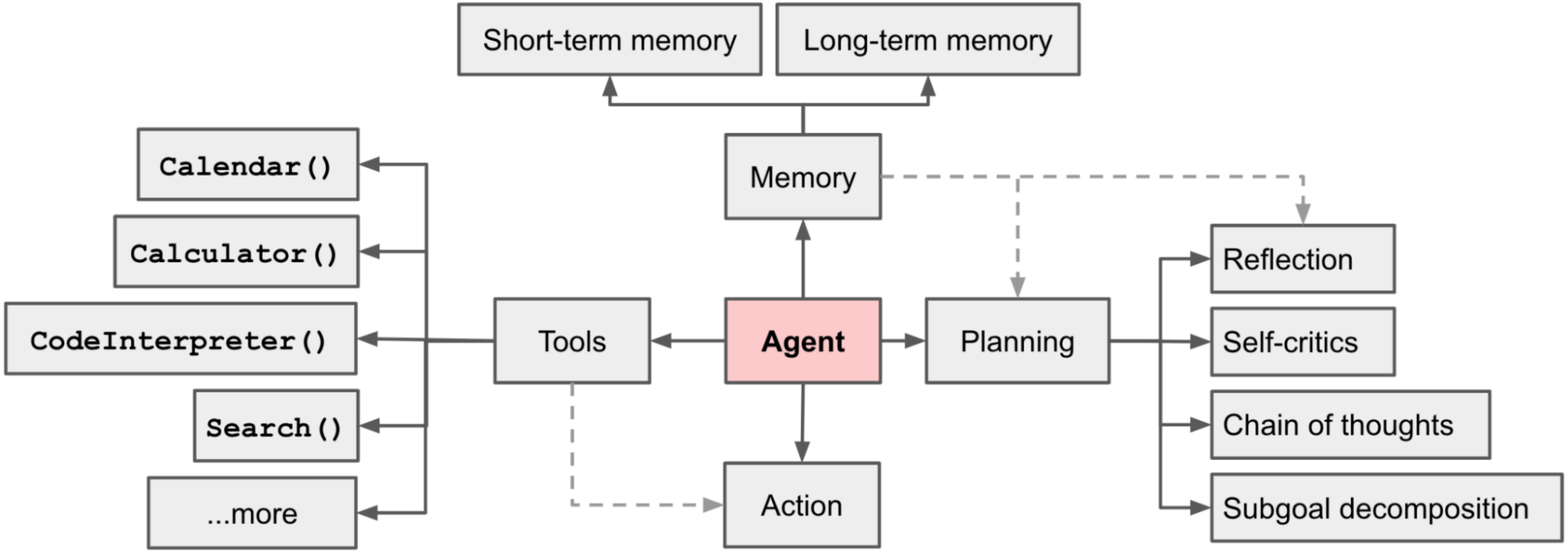
Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys] ✓

Yao et al， REAC T: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS

Augmented capabilities of LLMs

❑ LLM as an Agent

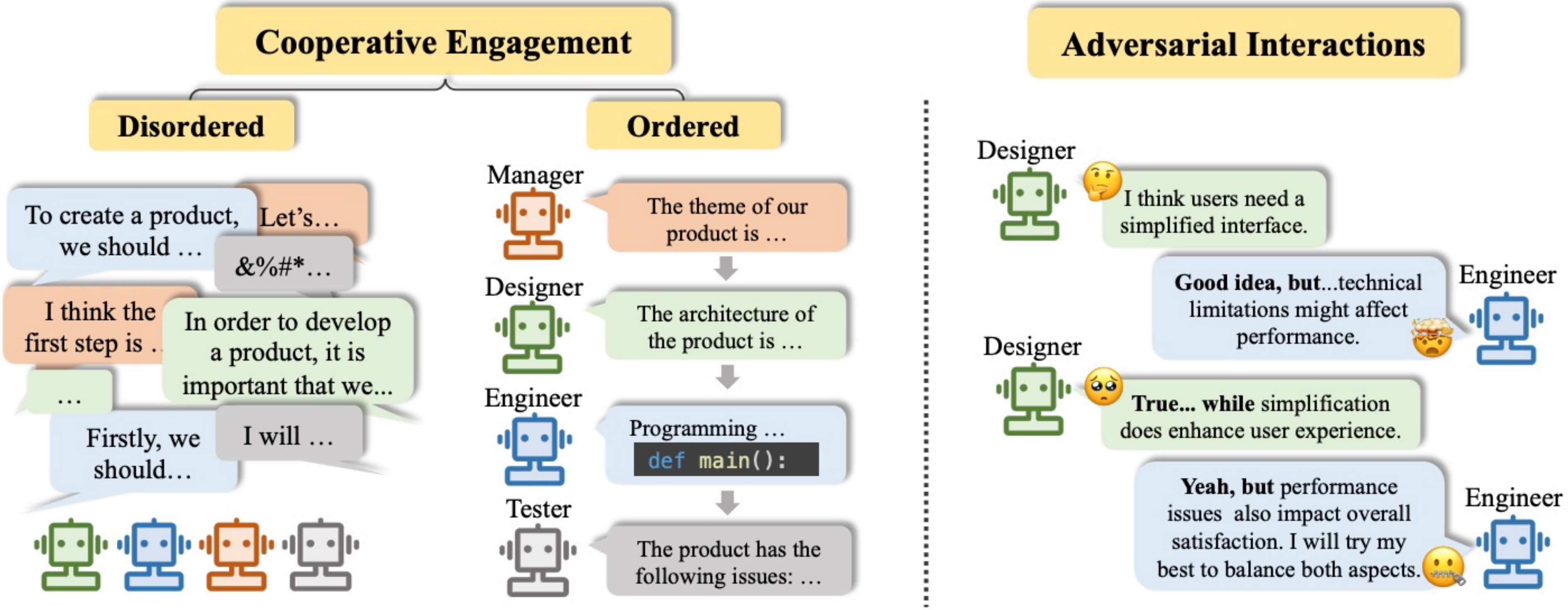


Images from https://lilianweng.github.io/posts/2023-06-23-agent/

Multi-Agent

❑ Multi-Agent

- ❑ Group intelligence surpasses individual intelligence
- ❑ Cooperative for complementary / Adversarial for advancement



Cooperative Engagement
Disordered
Ordered
To create a product, we should …
Let's…
&%#*…
I think the first step is …
In order to develop a product, it is important that we...
…
Firstly, we should…
I will …
Manager
The theme of our product is …
Designer
The architecture of the product is …
Engineer
Programming …
def main():
Tester
The product has the following issues: …
Adversarial Interactions
Designer
I think users need a simplified interface.
Good idea, but...technical limitations might affect performance.
Engineer
Designer
True... while simplification does enhance user experience.
Yeah, but performance issues also impact overall satisfaction. I will try my best to balance both aspects.
Engineer

Zhiheng Xi et al， The Rise and Potential of Large Language Model Based Agents: A Survey

LLMs for Recommendation

❑ How recommender systems benefit from LLMs

- **Representation:**

Textual feature,
item representation,
knowledge representation

- **Interaction:**

Acquire user information needs via dialog (**chat**)

- **Generalization:**

cross-domain, knowledge
compositional-generalization

- **Generation:**

Personalized content generation,
explanation generation

- **Learning paradigm:** Pretrain-finetune, Instruction-tuning, Preference-alignment

- **Model architecture:** Transformer、Self-attention,

LLMs for Recommendation

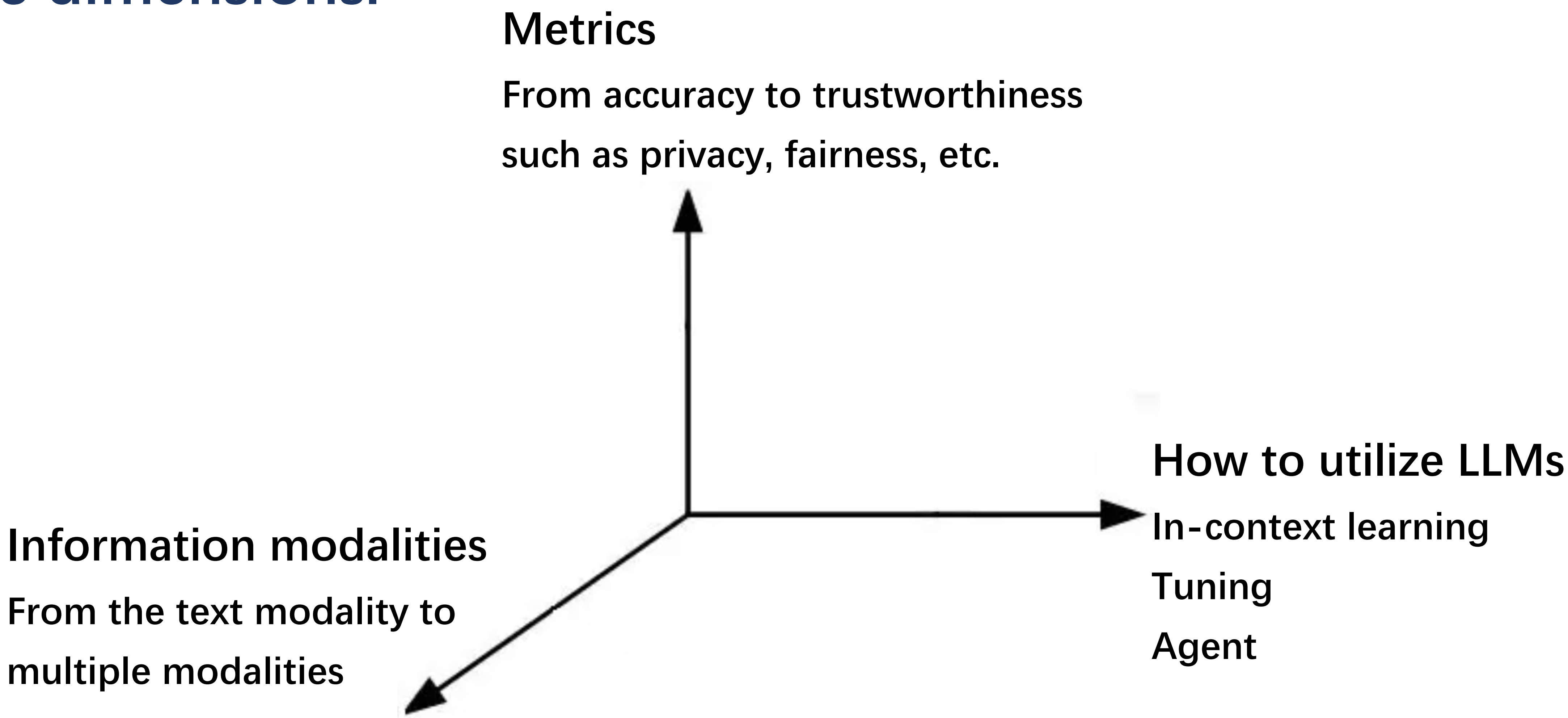
- ❑ Key Challenge
  - ❑ Mismatch between pretraining objective and recommendation
  - ❑ Tend to rely on semantics, and another important aspect of recommendation tasks is collaborative information.

Outline

- Introduction
- Background: LM & LM4Rec
- **The progress of LLM4Rec**
  - **Development of LLMs**
  - **LLMs for Recommendation**
    - **ICL**
    - **Tuning**
    - **Agent**
- Open Problems
- Conclusions

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

How to utilize LLMs

In-context learning

Tuning

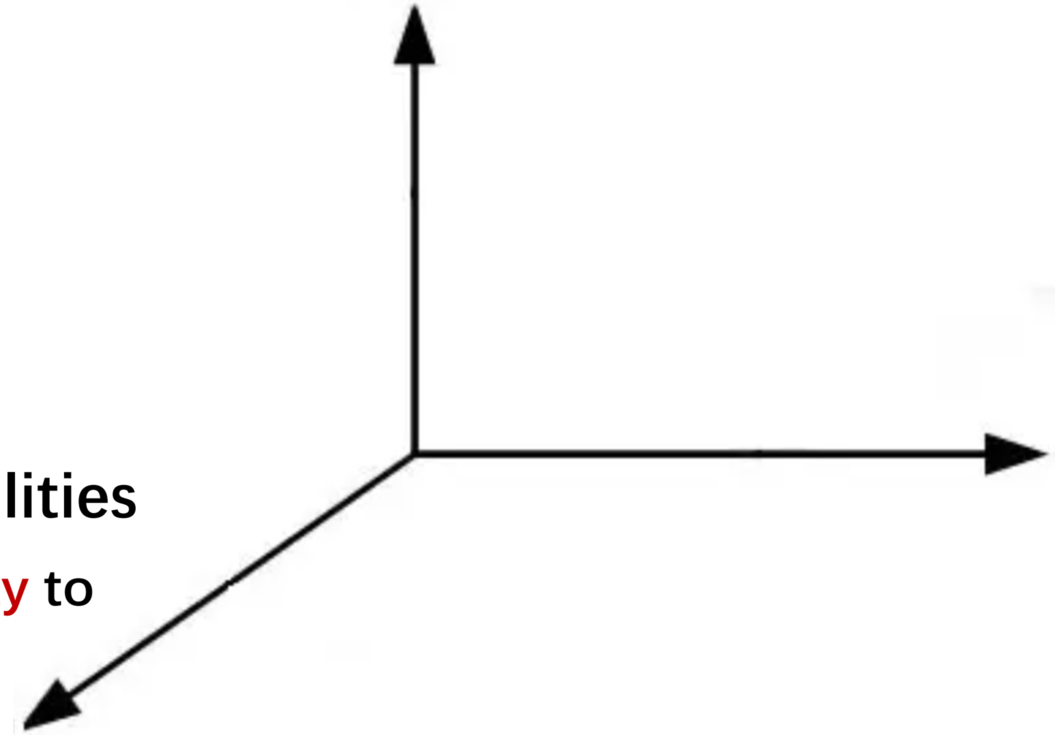
Agent

Information modalities

From the text modality to multiple modalities

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

How to utilize LLMs

- In-context learning
- Tuning
- Agent

Information modalities

From the text modality to multiple modalities

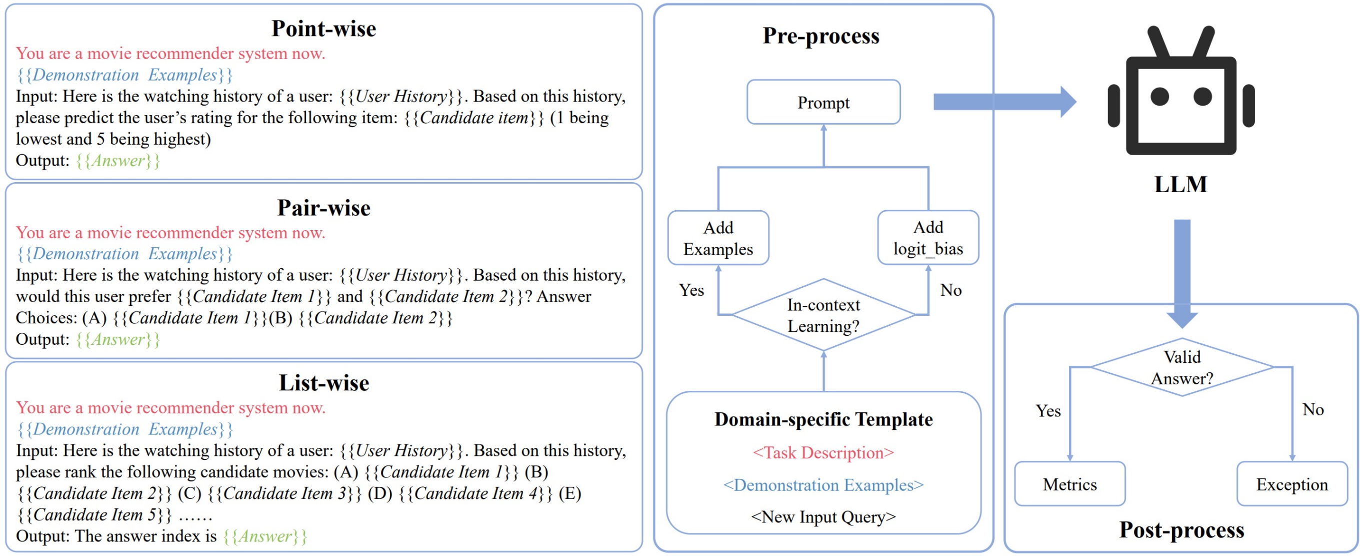
ICL

❑ In-context learning

- LLMs has rich world knowledge, wonderful abilities like reasoning, instruction following, in-context learning.
- The LLMs itself could be leveraged for recommendation by in context learning.
- Existing works on in-context learning:
  - Ask LLM for recommendation
  - Serving as knowledge augmentation for traditional recsys
  - Optimize the prompt used for recommendation
  - Directly used for conversational recommender system

ICL

- **In-context learning:** directly ask LLMs for recommendation
  - Prompt construction



Point-wise

You are a movie recommender system now.
{{*Demonstration Examples*}}
Input: Here is the watching history of a user: {{*User History*}}. Based on this history, please predict the user's rating for the following item: {{*Candidate item*}} (1 being lowest and 5 being highest)
Output: {{*Answer*}}

Pair-wise

You are a movie recommender system now.
{{*Demonstration Examples*}}
Input: Here is the watching history of a user: {{*User History*}}. Based on this history, would this user prefer {{*Candidate Item 1*}} and {{*Candidate Item 2*}}? Answer Choices: (A) {{*Candidate Item 1*}}(B) {{*Candidate Item 2*}}
Output: {{*Answer*}}

List-wise

You are a movie recommender system now.
{{*Demonstration Examples*}}
Input: Here is the watching history of a user: {{*User History*}}. Based on this history, please rank the following candidate movies: (A) {{*Candidate Item 1*}} (B) {{*Candidate Item 2*}} (C) {{*Candidate Item 3*}} (D) {{*Candidate Item 4*}} (E) {{*Candidate Item 5*}} ……
Output: The answer index is {{*Answer*}}

Figure 1: The overall evaluation framework of LLMs for recommendation. The left part demonstrates examples of how prompts are constructed to elicit each of the three ranking capabilities. The right part outlines the process of employing LLMs to perform different ranking tasks and conduct evaluations.

Three different ways of measuring ranking abilities:

$$\hat{y}'_i = LLM_{\text{point}}\left(I, \mathcal{D}, f(\mathbf{h}', \mathbf{c}' \mid u)\right)$$

$$\hat{y}'_{i_m > i_n} = LLM_{\text{pair}}\left(I, \mathcal{D}, f(\mathbf{h}', \mathbf{c}' \mid u)\right)$$

$$\hat{y}'_{i_1}, \hat{y}'_{i_2}, \cdots, \hat{y}'_{i_k} = LLM_{\text{list}}\left(I, \mathcal{D}, f(\mathbf{h}', \mathbf{c}' \mid u)\right)$$

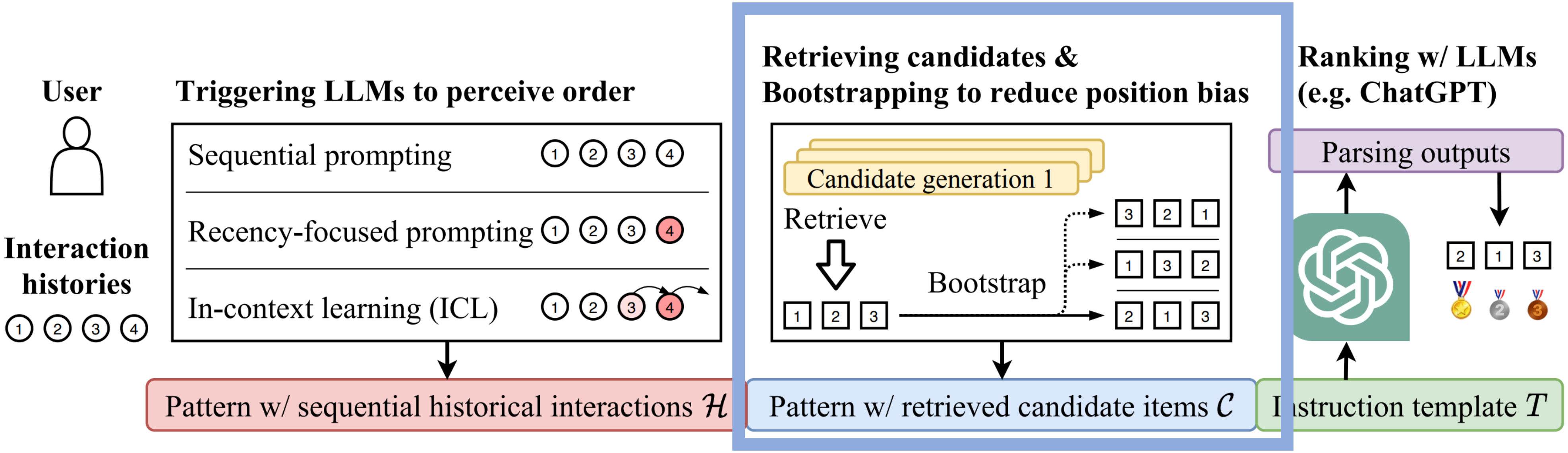
Uncovering ChatGPT's Capabilities in Recommender Systems, RecSys, 2023

ICL: LLMRank

❑ In-context learning: re-ranking given candidated items

❑ Task formulation:

- ***Using historical interaction to rank items retrieved by exsiting recsys.***
- ***Input:*** *language instructions created with historical interactions and candidate items*
- ***Output:*** *ranking of the candidate items*


User
Interaction histories
① ② ③ ④
Triggering LLMs to perceive order
Sequential prompting ① ② ③ ④
Recency-focused prompting ① ② ③ ④
In-context learning (ICL) ① ② ③ ④
Pattern w/ sequential historical interactions $\mathcal{H}$
Retrieving candidates & Bootstrapping to reduce position bias
Candidate generation 1
Retrieve
1 2 3
Bootstrap
3 2 1
1 3 2
2 1 3
Pattern w/ retrieved candidate items $\mathcal{C}$
Ranking w/ LLMs (e.g. ChatGPT)
Parsing outputs
2 1 3
1 2 3
Instruction template T

Yupeng Hou et al. Large Language Models are Zero-Shot Rankers for Recommender Systems ECIR 2024

ICL: LLMRank

❑ In-context learning: ranking given candidated items

❑ Tree types of prompts:

- **Sequential prompting: describing History using language**

  "I’ve watched the following movies in the past in order: ’0. Multiplicity’, ’1. Jurassic Park’, ..."

- **Recency-focused prompting**： **emphasize most recent interactions**

  “I’ve watched the following movies in the past in order: ’0. Multiplicity’, ’1. Jurassic Park’, ... Note that my most recently watched movie is Dead Presidents. . . .”

- **In-context learning (ICL)**： **providing recommendation example**

  “ If I’ve watched the following movies in the past in order: ’0. Multiplicity’, ’1. Jurassic Park’, . . ., then you should recommend Dead Presidents to me and now that I’ve watched Dead Presidents, then . . .”

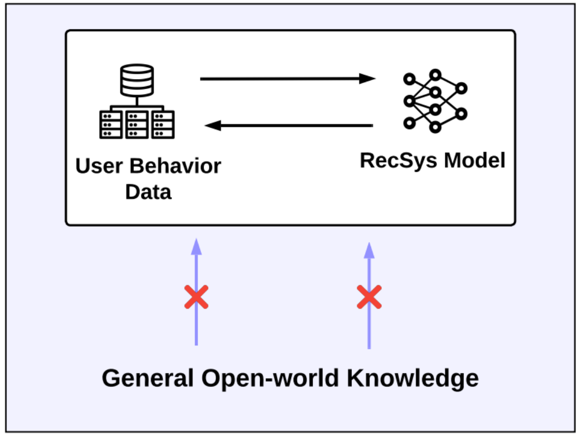
Yupeng Hou et al. Large Language Models are Zero-Shot Rankers for Recommender Systems ECIR 2024

ICL: KAR

❑ In-context learning: knowledge enhancement

❑ Traditional RecSys vs ICL-based RecSys

Traditional RecSys



User Behavior Data

RecSys Model

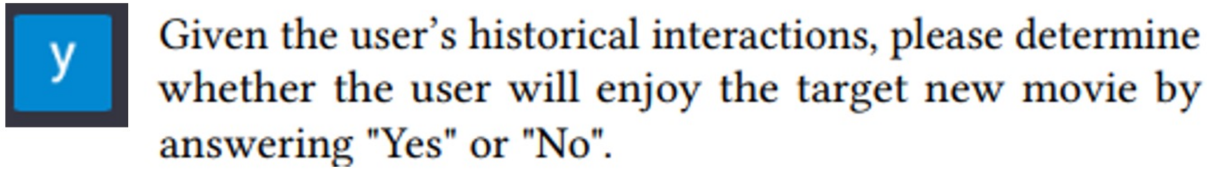
General Open-world Knowledge

Inference fast but being colsed system, generating recommendations relying on local dataset

Directly ask LLMs for recommendaiton

Instruction

LLM e.g. ChatGPT



Given the user's historical interactions, please determine whether the user will enjoy the target new movie by answering "Yes" or "No".

Could **leverage open-world knowledge**, but:

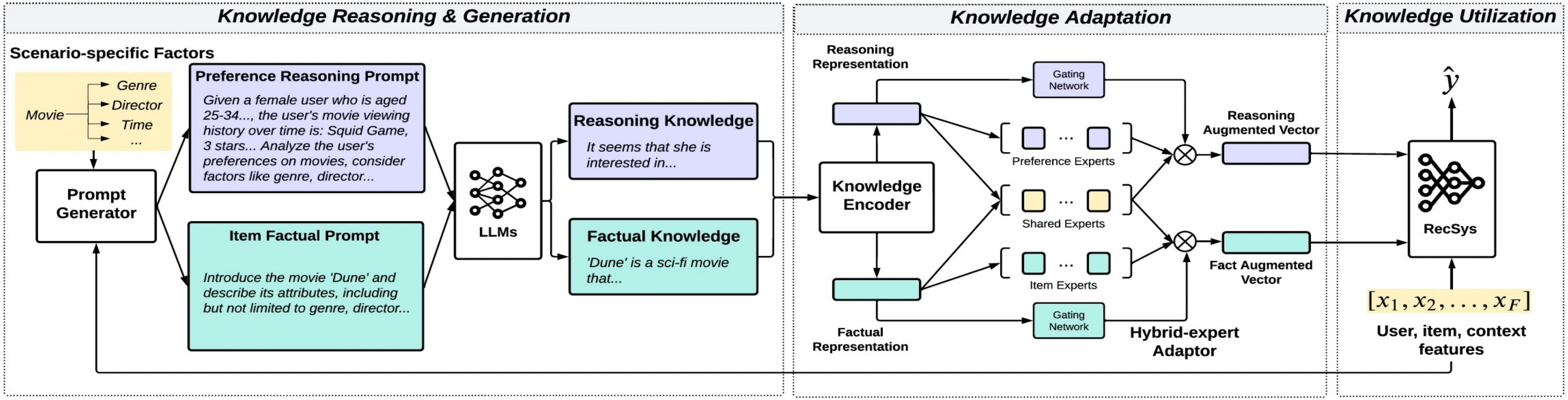
1) not trained on specific recommendation task
2) Inference slowly
3) hard to correctly answer compoitional questions

Extract and inject LLM's world knowledge into traditional recommender system

Xi Yunjia et al. "Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models“ arXiv 2023.

ICL: KAR

❑ In-context learning: knowledge enhancement



Knowledge Reasoning & Generation
Scenario-specific Factors
Movie
Genre
Director
Time
...
Prompt Generator
Preference Reasoning Prompt
Given a female user who is aged 25-34..., the user's movie viewing history over time is: Squid Game, 3 stars... Analyze the user's preferences on movies, consider factors like genre, director...
Item Factual Prompt
Introduce the movie 'Dune' and describe its attributes, including but not limited to genre, director...
LLMs
Reasoning Knowledge
It seems that she is interested in...
Factual Knowledge
'Dune' is a sci-fi movie that...
Knowledge Adaptation
Reasoning Representation
Knowledge Encoder
Factual Representation
Gating Network
Preference Experts
Shared Experts
Item Experts
Gating Network
Reasoning Augmented Vector
Fact Augmented Vector
Hybrid-expert Adaptor
Knowledge Utilization
$\hat{y}$
RecSys
$[x_1, x_2, ..., x_F]$
User, item, context features

Obtain knowledge beyond local rec dataset：

1） Generate reasoning knowledge on user preference (factors affect preference)

2） Generate factual knowledge about items

Knowledge Adaptation Stage

encode the textual knowledge and maping it into recommendation space

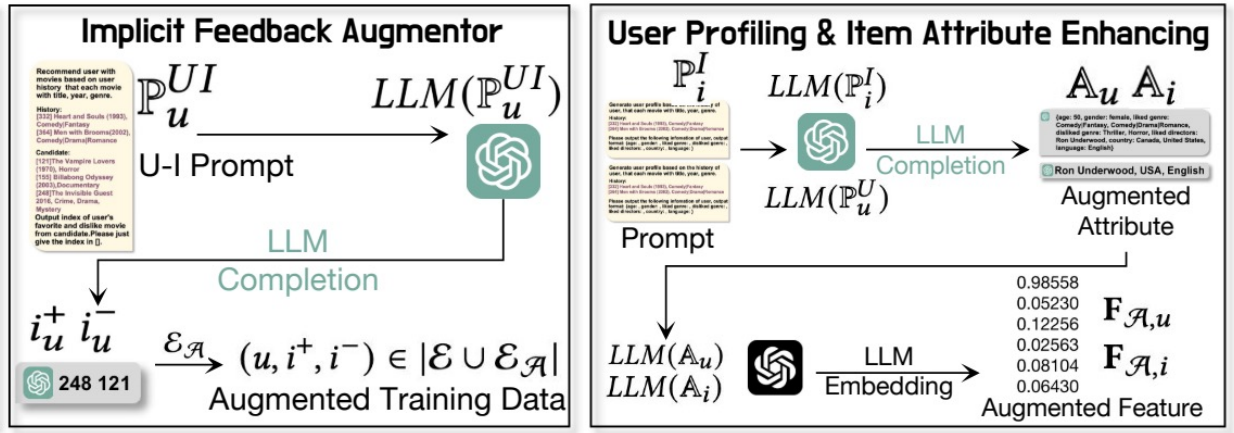
Knowledge Utilization

Use the knowledge obtained from LLMs as additional features

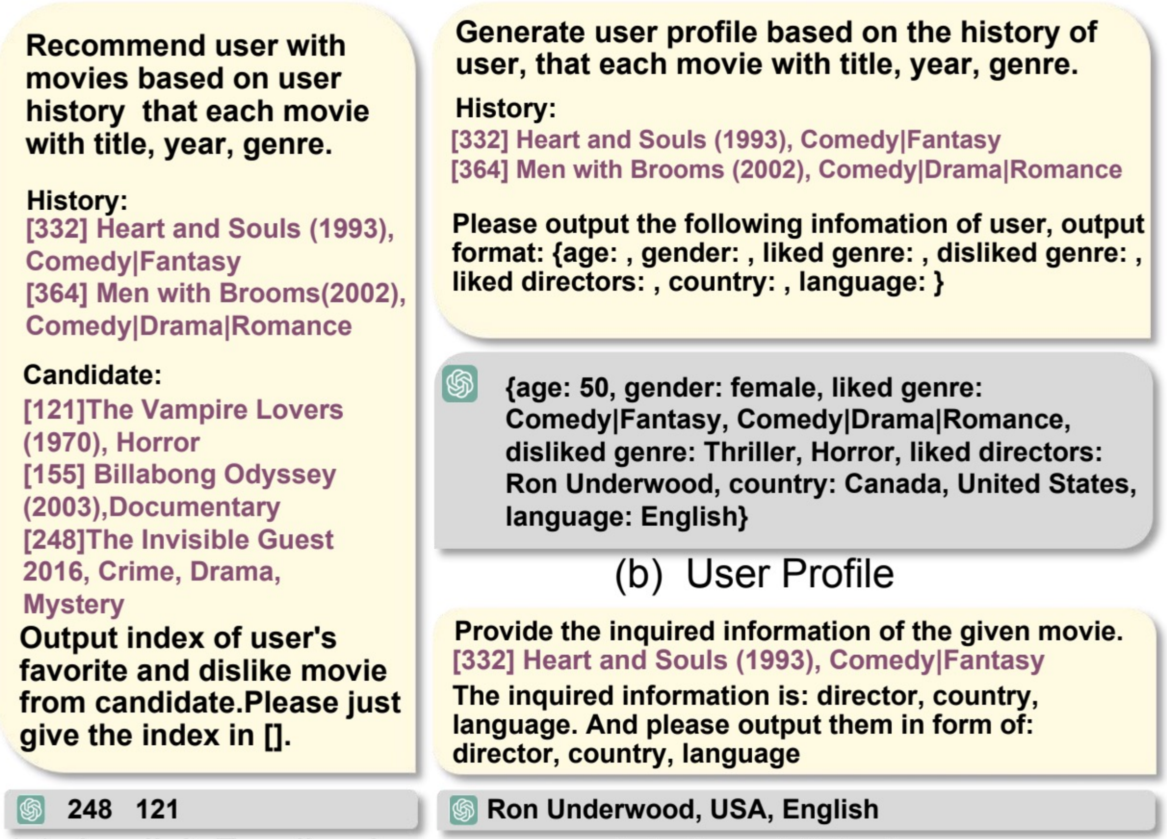
Xi Yunjia et al. "Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models“ arXiv 2023.

ICL: LLM4Rec

❑ ICL knowledge enhancement for Graph-based Recommendation



Implicit Feedback Augmentor
$\mathbb{P}_u^{UI}$
$LLM(\mathbb{P}_u^{UI})$
U-I Prompt
LLM
Completion
$i_u^+ \; i_u^-$
248 121
$\xrightarrow{\mathcal{E}_{\mathcal{A}}} (u, i^+, i^-) \in |\mathcal{E} \cup \mathcal{E}_{\mathcal{A}}|$
Augmented Training Data
User Profiling & Item Attribute Enhancing
$\mathbb{P}_i^I$
$LLM(\mathbb{P}_i^I)$
$LLM(\mathbb{P}_u^U)$
Prompt
LLM
Completion
$\mathbb{A}_u \; \mathbb{A}_i$
{age: 50, gender: female, liked genre: Comedy|Fantasy, Comedy|Drama|Romance, disliked genre: Thriller, Horror, liked directors: Ron Underwood, country: Canada, United States, language: English}
Ron Underwood, USA, English
Augmented Attribute
$LLM(\mathbb{A}_u)$
$LLM(\mathbb{A}_i)$
LLM
Embedding
0.98558
0.05230
0.12256
0.02563
0.08104
0.06430
$\mathbf{F}_{\mathcal{A},u}$
$\mathbf{F}_{\mathcal{A},i}$
Augmented Feature



Recommend user with movies based on user history that each movie with title, year, genre.

History:
[332] Heart and Souls (1993), Comedy|Fantasy
[364] Men with Brooms(2002), Comedy|Drama|Romance

Candidate:
[121]The Vampire Lovers (1970), Horror
[155] Billabong Odyssey (2003),Documentary
[248]The Invisible Guest 2016, Crime, Drama, Mystery
Output index of user's favorite and dislike movie from candidate.Please just give the index in [].

248 121

(a) Implicit Feedback

Generate user profile based on the history of user, that each movie with title, year, genre.

History:
[332] Heart and Souls (1993), Comedy|Fantasy
[364] Men with Brooms (2002), Comedy|Drama|Romance

Please output the following infomation of user, output format: {age: , gender: , liked genre: , disliked genre: , liked directors: , country: , language: }

{age: 50, gender: female, liked genre: Comedy|Fantasy, Comedy|Drama|Romance, disliked genre: Thriller, Horror, liked directors: Ron Underwood, country: Canada, United States, language: English}

(b) User Profile

Provide the inquired information of the given movie.
[332] Heart and Souls (1993), Comedy|Fantasy
The inquired information is: director, country, language. And please output them in form of: director, country, language

Ron Underwood, USA, English

(c) Item Attribute

1) Augmenting user-item interactions

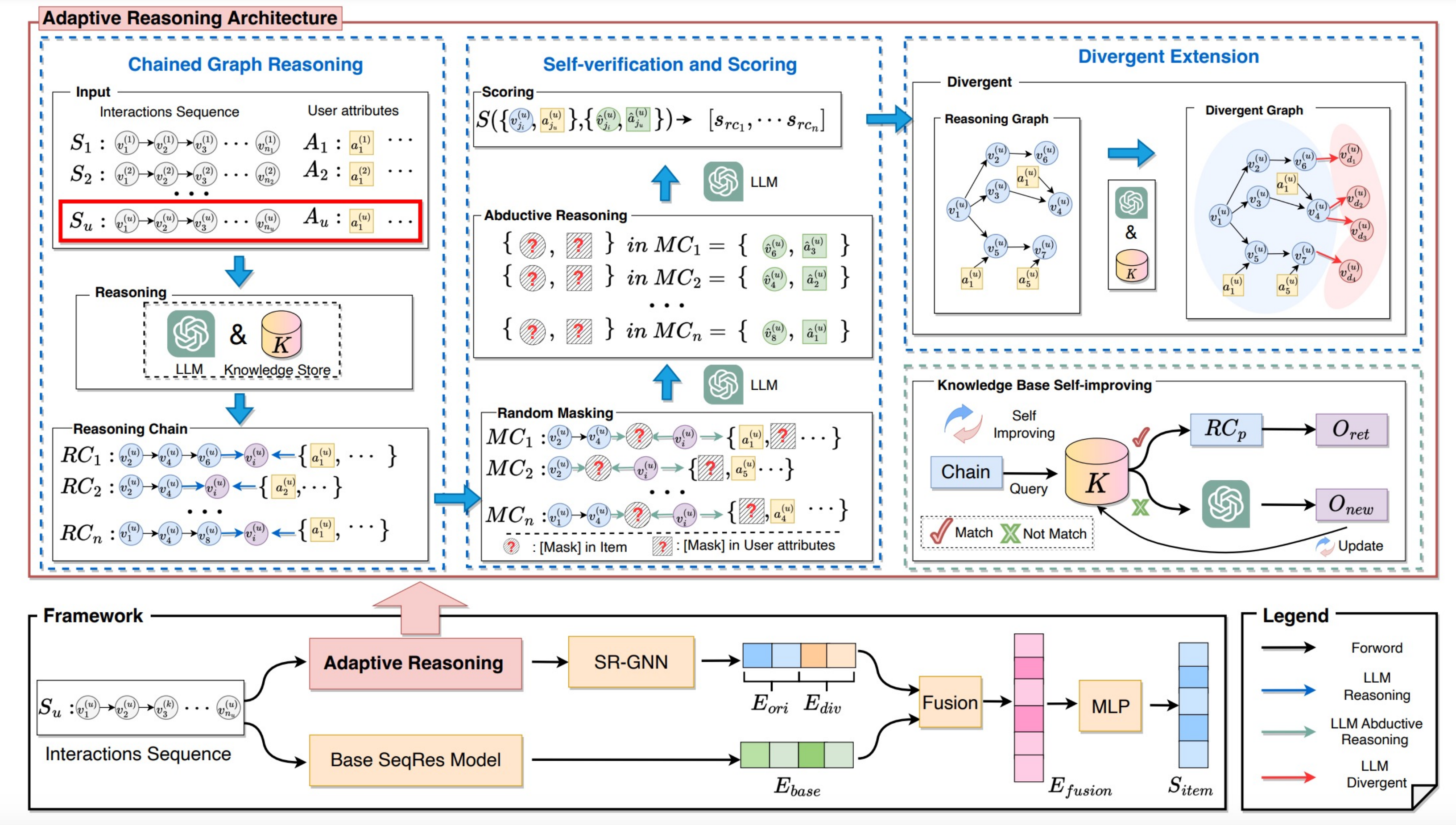
2) Enhancing item attributes

3) User profiling

Wei Wei et al. LLMRec: Large Language Models with Graph Augmentation for Recommendation WSDM 2024

ICL: LLMRG

❑ ICL → reasoning graph → enhancement for recommender systems

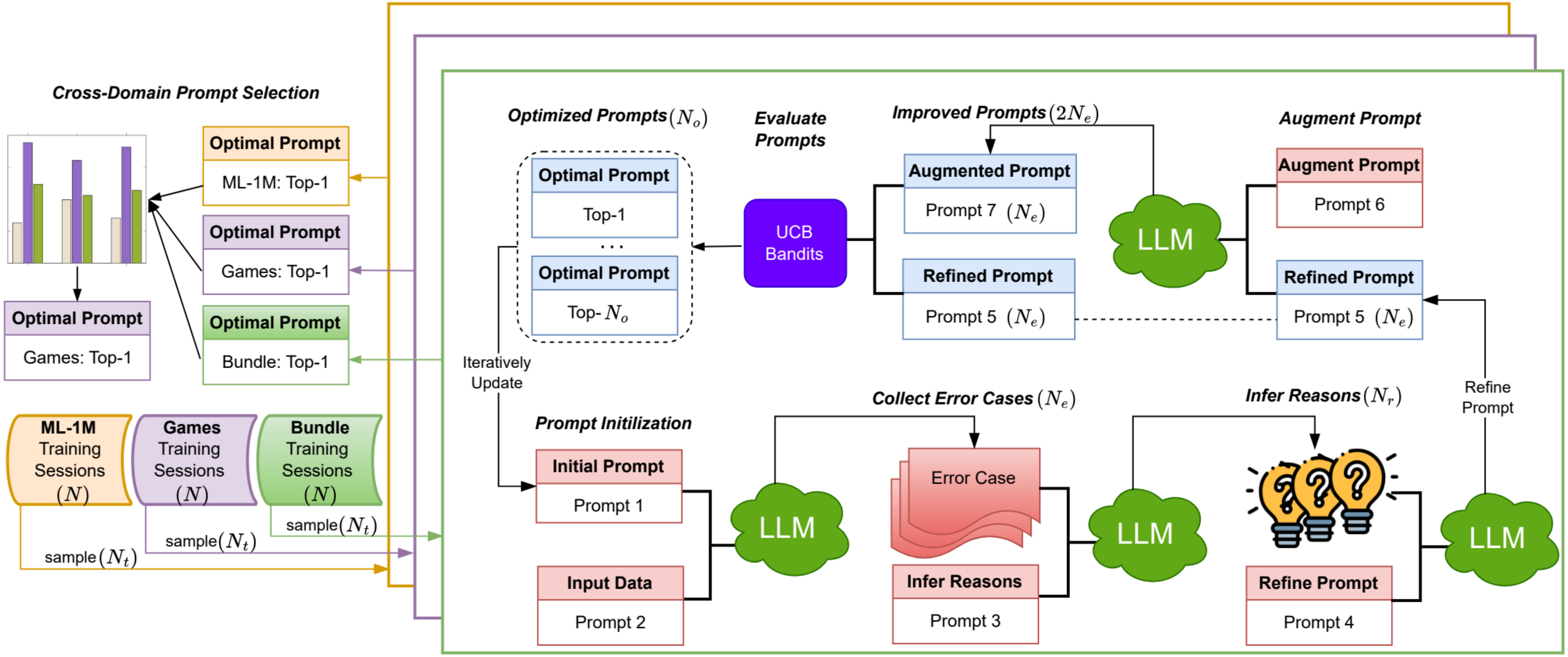


1) Generate reasoning chains for user interactions
2) Extend the user interaction to multiple recommendations via LLM reasoning
3) Score, verify and then filter the reasoning chains
4) Maintain a DB of reasoning chains for inference

Yan Wang et al. Enhancing Recommender Systems with Large Language Model Reasoning Graphs SIGIR 2024

ICL: PO4SIR

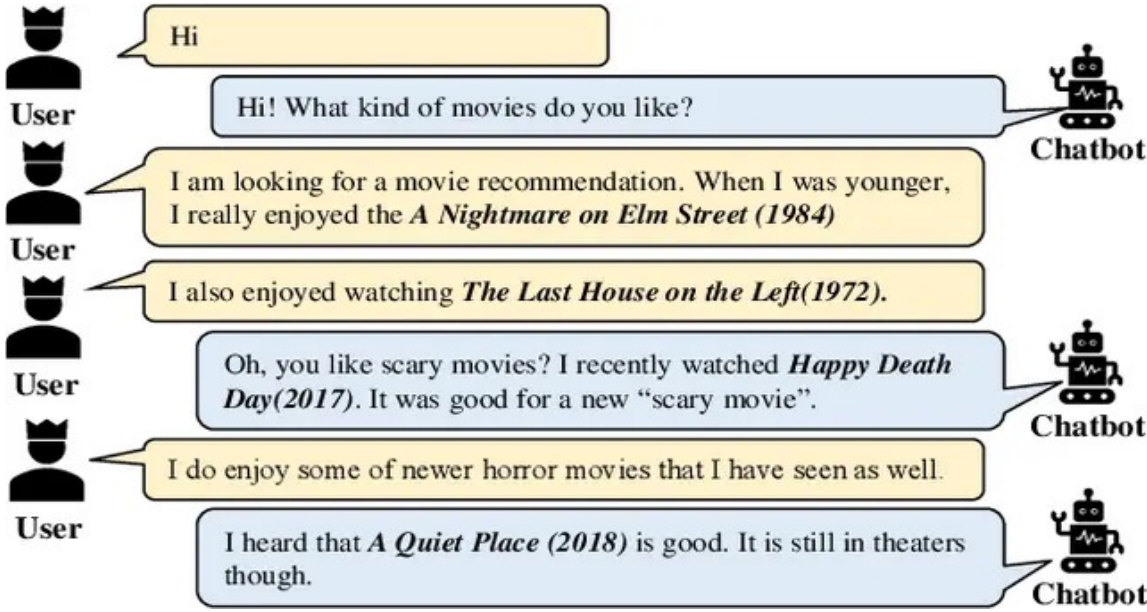
❑ ICL: Automaticaly adjust and optimize prompts for recommendation



Zhu Sun et al. Large Language Models for Intent-Driven Session Recommendations

ICL: KECRS

- ICL for conversational recommender system
  - ➢ Users chat with chatbot with natural language
  - ➢ Chatbot analyses user interest
  - ➢ Chatbot provide recommendaiton


User: Hi

Chatbot: Hi! What kind of movies do you like?

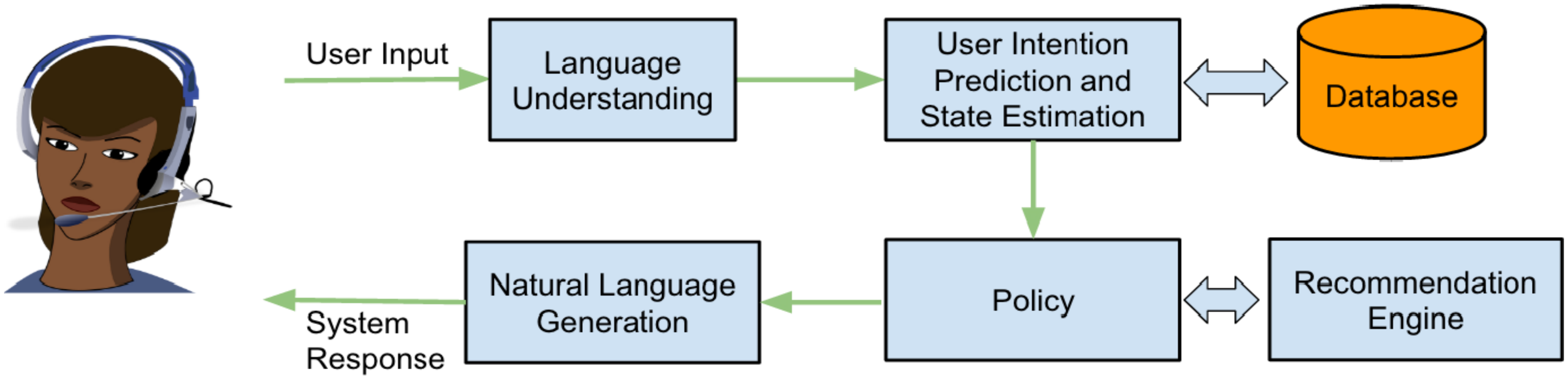
User: I am looking for a movie recommendation. When I was younger, I really enjoyed the ***A Nightmare on Elm Street (1984)***

User: I also enjoyed watching ***The Last House on the Left(1972).***

Chatbot: Oh, you like scary movies? I recently watched ***Happy Death Day(2017)***. It was good for a new "scary movie".

User: I do enjoy some of newer horror movies that I have seen as well.

Chatbot: I heard that ***A Quiet Place (2018)*** is good. It is still in theaters though.


User Input

Language Understanding

User Intention Prediction and State Estimation

Database

Policy

Recommendation Engine

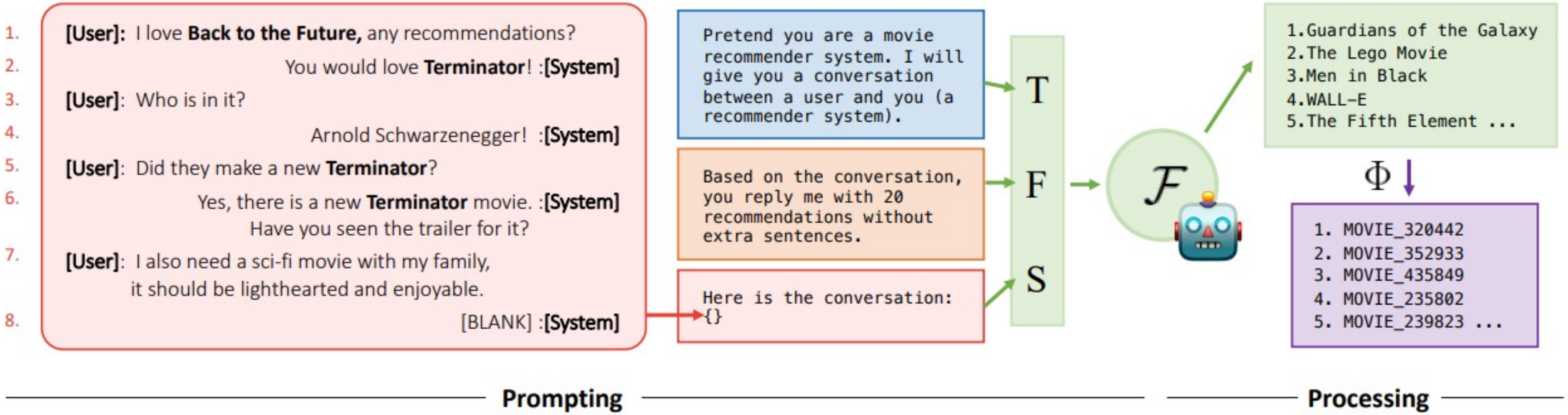
Natural Language Generation

System Response

KECRS: Towards Knowledge-Enriched Conversational Recommendation System

ICL: LLMCRS

■ Framework



1. **[User]:** I love **Back to the Future,** any recommendations?
2. You would love **Terminator**! :**[System]**
3. **[User]:** Who is in it?
4. Arnold Schwarzenegger! :**[System]**
5. **[User]:** Did they make a new **Terminator**?
6. Yes, there is a new **Terminator** movie. :**[System]** Have you seen the trailer for it?
7. **[User]:** I also need a sci-fi movie with my family, it should be lighthearted and enjoyable.
8. [BLANK] :**[System]**

```
Pretend you are a movie
recommender system. I will
give you a conversation
between a user and you (a
recommender system).
```

```
Based on the conversation,
you reply me with 20
recommendations without
extra sentences.
```

```
Here is the conversation:
{}
```

```
1.Guardians of the Galaxy
2.The Lego Movie
3.Men in Black
4.WALL-E
5.The Fifth Element ...
```

```
1. MOVIE_320442
2. MOVIE_352933
3. MOVIE_435849
4. MOVIE_235802
5. MOVIE_239823 ...
```

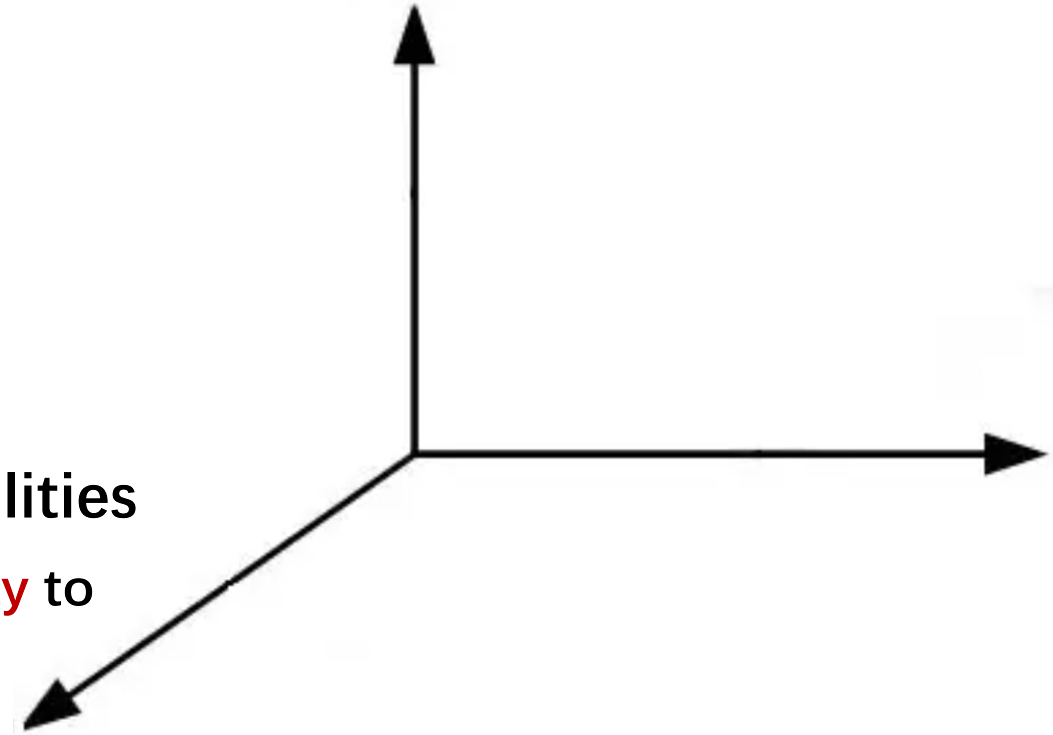
ICL for conversational recommender system

- ➢Input: task description T , format requirement F and conversation context S
- ➢ LLMs analys the input data
- ➢ LLMs generate the recommendation list

He Z, Xie Z, Jha R, et al. Large language models as zero-shot conversational recommenders, CIKM 2023.

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

How to utilize LLMs

In-context learning

Tuning

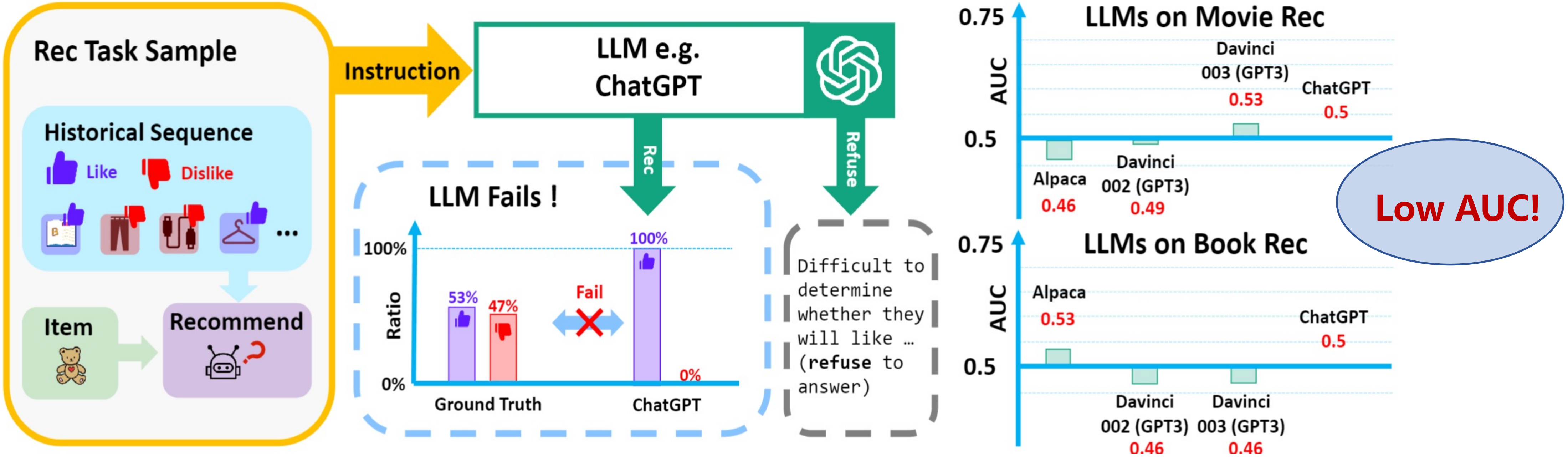
Agent

Information modalities

From the text modality to multiple modalities

Tuning LLM4Rec

- ❑ **In-context learning is not enough.**
- ❑ In complex scenarios, ChatGPT usually gives positive ratings or refuse to answer.



Need to **align** LLM with recommendation task!

Keqin Bao et al. Recsys, TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. 2023

Tuning LLM4Rec

Motivation: lack of recommendation task tuning in LLM pre-training

→ tune LLMs with the recommendation data to align with the recommendation task

Existing work on tuning LLMs for recommendation:

Discriminative manner

Following **traditional rec task**,
provide candidates:
pointwise, pairwise, listwise

PEFT tuning

TALLRec [1]

LLM-TRSR [5]

LLamaRec [4]

GLRec[8] ……

Full tuning

InstructRec[2]

LLMunderPre[3]

……

Generative manner

Following **the pretraining task**,
do not provide candidates:
directly generate items

BigRec [6]

TransRec [7]

LC-Rec [10]

GIRL[9] ……

[1] Bao et al. Recsys, TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. 2023

[2] Zhang et al. Recommendation as instruction following: a large language model empowered recommendation approach. 2023.

[3] Kang el al. Do LLMs Understand User Preferences? Evaluating LLMs on User Rating Prediction. 2023.

[4] Yue et al. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. 2023.

[5] Zhi Zheng et al. Harnessing Large Language Models for Text-Rich Sequential Recommendation. 2024 .....

[6] Bao et al. A Bi-step Grounding Paradigm for Large Language Models in Recommender system. 2023.

[7] Lin et al. A Multi-facet Paradigm to Bridge Large Language Model and Recommendation. KDD 2024.

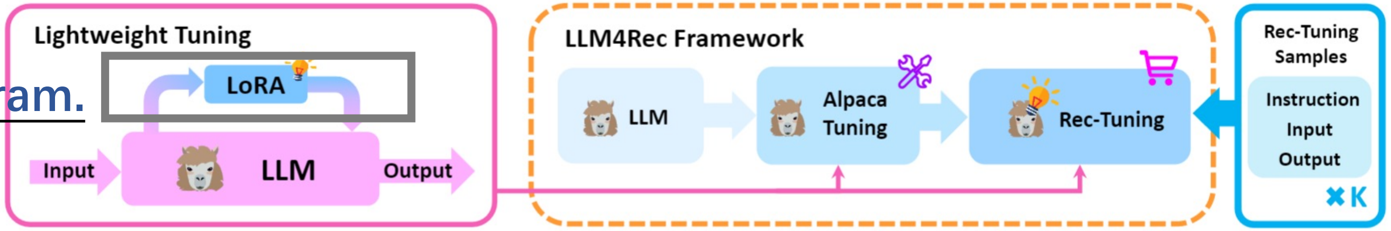
[8] Wu et al. Exploring Large Language Model for Graph Data Understanding in online Job Recommendation. 2023

[9] Zheng et al. Generative job recommendations with large language mode. 2023.

[10] Bowen Zheng et al. "Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation“ ICDE 2024.

Tuning LLM4Rec: TALLRec

❑ TALLRec: Instruction-tuning

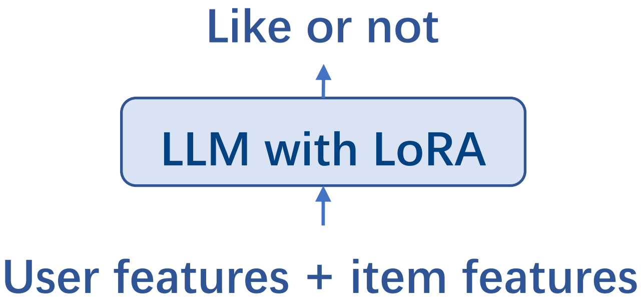


4M param.

7B LLAMA

$$\max_{\Theta} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log\left(P_{\Phi+\Theta}(y_t|x, y_{<t})\right),$$

- Fine-tune 4M parameters by few-shot samples via the **generative loss**
- **Quickly adapt to new tasks**



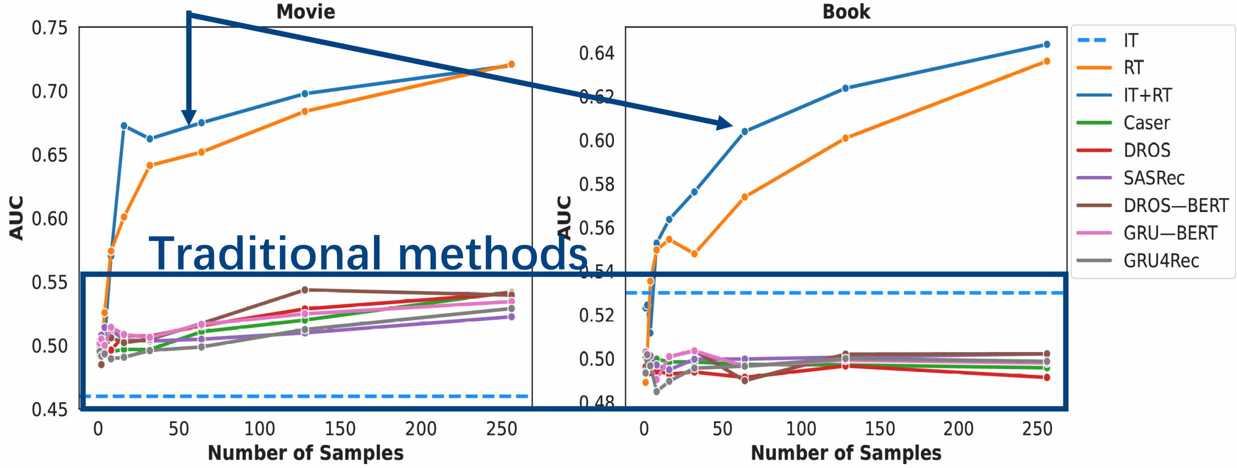
Like or not

LLM with LoRA

User features + item features

- Use item titles as the input
- Better for cold-start recommendation

Performance significantly improves by fine-tuning few-shot samples.

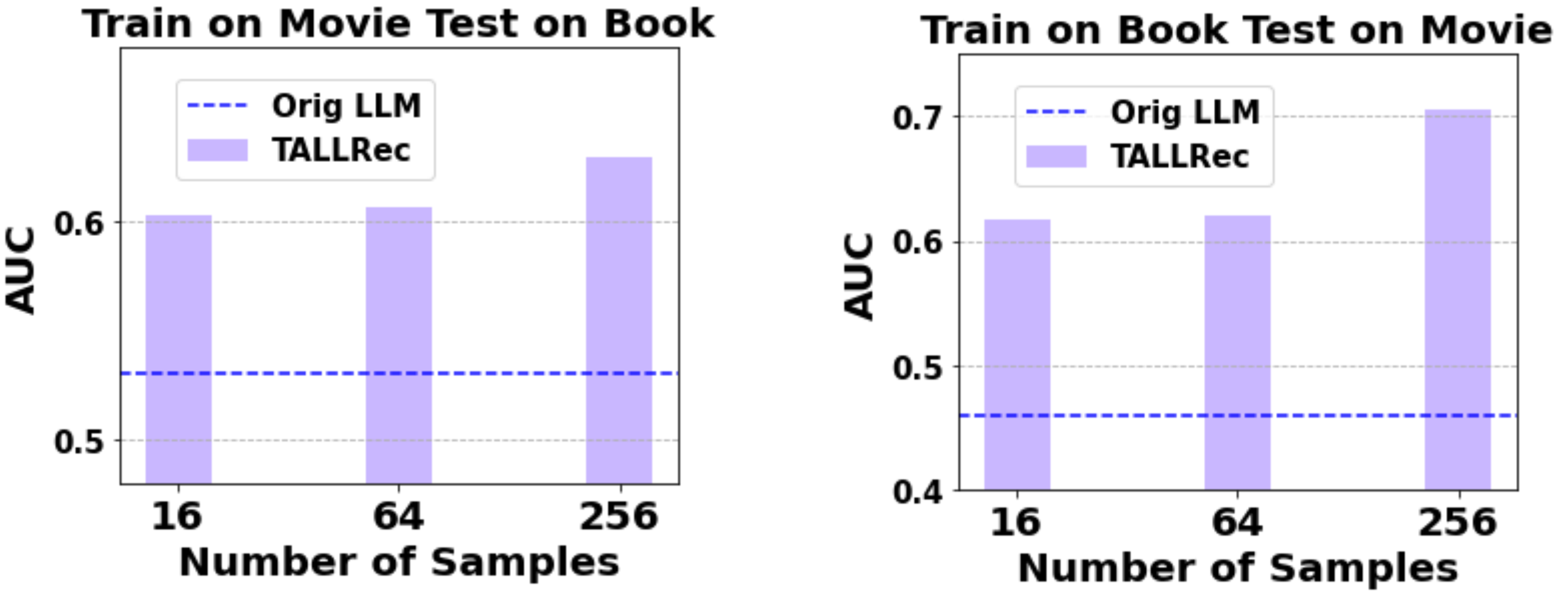


Keqin Bao et al. Recsys, TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. 2023

Tuning LLM4Rec: TALLRec

❑ TALLRec: Cross-domain generalization

- Learning from movie scenario can directly recommend on books, and vice versa
- LLM can leverage domain knowledge to accomplish recommendation tasks after acquiring the ability to recommend.

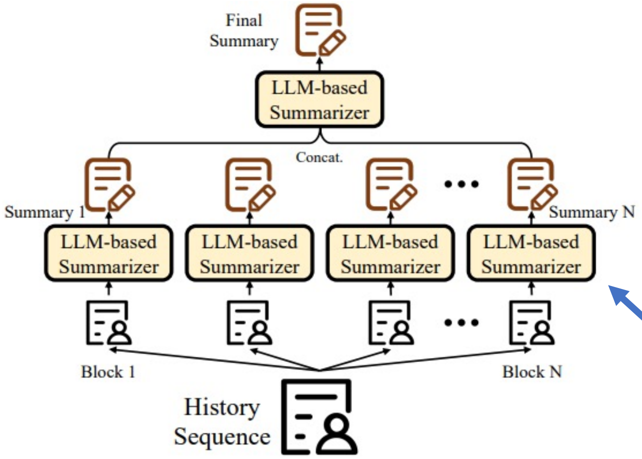


Keqin Bao et al. Recsys, TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. 2023

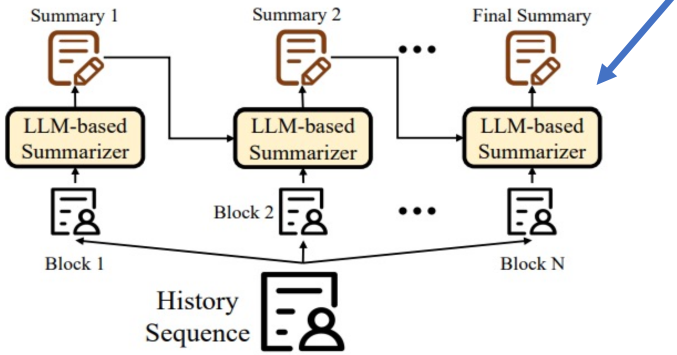
Tuning LLM4Rec: LLM-TRSR

❑ Text-Rich Sequential Recommendation

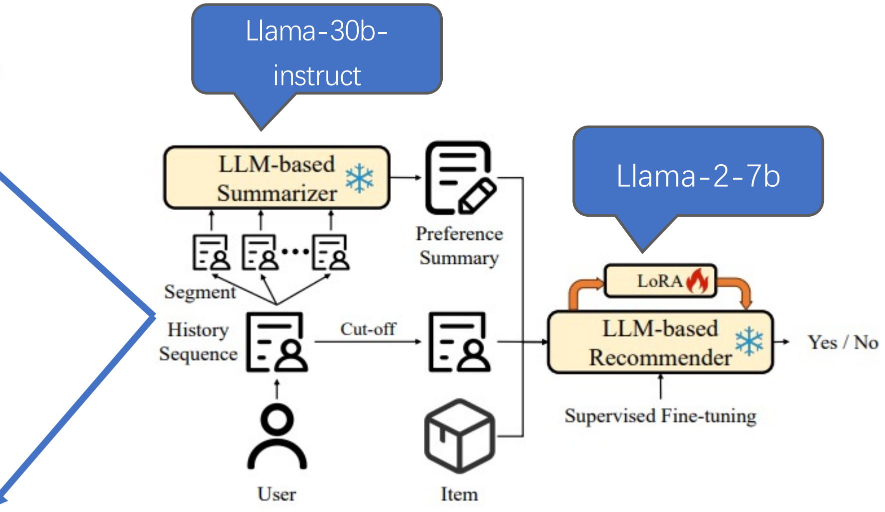
- LLM for preference summary
  - Hierarchical summarization
  - Recurrent summarization
- Supervised fine-tuning
  - Given user preference summary, recently interacted items, and candidate items, LLMs are tuned for recommendation



Hierarchical summarization



Recurrent summarization

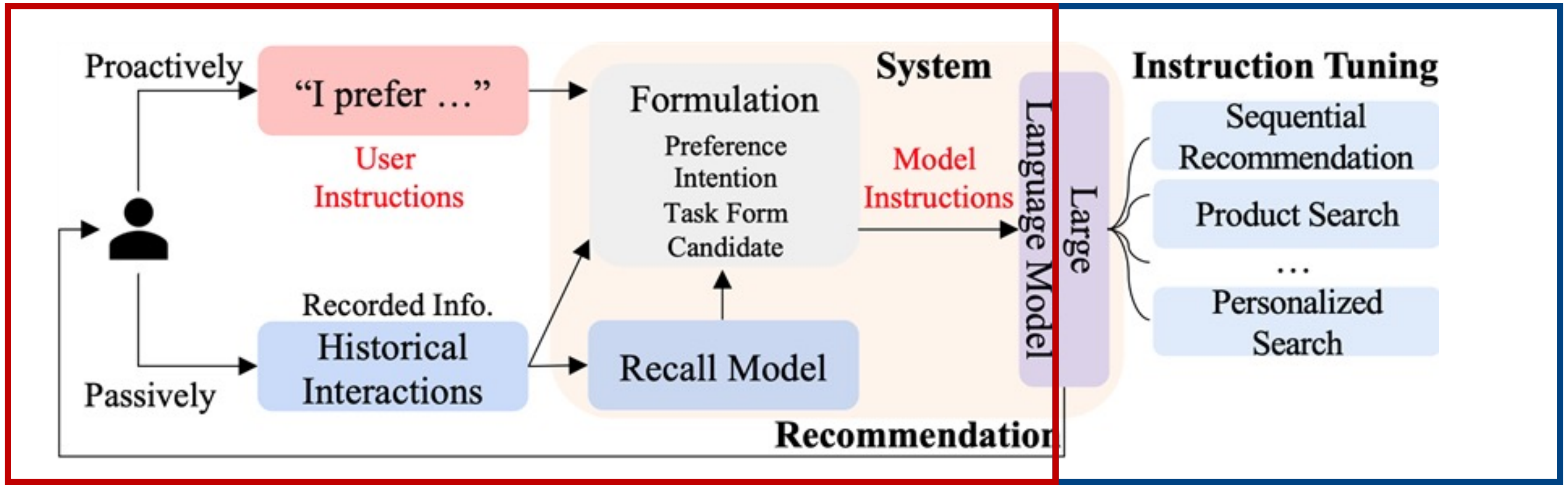


Zhi Zheng et al. Harnessing Large Language Models for Text-Rich Sequential Recommendation. Arxiv 2024

Tuning LLM4Rec: InstructRec

❑ InstructRec

- User could express their needs diversely: vague or specific; implicit or explicit
- LLM should understand and follow different instructions for recommendation



Recommendation instruction construction

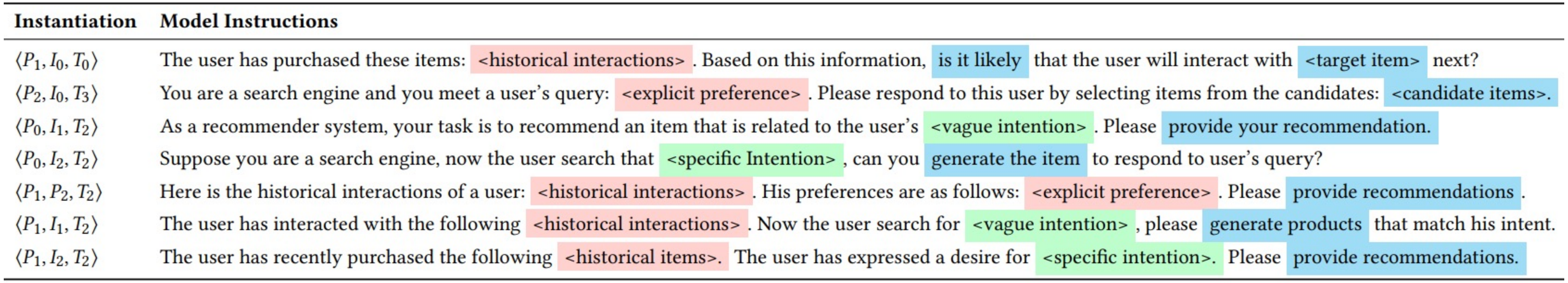
Instruction tuning: tuning LLMs with the instruction data

Junjie Zhang et al. Arxiv, Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach. 2023

Tuning LLM4Rec: InstructRec

❑ InstructRec: Instruction construction:

- **Format:** Preference: none/implicit/explicit Intention: none/vague/specific task: pointwise/pairwise/listwise



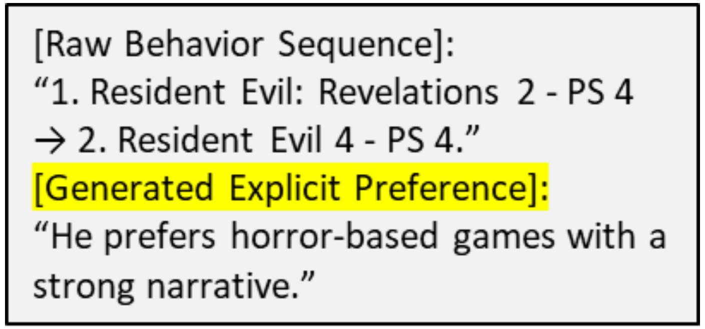
Instantiation	Model Instructions
$\langle P_1, I_0, T_0\rangle$	The user has purchased these items: <historical interactions> . Based on this information, is it likely that the user will interact with <target item> next?
$\langle P_2, I_0, T_3\rangle$	You are a search engine and you meet a user's query: <explicit preference> . Please respond to this user by selecting items from the candidates: <candidate items>.
$\langle P_0, I_1, T_2\rangle$	As a recommender system, your task is to recommend an item that is related to the user's <vague intention> . Please provide your recommendation.
$\langle P_0, I_2, T_2\rangle$	Suppose you are a search engine, now the user search that <specific Intention> , can you generate the item to respond to user's query?
$\langle P_1, P_2, T_2\rangle$	Here is the historical interactions of a user: <historical interactions> . His preferences are as follows: <explicit preference> . Please provide recommendations .
$\langle P_1, I_1, T_2\rangle$	The user has interacted with the following <historical interactions> . Now the user search for <vague intention> , please generate products that match his intent.
$\langle P_1, I_2, T_2\rangle$	The user has recently purchased the following <historical items>. The user has expressed a desire for <specific intention>. Please provide recommendations.

- **Instruction generation: #1** using ChatGPT to generate user preferences and intentions based on interactions

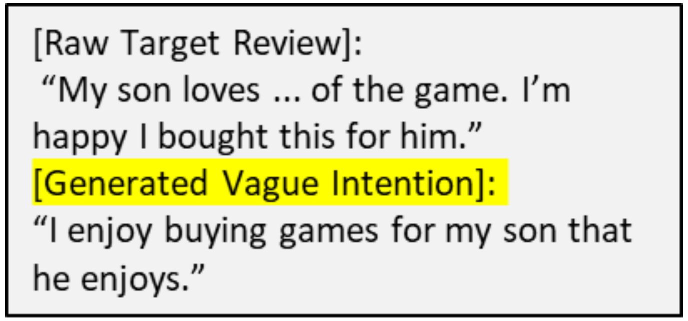
Interaction

↓

Explicit preference



[Raw Behavior Sequence]:
"1. Resident Evil: Revelations 2 - PS 4
→ 2. Resident Evil 4 - PS 4."
[Generated Explicit Preference]:
"He prefers horror-based games with a strong narrative."



[Raw Target Review]:
"My son loves ... of the game. I'm happy I bought this for him."
[Generated Vague Intention]:
"I enjoy buying games for my son that he enjoys."

review

↓

vague intention

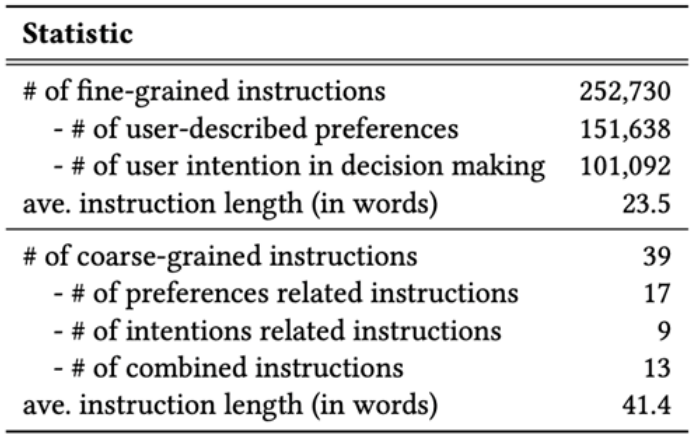
#2 Increasing the instruction diversity via multiple strategies such as CoT

Junjie Zhang et al. Arxiv, Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach. 2023

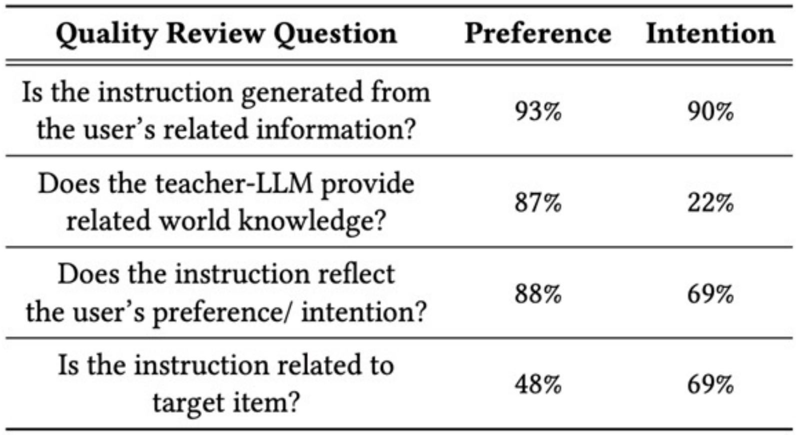
Tuning LLM4Rec: InstructRec

InstructRec

- Instruction construction
  - Quality: human evaluation

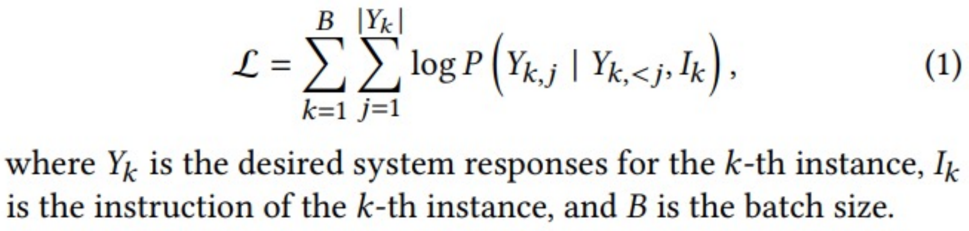


Statistic	
# of fine-grained instructions	252,730
- # of user-described preferences	151,638
- # of user intention in decision making	101,092
ave. instruction length (in words)	23.5
# of coarse-grained instructions	39
- # of preferences related instructions	17
- # of intentions related instructions	9
- # of combined instructions	13
ave. instruction length (in words)	41.4



Quality Review Question	Preference	Intention
Is the instruction generated from the user's related information?	93%	90%
Does the teacher-LLM provide related world knowledge?	87%	22%
Does the instruction reflect the user's preference/ intention?	88%	69%
Is the instruction related to target item?	48%	69%

- Instruction tuning:
  - Supervised fine-tuning, tuning all model parameters (3B Flan-T5-XL)



$$\mathcal{L} = \sum_{k=1}^{B} \sum_{j=1}^{|Y_k|} \log P\left(Y_{k,j} \mid Y_{k,<j}, I_k\right), \qquad (1)$$

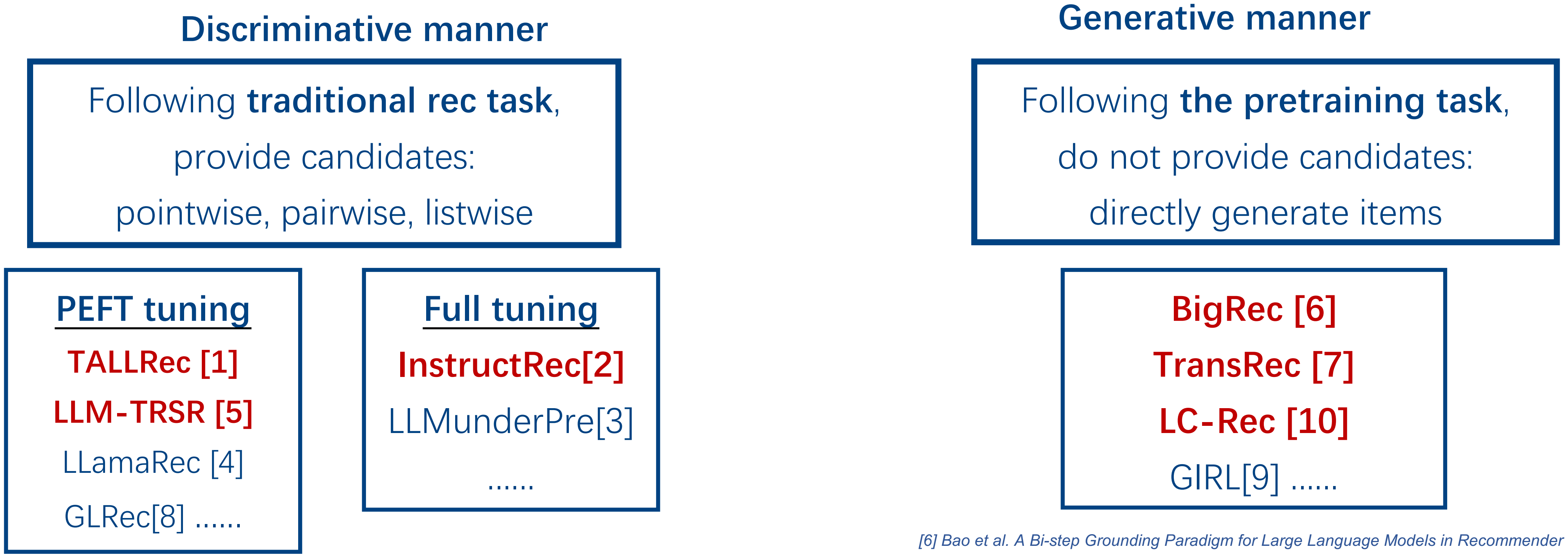
where Y_k is the desired system responses for the k-th instance, I_k is the instruction of the k-th instance, and B is the batch size.

Tuning LLM4Rec

Motivation: lack of recommendation task tuning in LLM pre-training

→ tune LLMs with the recommendation data to align with the recommendation task

Existing work on tuning LLMs for recommendation:



Discriminative manner

Following **traditional rec task**, provide candidates: pointwise, pairwise, listwise

PEFT tuning

TALLRec [1]

LLM-TRSR [5]

LLamaRec [4]

GLRec[8] ……

Full tuning

InstructRec[2]

LLMunderPre[3]

……

Generative manner

Following **the pretraining task**, do not provide candidates: directly generate items

BigRec [6]

TransRec [7]

LC-Rec [10]

GIRL[9] ……

[1] *Bao et al. Recsys, TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. 2023*

[2] *Zhang et al. Recommendation as instruction following: a large language model empowered recommendation approach. 2023.*

[3] *Kang el al. Do LLMs Understand User Preferences? Evaluating LLMs on User Rating Prediction. 2023.*

[4] *Yue et al. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. 2023.*

[5] *Zhi Zheng et al. Harnessing Large Language Models for Text-Rich Sequential Recommendation. 2024 .....*

[6] *Bao et al. A Bi-step Grounding Paradigm for Large Language Models in Recommender system. 2023.*

[7] *Lin et al. A Multi-facet Paradigm to Bridge Large Language Model and Recommendation. KDD 2024.*

[8] *Wu et al. Exploring Large Language Model for Graph Data Understanding in online Job Recommendation. 2023*

[9] *Zheng et al. Generative job recommendations with large language mode. 2023.*

[10] *Bowen Zheng et al. "Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation“ ICDE 2024.*

Tuning LLM4Rec: BIGRec

- **BIGRec**
  - **Generation + Grounding**
    - Given user interaction history in natural language, LLMs aim to generate the next item as recommendation.
    - However, LLMs do not know how to represent an item via token sequence in the recommendation scenario.
    - Besides, the item generated by the LLM may not exist in **the actual world**.

Grounding Paradigm

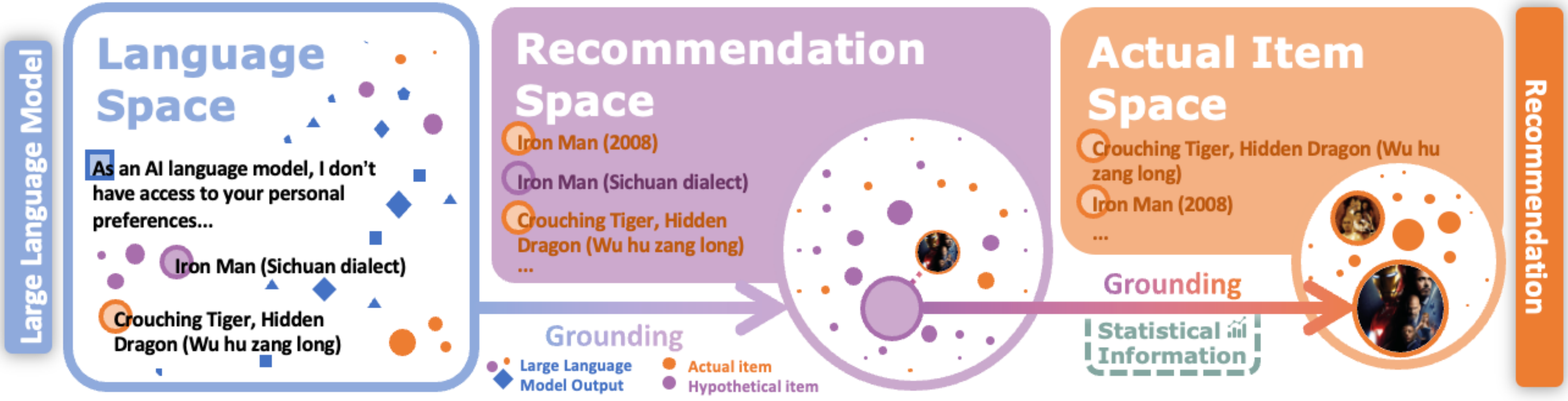
Language Space

↓ *Step1: instruction tuning*

Recommendation Space

↓ *Step2: L2 distance grounding*

Actual Item Space



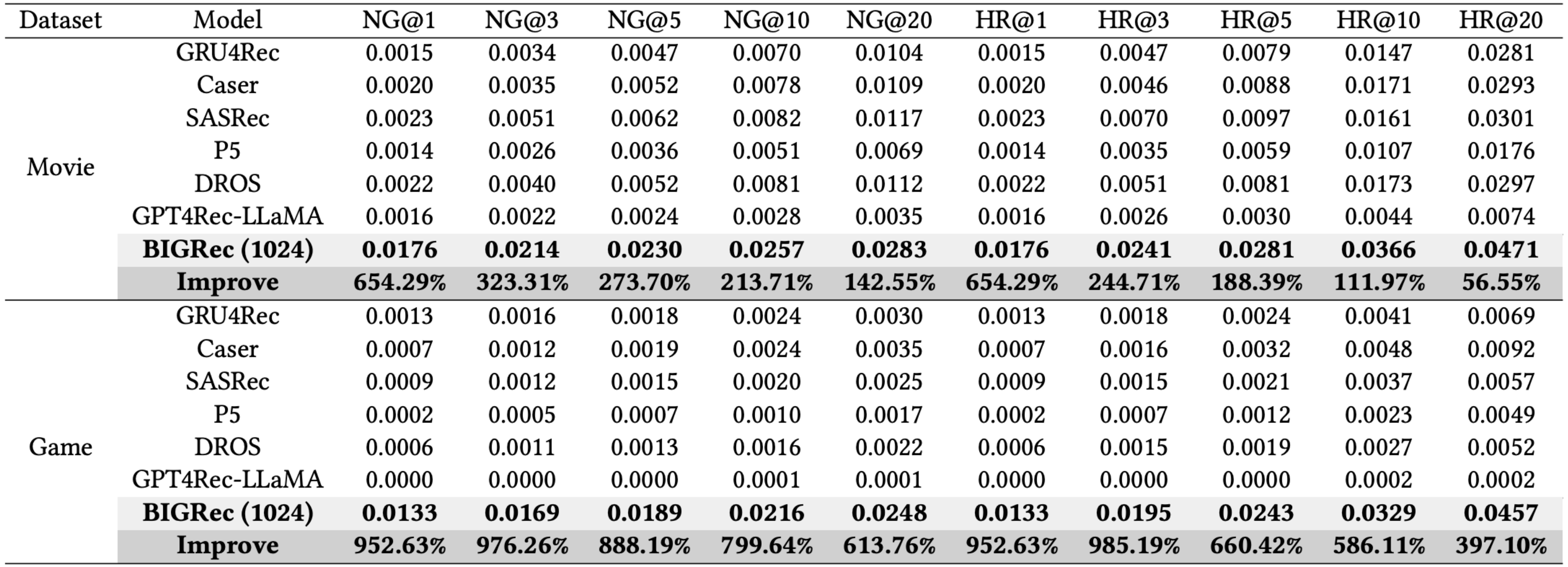
Large Language Model
Language Space
As an AI language model, I don't have access to your personal preferences...
Iron Man (Sichuan dialect)
Crouching Tiger, Hidden Dragon (Wu hu zang long)
Recommendation Space
Iron Man (2008)
Iron Man (Sichuan dialect)
Crouching Tiger, Hidden Dragon (Wu hu zang long)
...
Actual Item Space
Crouching Tiger, Hidden Dragon (Wu hu zang long)
Iron Man (2008)
...
Grounding
Statistical Information
Grounding
Large Language Model Output
Actual item
Hypothetical item
Recommendation

Bao Keqin et al. " A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems" arXiv 2023.

Tuning LLM4Rec: BIGRec

❑ BIGRec

- Few-shot tuning



Dataset	Model	NG@1	NG@3	NG@5	NG@10	NG@20	HR@1	HR@3	HR@5	HR@10	HR@20
Movie	GRU4Rec	0.0015	0.0034	0.0047	0.0070	0.0104	0.0015	0.0047	0.0079	0.0147	0.0281
	Caser	0.0020	0.0035	0.0052	0.0078	0.0109	0.0020	0.0046	0.0088	0.0171	0.0293
	SASRec	0.0023	0.0051	0.0062	0.0082	0.0117	0.0023	0.0070	0.0097	0.0161	0.0301
	P5	0.0014	0.0026	0.0036	0.0051	0.0069	0.0014	0.0035	0.0059	0.0107	0.0176
	DROS	0.0022	0.0040	0.0052	0.0081	0.0112	0.0022	0.0051	0.0081	0.0173	0.0297
	GPT4Rec-LLaMA	0.0016	0.0022	0.0024	0.0028	0.0035	0.0016	0.0026	0.0030	0.0044	0.0074
	BIGRec (1024)	**0.0176**	**0.0214**	**0.0230**	**0.0257**	**0.0283**	**0.0176**	**0.0241**	**0.0281**	**0.0366**	**0.0471**
	Improve	**654.29%**	**323.31%**	**273.70%**	**213.71%**	**142.55%**	**654.29%**	**244.71%**	**188.39%**	**111.97%**	**56.55%**
Game	GRU4Rec	0.0013	0.0016	0.0018	0.0024	0.0030	0.0013	0.0018	0.0024	0.0041	0.0069
	Caser	0.0007	0.0012	0.0019	0.0024	0.0035	0.0007	0.0016	0.0032	0.0048	0.0092
	SASRec	0.0009	0.0012	0.0015	0.0020	0.0025	0.0009	0.0015	0.0021	0.0037	0.0057
	P5	0.0002	0.0005	0.0007	0.0010	0.0017	0.0002	0.0007	0.0012	0.0023	0.0049
	DROS	0.0006	0.0011	0.0013	0.0016	0.0022	0.0006	0.0015	0.0019	0.0027	0.0052
	GPT4Rec-LLaMA	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0002	0.0002
	BIGRec (1024)	**0.0133**	**0.0169**	**0.0189**	**0.0216**	**0.0248**	**0.0133**	**0.0195**	**0.0243**	**0.0329**	**0.0457**
	Improve	**952.63%**	**976.26%**	**888.19%**	**799.64%**	**613.76%**	**952.63%**	**985.19%**	**660.42%**	**586.11%**	**397.10%**

- BIGRec significantly surpasses baselines by few-shot tuning.

- Improvement of BIGRec is significantly higher on Game compared to on Movie.

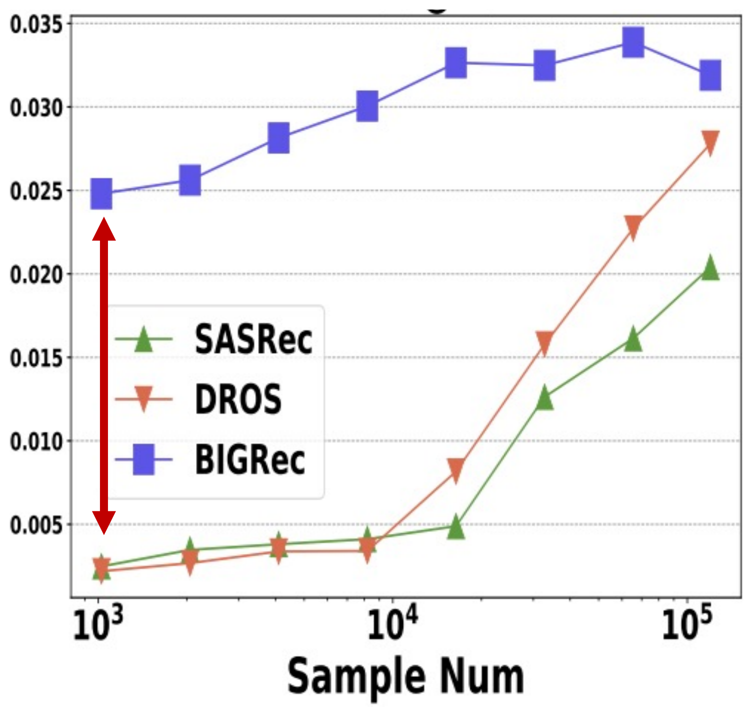
  - possibly due to the varying properties of popularity bias between the two datasets.

Bao Keqin et al. " A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems“ arXiv 2023.

Tuning LLM4Rec: BIGRec

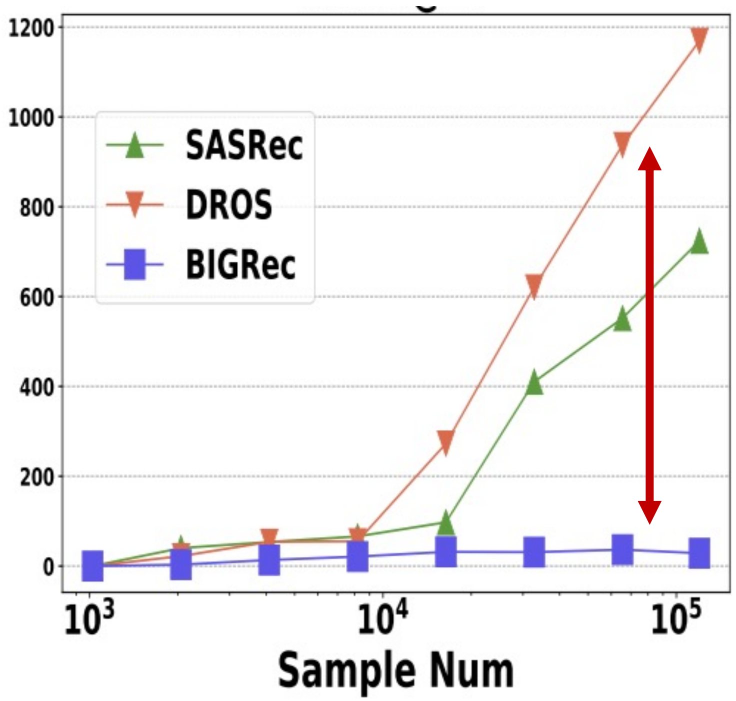
❑ BIGRec

Performance w.r.t. NDCG@20



Quickly adapt to recommendation!

Relative improvements w.r.t. NDCG@20



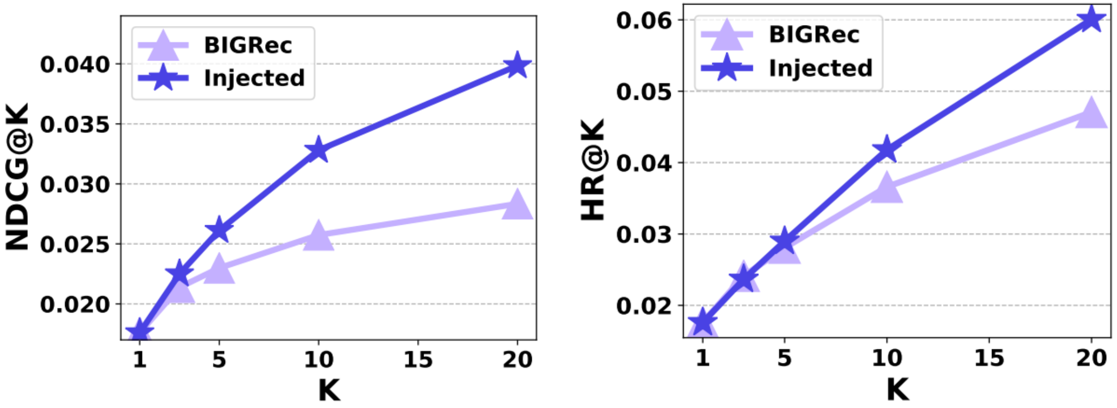
Not proficient in utilizing collaborative filtering signals in interactions!

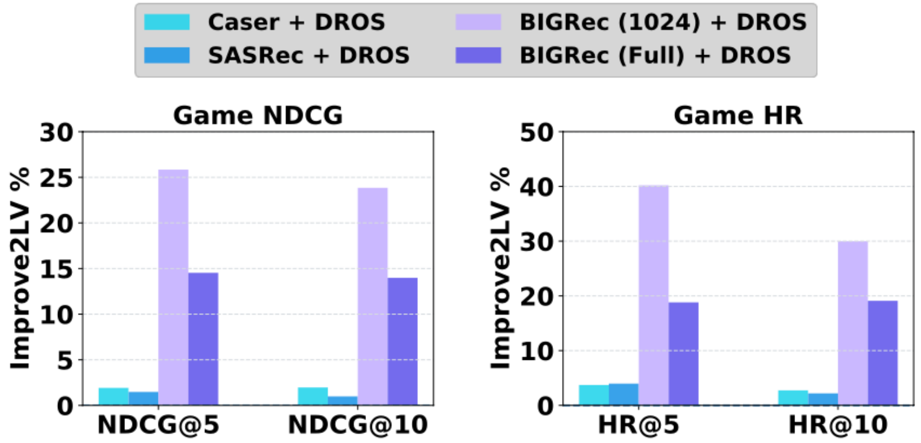
Bao Keqin et al. " A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems“ arXiv 2023.

Tuning LLM4Rec: BIGRec

❑ BIGRec

- Injecting statistical information into BIGRec at Step 2: L2 distance grounding





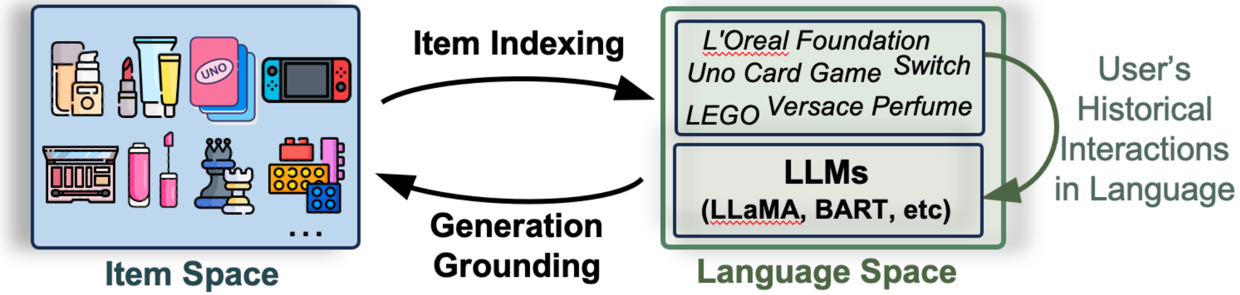
- By incorporating popularity, BIGRec achieves significant improvements *w.r.t.* NDCG@*K* and HR@*K*, particularly for a larger *K*.

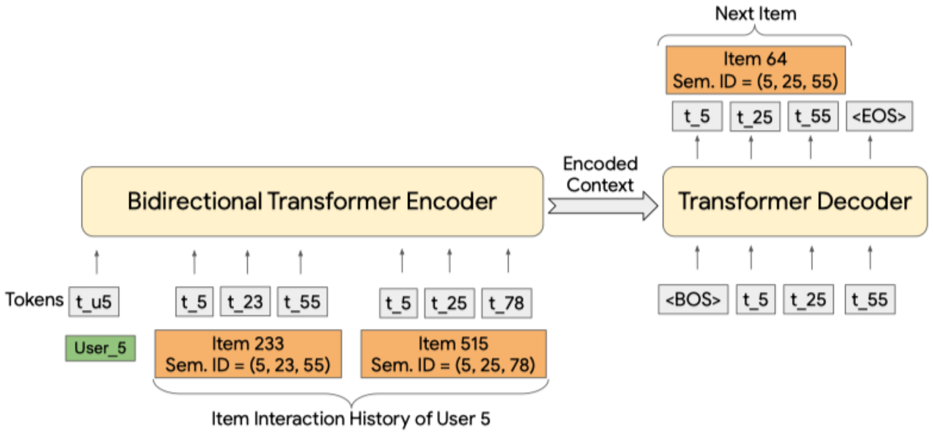
- Incorporating collaborative information into BIGRec yields more significant enhancements than conventional models.

Bao Keqin et al. " A Bi-Step Grounding Paradigm for Large Language Models in Recommendation Systems“ arXiv 2023.

Tuning LLM4Rec

❑ LLM for generative recommendation



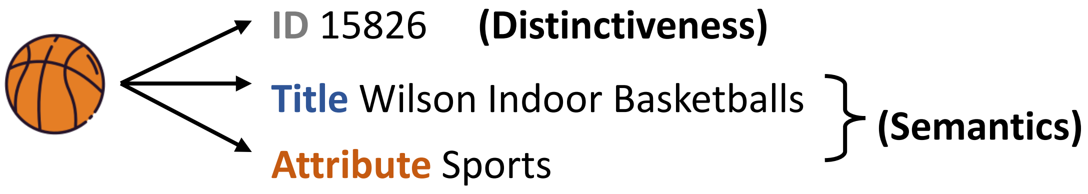


- Two key problems of LLM4Rec
  - Item tokenization: index items into language space
  - Item generation: generate items as recommendations

Lin Xinyu et al. " A Multi-facet Paradigm to Bridge Large Language Model and Recommendation “ KDD 2024.

Tuning LLM4Rec: TransRec

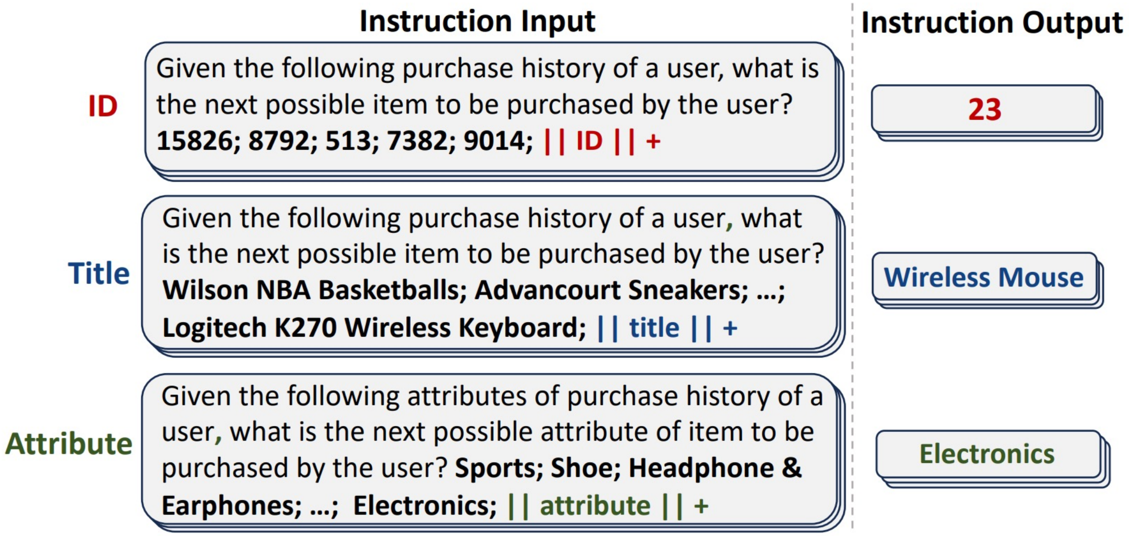
- **Item indexing: multi-facet identifier**



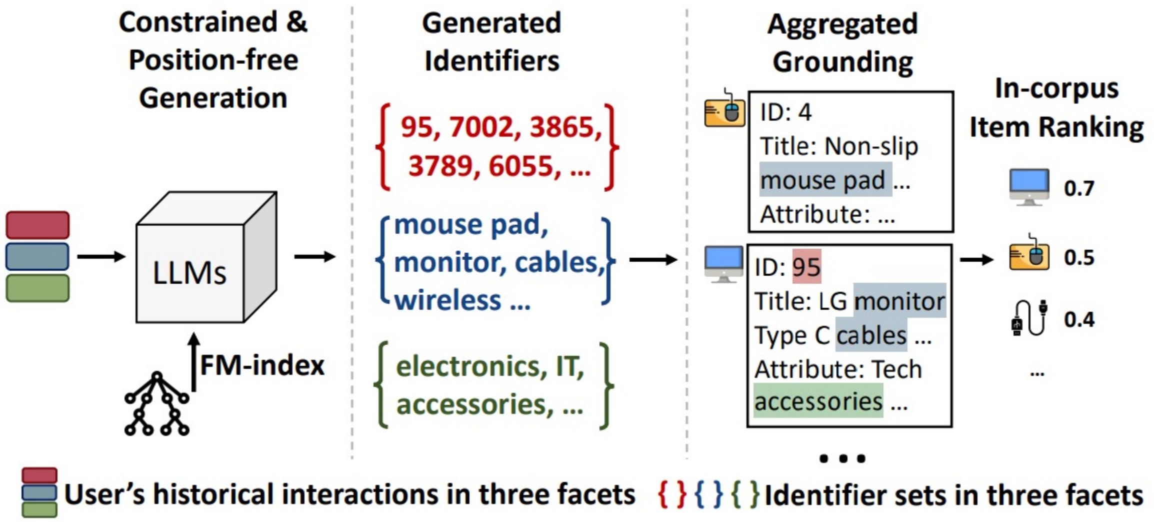
- **Generation grounding:**
  - Position-free constrained generation

FM-index: special prefix tree that supports search from any position of the identifier corpus.

- **Instruction data reconstruction**



	Instruction Input	Instruction Output
ID	Given the following purchase history of a user, what is the next possible item to be purchased by the user? **15826; 8792; 513; 7382; 9014; || ID || +**	23
Title	Given the following purchase history of a user, what is the next possible item to be purchased by the user? **Wilson NBA Basketballs; Advancourt Sneakers; …; Logitech K270 Wireless Keyboard; || title || +**	Wireless Mouse
Attribute	Given the following attributes of purchase history of a user, what is the next possible attribute of item to be purchased by the user? **Sports; Shoe; Headphone & Earphones; …; Electronics; || attribute || +**	Electronics

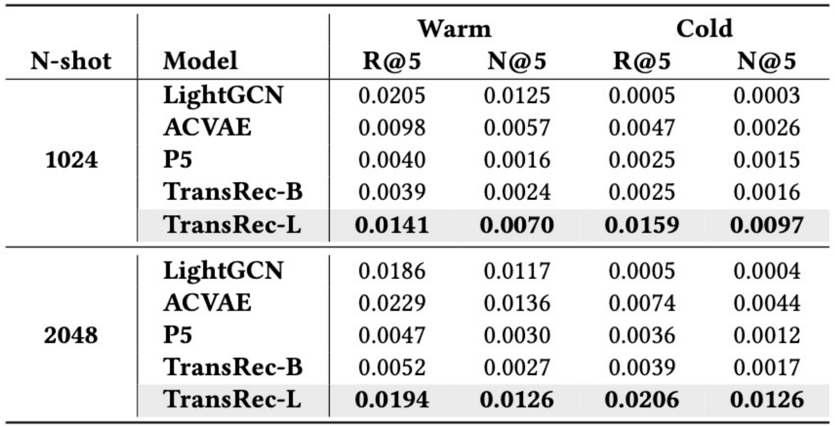

Constrained & Position-free Generation
LLMs
FM-index
Generated Identifiers
95, 7002, 3865, 3789, 6055, …
mouse pad, monitor, cables, wireless …
electronics, IT, accessories, …
Aggregated Grounding
ID: 4
Title: Non-slip mouse pad …
Attribute: …
ID: 95
Title: LG monitor Type C cables …
Attribute: Tech accessories …
• • •
In-corpus Item Ranking
0.7
0.5
0.4
…
User's historical interactions in three facets
{ } { } { } Identifier sets in three facets

Lin Xinyu et al. " A Multi-facet Paradigm to Bridge Large Language Model and Recommendation " KDD 2024.

Tuning LLM4Rec: TransRec

❑ Strong generalization ability

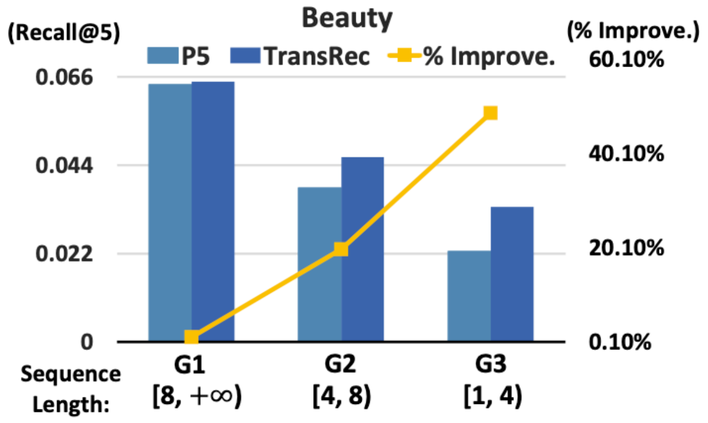
- **Item group analysis**
  - From warm to cold items



N-shot	Model	Warm R@5	Warm N@5	Cold R@5	Cold N@5
1024	LightGCN	0.0205	0.0125	0.0005	0.0003
	ACVAE	0.0098	0.0057	0.0047	0.0026
	P5	0.0040	0.0016	0.0025	0.0015
	TransRec-B	0.0039	0.0024	0.0025	0.0016
	TransRec-L	**0.0141**	**0.0070**	**0.0159**	**0.0097**
2048	LightGCN	0.0186	0.0117	0.0005	0.0004
	ACVAE	0.0229	0.0136	0.0074	0.0044
	P5	0.0047	0.0030	0.0036	0.0012
	TransRec-B	0.0052	0.0027	0.0039	0.0017
	TransRec-L	**0.0194**	**0.0126**	**0.0206**	**0.0126**

* The bold results highlight the superior performance compared to the best LLM-based recommender baseline.

- **User group analysis**
  - From dense users to sparse users

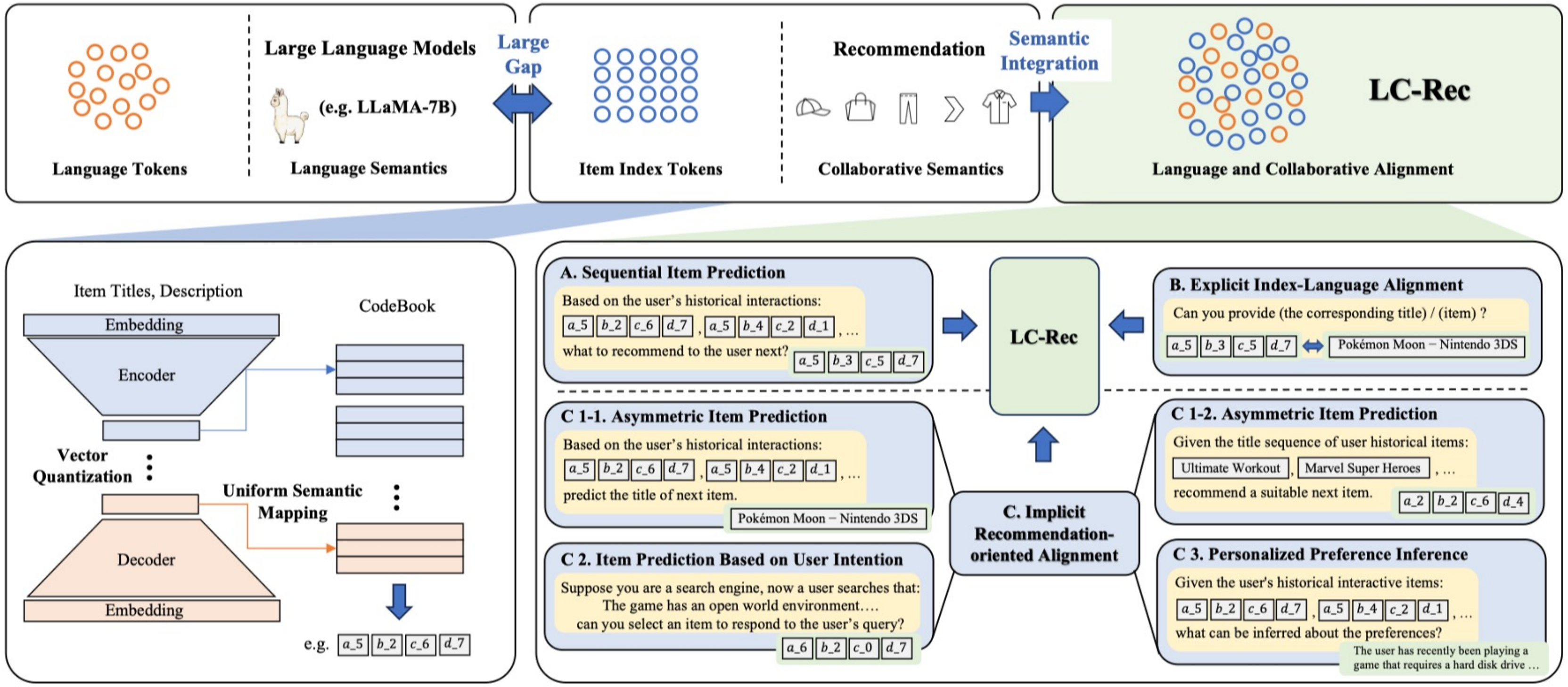


- On the item side, TransRec-L with LLMs has remarkable generalization ability with vase knowledge base, especially on cold-start recommendation under limited data.
- On the user side, TransRec significantly **improves the performance of sparse users** with fewer interactions.

Lin Xinyu et al. " A Multi-facet Paradigm to Bridge Large Language Model and Recommendation " KDD 2024.

Tuning LLM4Rec: LC-Rec

- **LC-Rec**
  - Item indexing: utilize Residual-Quantized Variational AutoEncoder (RQ-VAE) to encode item semantic information as identifiers.
  - Multiple alignment tasks to inject collaborative signals

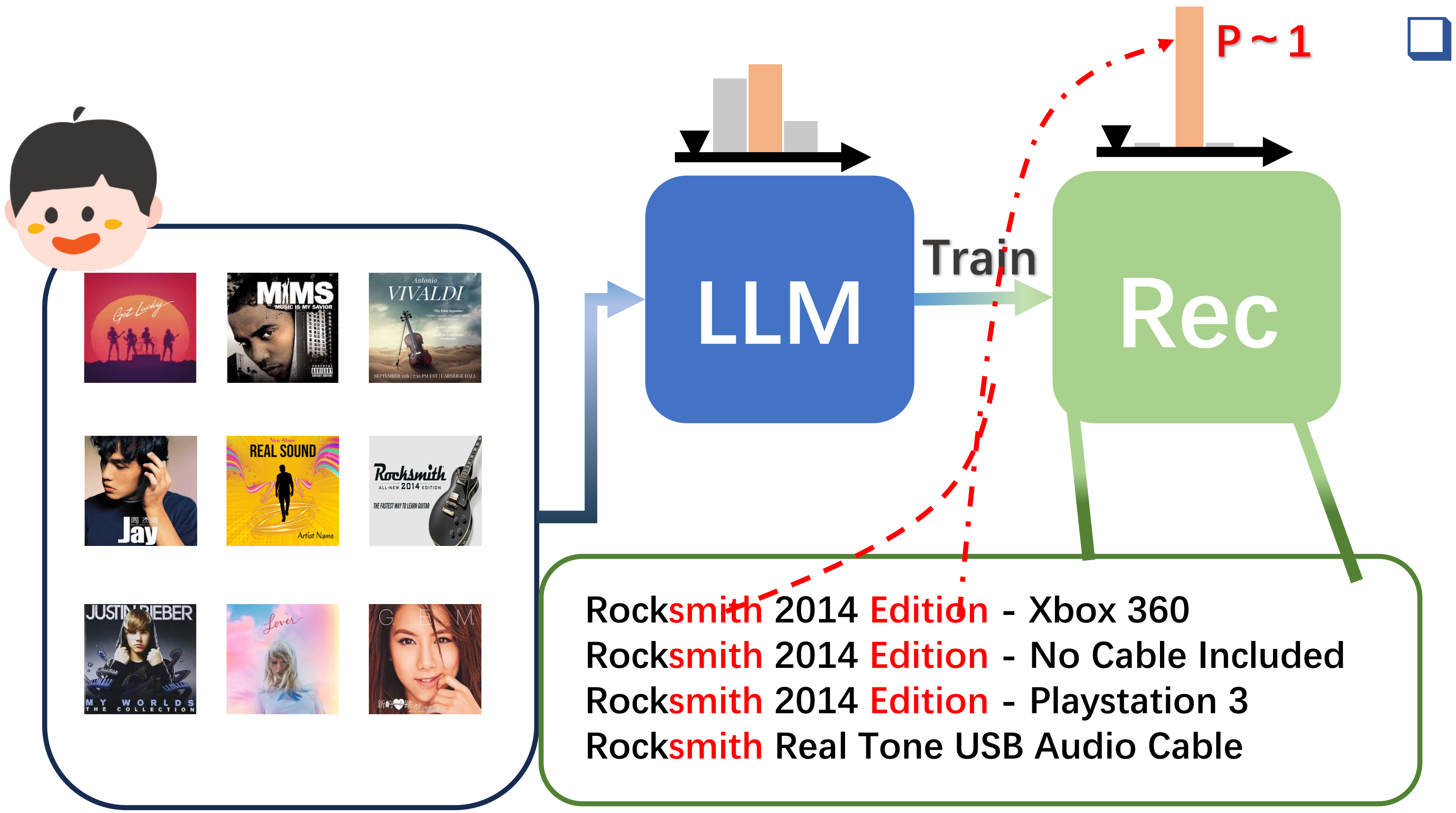


Language Tokens
Large Language Models
(e.g. LLaMA-7B)
Language Semantics
Large Gap
Item Index Tokens
Recommendation
Collaborative Semantics
Semantic Integration
LC-Rec
Language and Collaborative Alignment
Item Titles, Description
Embedding
Encoder
CodeBook
Vector Quantization
Uniform Semantic Mapping
Decoder
Embedding
e.g. a_5 b_2 c_6 d_7
A. Sequential Item Prediction
Based on the user's historical interactions:
a_5 b_2 c_6 d_7 , a_5 b_4 c_2 d_1 , …
what to recommend to the user next?
a_5 b_3 c_5 d_7
LC-Rec
B. Explicit Index-Language Alignment
Can you provide (the corresponding title) / (item) ?
a_5 b_3 c_5 d_7
Pokémon Moon – Nintendo 3DS
C 1-1. Asymmetric Item Prediction
Based on the user's historical interactions:
a_5 b_2 c_6 d_7 , a_5 b_4 c_2 d_1 , …
predict the title of next item.
Pokémon Moon – Nintendo 3DS
C 1-2. Asymmetric Item Prediction
Given the title sequence of user historical items:
Ultimate Workout , Marvel Super Heroes , …
recommend a suitable next item.
a_2 b_2 c_6 d_4
C. Implicit Recommendation-oriented Alignment
C 2. Item Prediction Based on User Intention
Suppose you are a search engine, now a user searches that:
The game has an open world environment.…
can you select an item to respond to the user's query?
a_6 b_2 c_0 d_7
C 3. Personalized Preference Inference
Given the user's historical interactive items:
a_5 b_2 c_6 d_7 , a_5 b_4 c_2 d_1 , …
what can be inferred about the preferences?
The user has recently been playing a game that requires a hard disk drive …

Bowen Zheng et al. "Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation" ICDE 2024.

LLM4Rec Decoding

- ❑ **Language decoding ≠ Recommendation decoding**



Rocksmith 2014 Edition - Xbox 360
Rocksmith 2014 Edition - No Cable Included
Rocksmith 2014 Edition - Playstation 3
Rocksmith Real Tone USB Audio Cable

❑ Amplification Bias

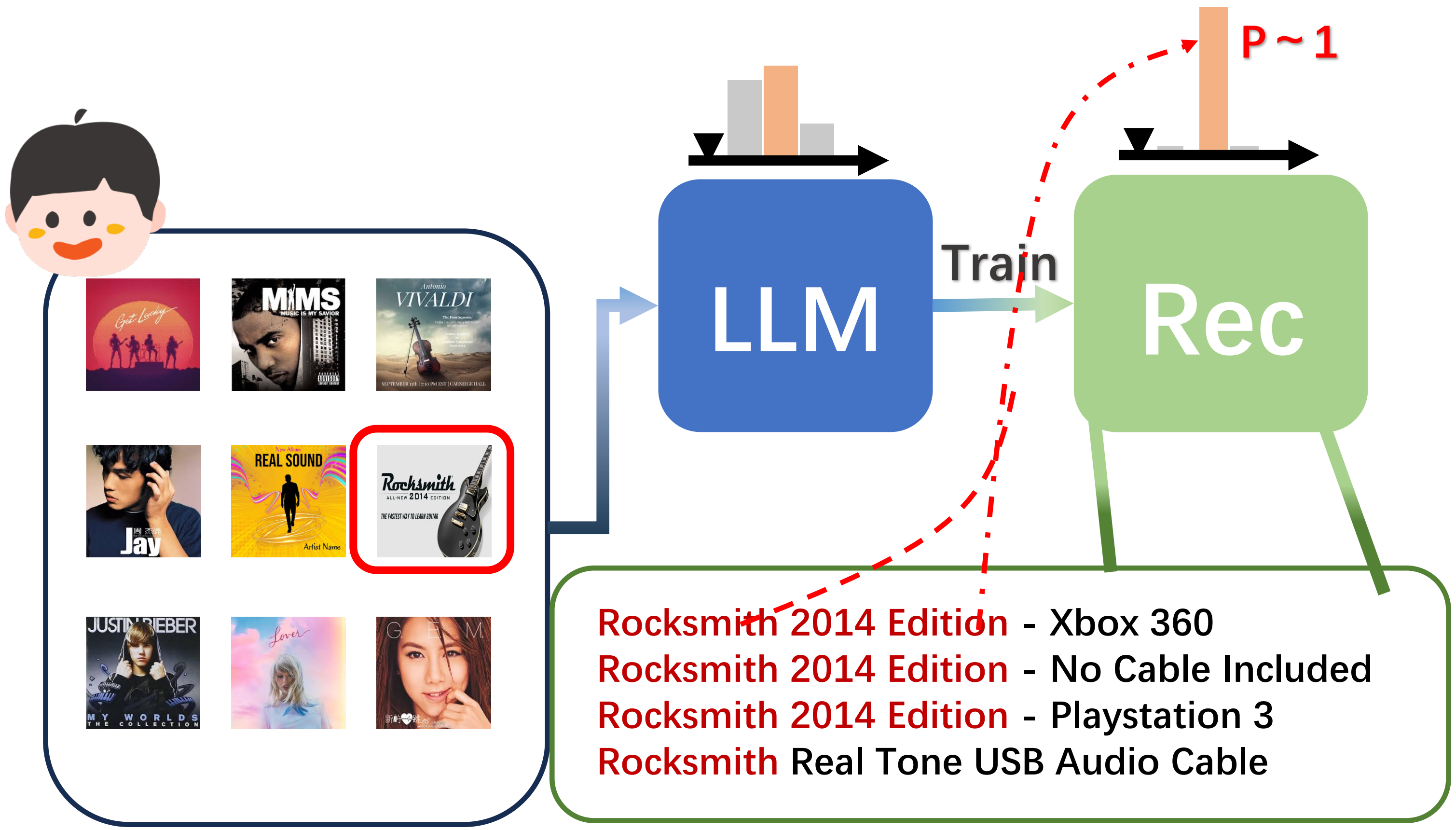
- ❑ The generation probabilities of some tokens are **close to 1** under the condition of already generated tokens. (e.g. “smith” & “Edition”)
- ❑ Length normalization tends to enhance scores for items contains more of those tokens

→ Remove length normalization

Keqin Bao et al. " Decoding Matters: Addressing Amplification Bias and Homogeneity Issue for LLM-based Recommendation" arXiv 2024.

LLM4Rec Decoding

❑ Language decoding ≠ Recommendation decoding



Rocksmith 2014 Edition - Xbox 360
Rocksmith 2014 Edition - No Cable Included
Rocksmith 2014 Edition - Playstation 3
Rocksmith Real Tone USB Audio Cable

❑ Homogeneity Issue

- Recommend items with similar content and structures
- Frequently repeats item features based on past user interactions

→ Use a text-free model to assist the LLM to decode

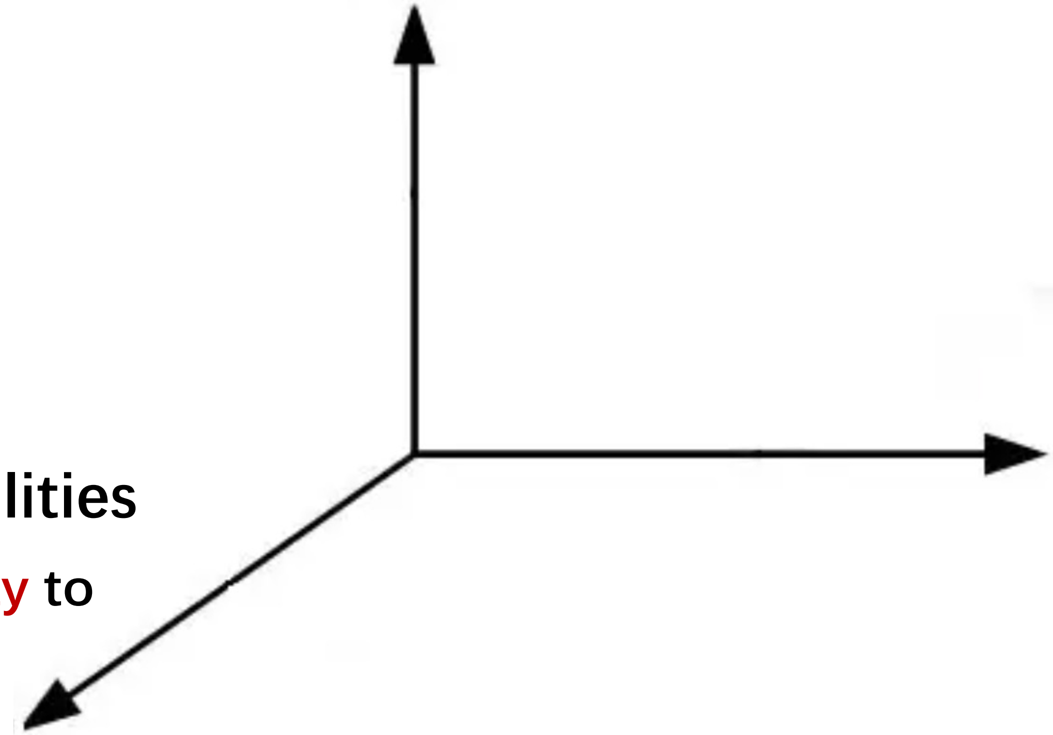
$$L_{TF}(h_t|h_{\le t-1}) = \log\left(\frac{\sum_{i\in I_{h_{<t},h_t}} p_{TF}(I_i)}{\sum_{i\in I_{h_{<t}}} p_{TF}(I_i)}\right)$$

$$S_{TF}(h_{\le t}) = S_{TF}(h_{\le t-1}) + L_{TF}(h_t|h_{\le t\,-1})$$

Keqin Bao et al. " Decoding Matters: Addressing Amplification Bias and Homogeneity Issue for LLM-based Recommendation" arXiv 2024.

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

How to utilize LLMs

In-context learning

Tuning

Agent

Information modalities

From the text modality to multiple modalities

Agent for Recommendation

❑ LLM-empowered Agents for Recommendation

❑ Agent as User Simulator

- Main idea: using agents to simulate user behavior for real-world recommendation.
- RecAgent[1], Agent4Rec[2]

❑ Agent for Recommendation

- Main idea: harnessing the powerful capabilities of LLMs, such as reasoning, reflection, planning and tool usage, for recommendation.
- RecMind[3], InteRecAgent[4], BiLLP[5], Multi-Agent Collaboration[6]

[1] Lei Wang et al. "When Large Language Model based Agent Meets User Behavior Analysis: A Novel User Simulation Paradigm" arXiv 2023.

[2] Zhang An et al. "On Generative Agents in Recommendation" arXiv 2023.

[3] Wang Yancheng et al. "RecMind: Large Language Model Powered Agent For Recommendation" arXiv 2023.

[4] Xu Huang et al. "Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations" arxiv 2023.

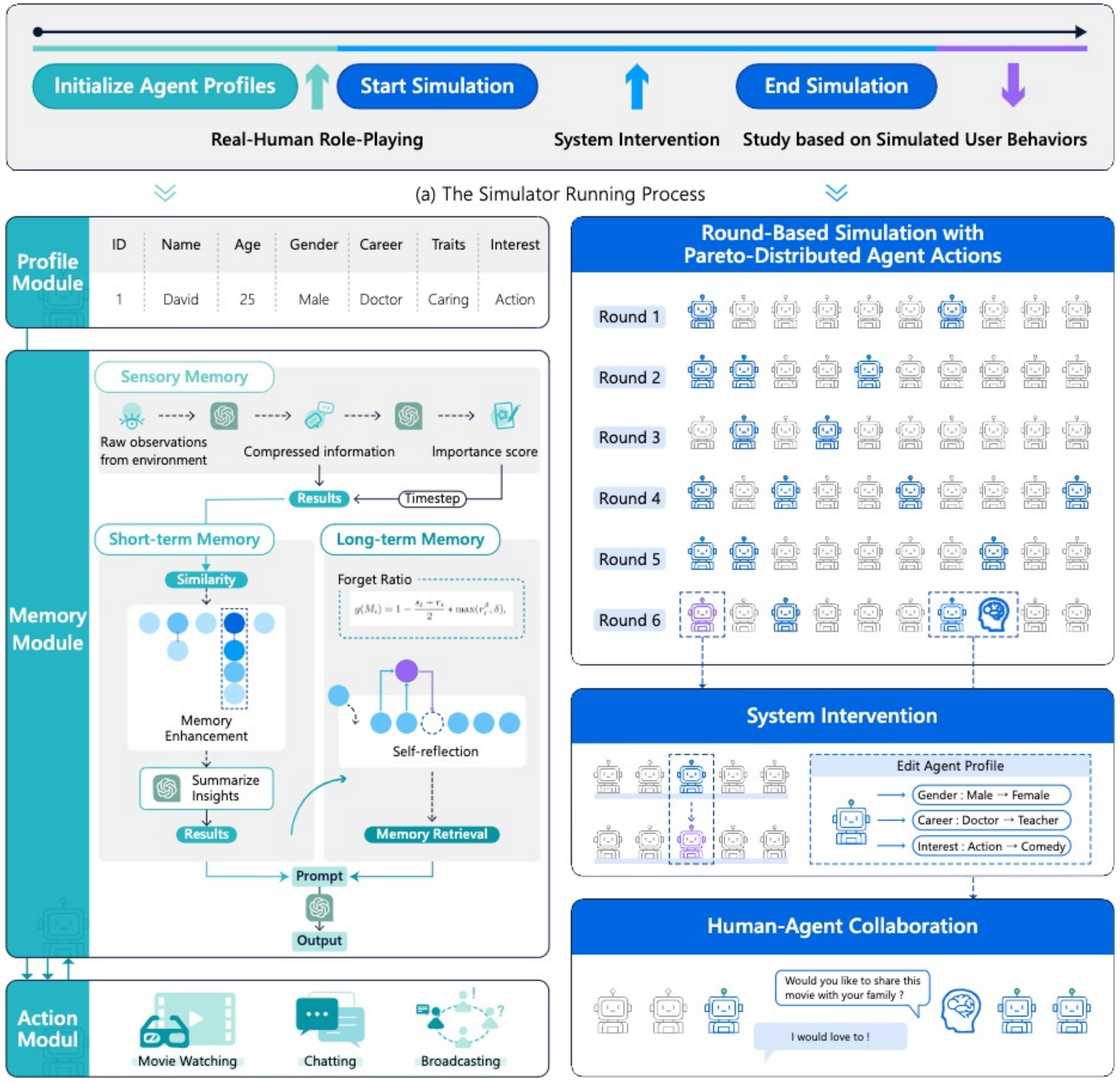
[5] Wentao Shi et al. 2023. Large Language Models are Learnable Planners for Long-Term Recommendation. in SIGIR 2024.

[6] Jiabao Fang et al. A Multi-Agent Conversational Recommender System. Arxiv 2024

Agent: RecAgent

❑ LLM-based agent for user simulation

- User simulation is a fundamental problem in human-centered applications.
- Traditional methods struggle to simulate complex user behaviors.
- LLMs show potential in human-level intelligence and generalization capabilities.



(a) The Simulator Running Process

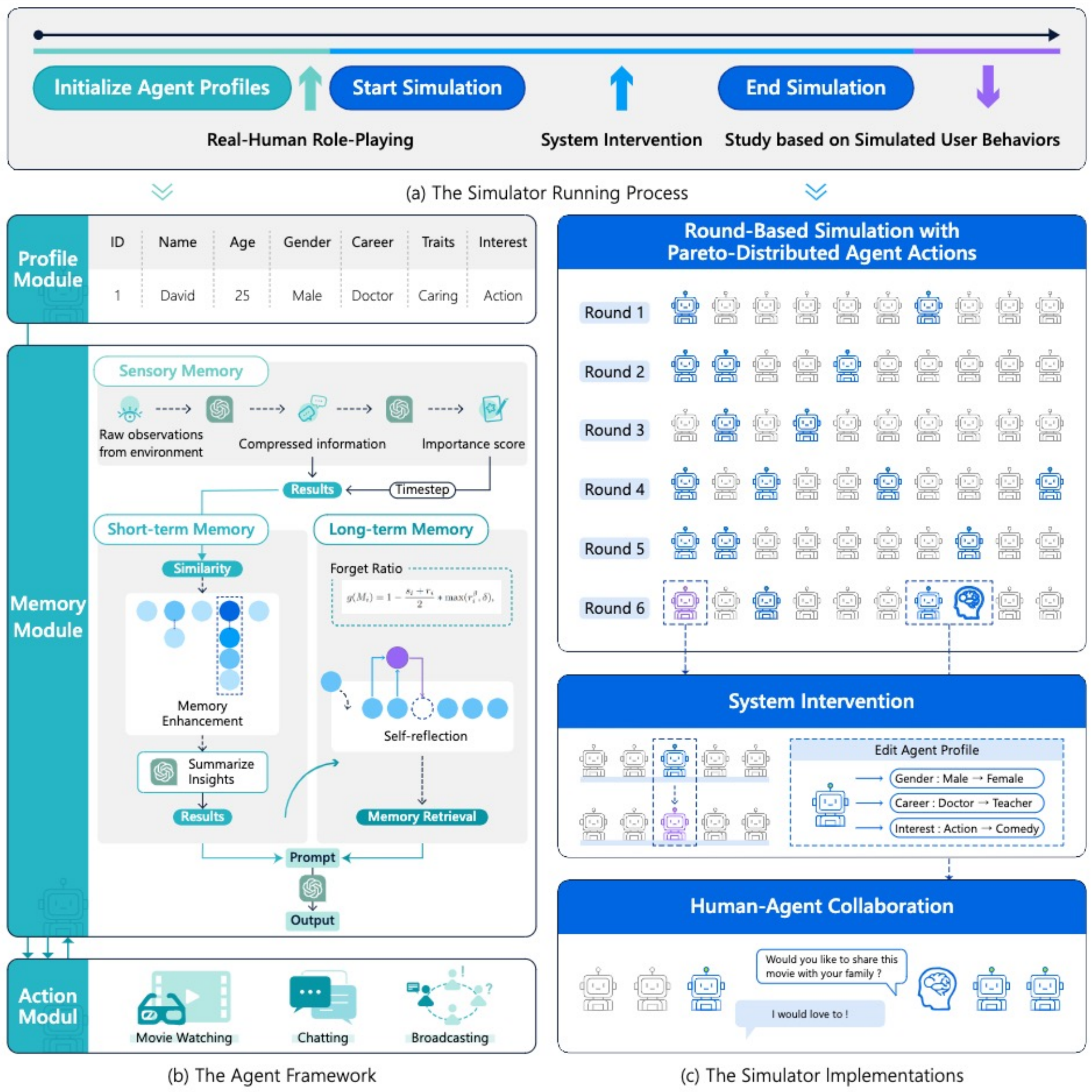
(b) The Agent Framework

(c) The Simulator Implementations

Lei Wang et al. " User Behavior Simulation with Large Language Model based Agents" arXiv 2023.

Agent: RecAgent

- ❑ **Recommendation Behaviors**
  Agent chooses to **search or receive recommendations**, selects movies, and **stores** feelings after watching.
- ❑ **Chatting Behaviors**
  Two agents **discuss and stored** the conversation in their memories.
- ❑ **Broadcasting Behaviors**
  An agent **posts** a message on social media, **received by friends** and stored in their memories.



(a) The Simulator Running Process

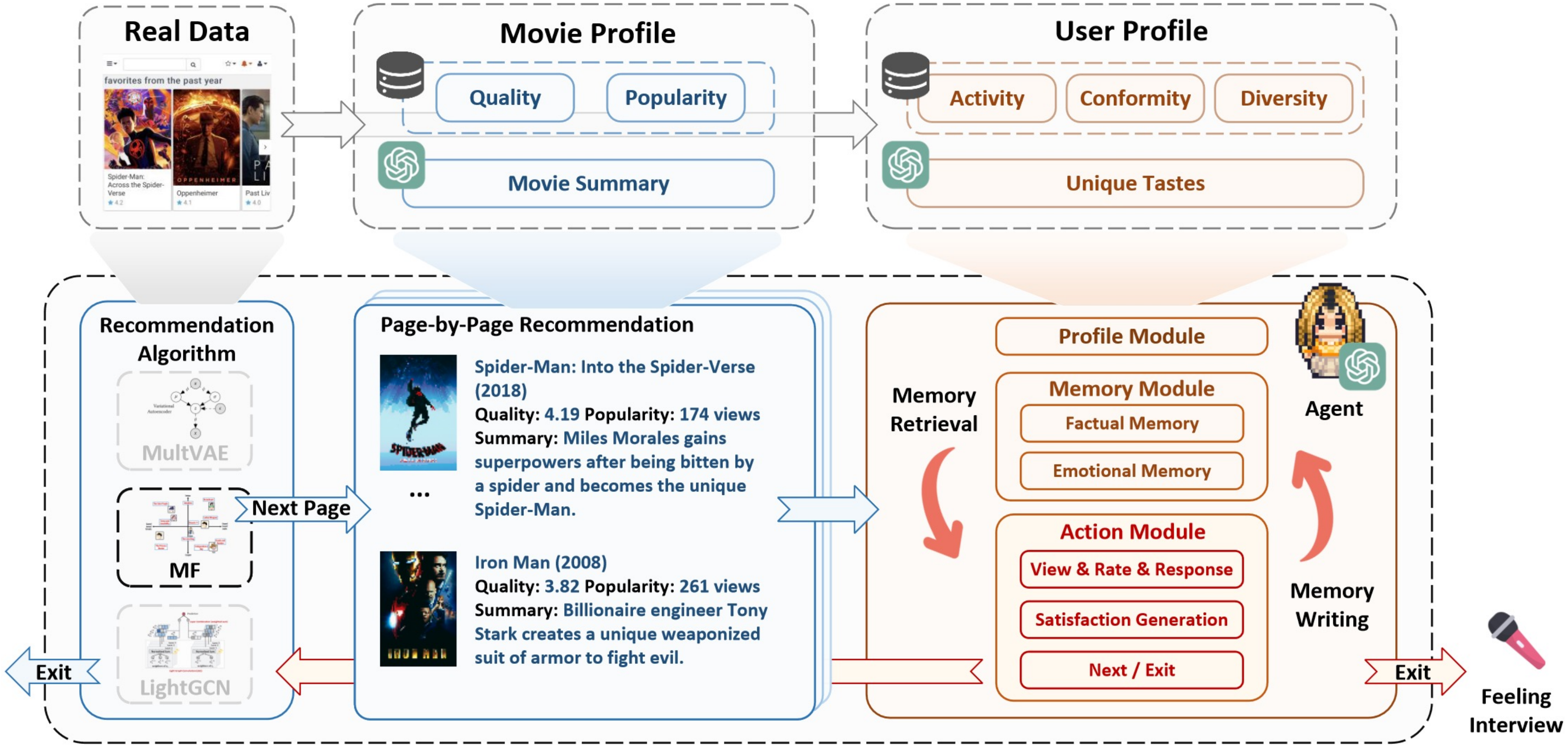
(b) The Agent Framework

(c) The Simulator Implementations

Lei Wang et al. " User Behavior Simulation with Large Language Model based Agents" arXiv 2023.

Agent: Agent4Rec

- ❑ Agent4Rec, a simulator with 1,000 LLM-empowered generative agents.
- ❑ Agents are trained by the MovieLens-1M dataset, embodying varied social traits and preferences.
- ❑ Each agent interacts with personalized movie recommendations in a page-by-page manner and undertakes various actions such as watching, rating, evaluating, exiting, and interviewing.



Real Data
favorites from the past year
Spider-Man: Across the Spider-Verse
Oppenheimer
Past Liv
Movie Profile
Quality
Popularity
Movie Summary
User Profile
Activity
Conformity
Diversity
Unique Tastes
Recommendation Algorithm
MultVAE
MF
LightGCN
Next Page
Exit
Page-by-Page Recommendation
Spider-Man: Into the Spider-Verse (2018)
Quality: 4.19 Popularity: 174 views
Summary: Miles Morales gains superpowers after being bitten by a spider and becomes the unique Spider-Man.
...
Iron Man (2008)
Quality: 3.82 Popularity: 261 views
Summary: Billionaire engineer Tony Stark creates a unique weaponized suit of armor to fight evil.
Memory Retrieval
Profile Module
Memory Module
Factual Memory
Emotional Memory
Action Module
View & Rate & Response
Satisfaction Generation
Next / Exit
Agent
Memory Writing
Exit
Feeling Interview

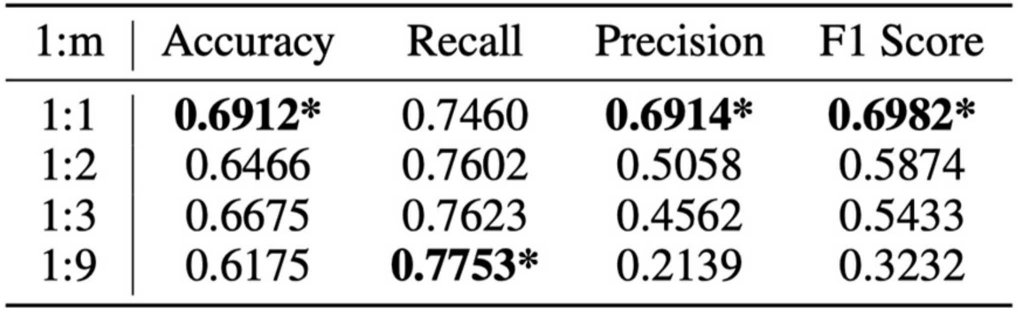
Zhang An et al. "On Generative Agents in Recommendation" arXiv 2023.

Agent: Agent4Rec

- ❑ To what extent can LLM-empowered generative agents truly simulate the behavior of genuine, independent humans in recommender systems?

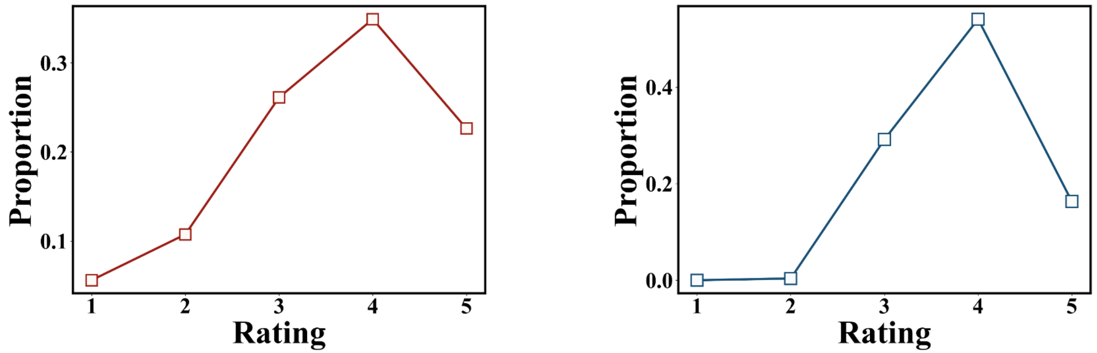
- ❑ User Taste Alignment

Table 1: User taste discrimination.



1:m	Accuracy	Recall	Precision	F1 Score
1:1	**0.6912***	0.7460	**0.6914***	**0.6982***
1:2	0.6466	0.7602	0.5058	0.5874
1:3	0.6675	0.7623	0.4562	0.5433
1:9	0.6175	**0.7753***	0.2139	0.3232

- ❑ Rating Distribution Alignment



(a) Distribution on MovieLens

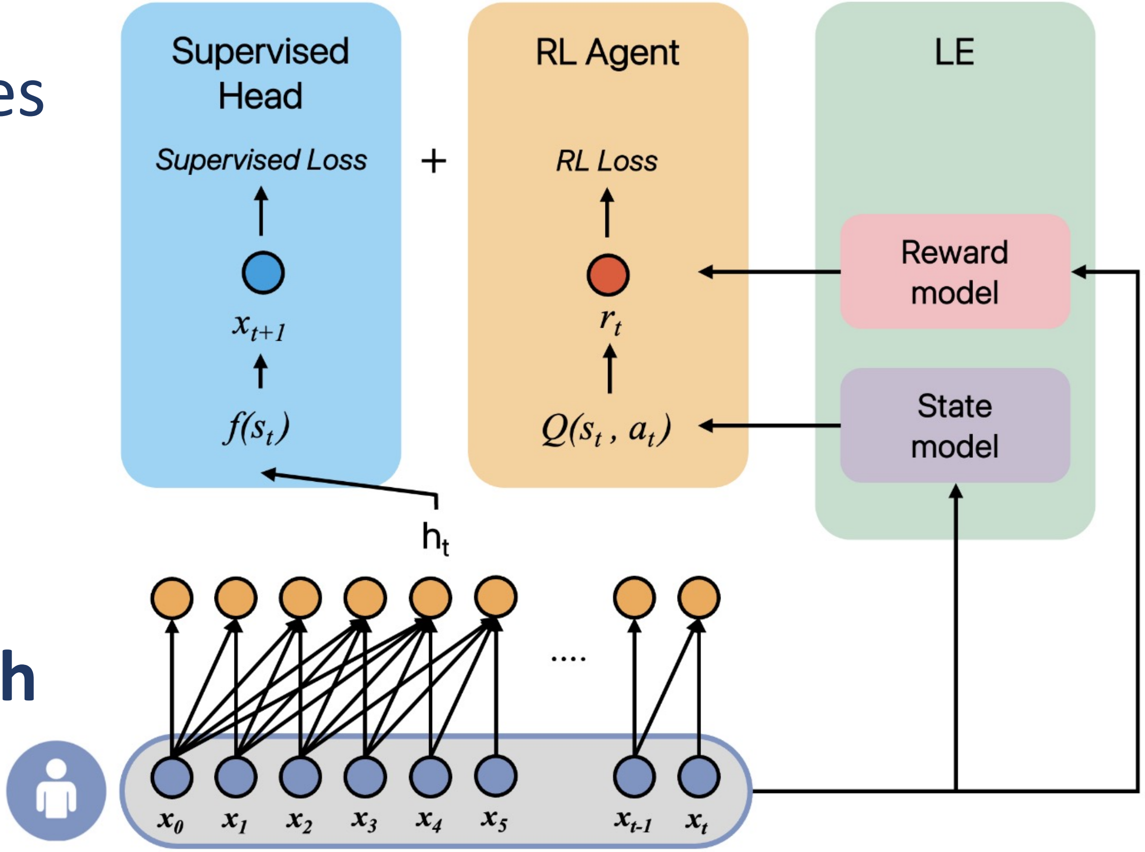
(b) Agent-simulated distribution

Zhang An et al. "On Generative Agents in Recommendation" arXiv 2023.

Agent: Environment Simulation

- ❑ **LLM as environment simulator**
  - ❑ Act as a state model that produces high quality states
  - ❑ Function as a reward model to simulate user feedback on actions

- ❑ **Application: interactive training with RL-based recommender models**



Jie Wang et al. "Reinforcement Learning-based Recommender Systems with Large Language Models for State Reward and Action Modeling" arXiv 2024.

Agent for Recommendation

❑ LLM-empowered Agents for Recommendation

❑ Agent as User Simulator

- Main idea: using agents to simulate user behavior for real-world recommendation.
- RecAgent[1], Agent4Rec[2]

❑ Agent for Recommendation

- Main idea: harnessing the powerful capabilities of LLMs, such as reasoning, reflection, planning and tool usage, for recommendation.
- RecMind[3], InteRecAgent[4], BiLLP[5], Multi-Agent Collaboration[6]

[1] Lei Wang et al. "When Large Language Model based Agent Meets User Behavior Analysis: A Novel User Simulation Paradigm" arXiv 2023.

[2] Zhang An et al. "On Generative Agents in Recommendation" arXiv 2023.

[3] Wang Yancheng et al. "RecMind: Large Language Model Powered Agent For Recommendation" arXiv 2023.

[4] Xu Huang et al. "Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations" arxiv 2023.

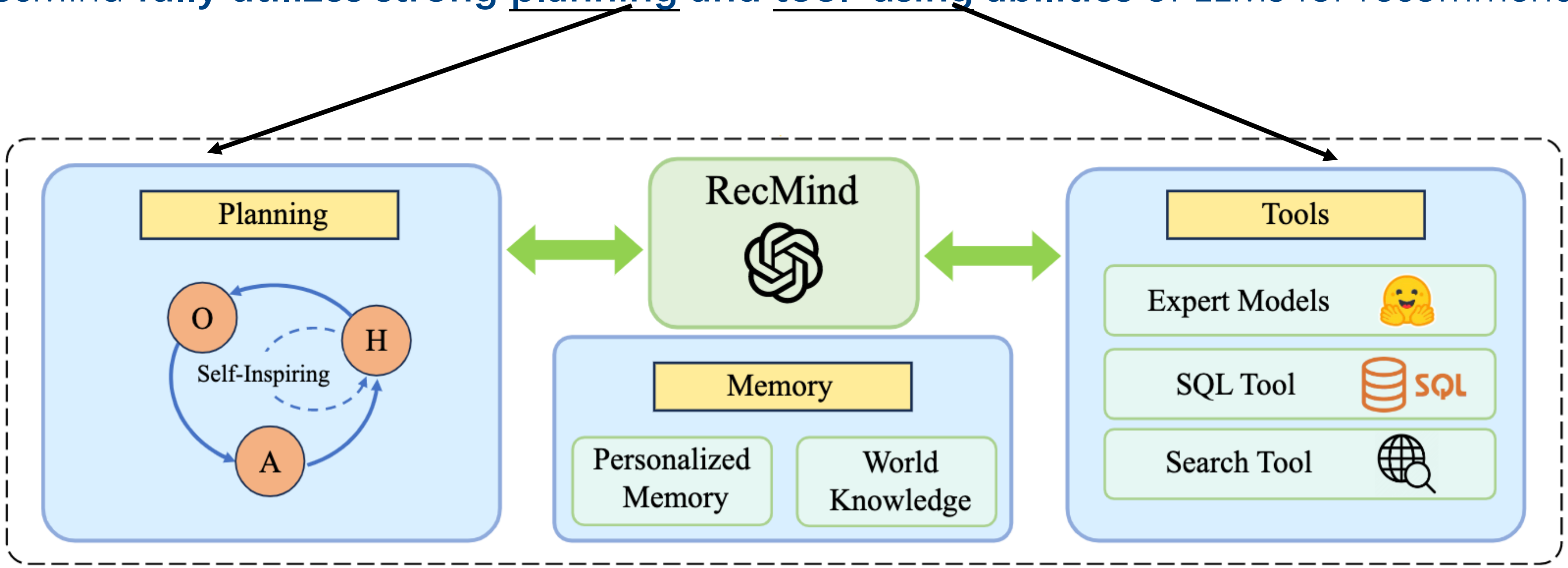
[5] Wentao Shi et al. 2023. Large Language Models are Learnable Planners for Long-Term Recommendation. in SIGIR 2024.

[6] Jiabao Fang et al. A Multi-Agent Conversational Recommender System. Arxiv 2024

Agent: RecMind

❑ LLM-based agent for recommendation

- ❑ **Traditional methods** train and fine-tune models on **task-specific** datasets, **struggle to** leverage **external knowledge** and **lack generalizability across tasks and domains.**
- ❑ Existing LLM4Rec methods primarily rely on **internal knowledge in LLM weights**.
- ❑ RecMind **fully utilizes strong planning and tool-using abilities** of LLMs for recommendation.

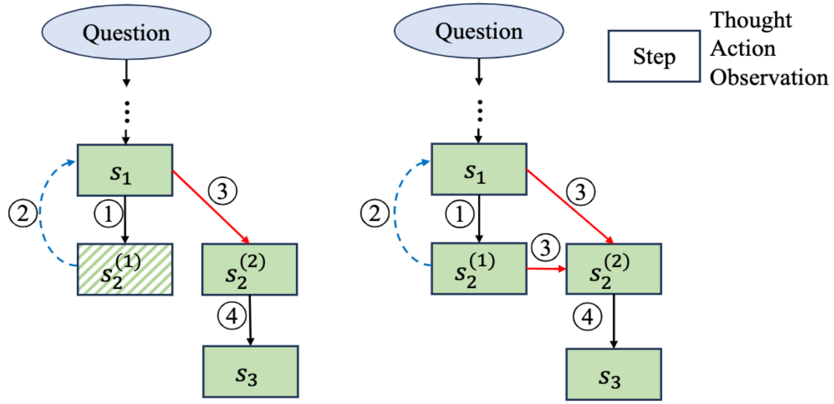


Yancheng Wang et al. "RecMind: Large Language Model Powered Agent For Recommendation" arXiv 2023.

Agent: RecMind

❑ Planning ability

- ❑ To break complex tasks into smaller sub-tasks.
- ❑ **Self-inspiring** to integrates multiple reasoning paths.



(a) Tree-of-Thoughts (DFS) (b) Self-Inspiring

❑ Tool-using ability

- ❑ **Database tool** to access domain-specific knowledge.
- ❑ **Search tool** to access real-time information.
- ❑ **Text summarization** tool to summarize lengthy texts.

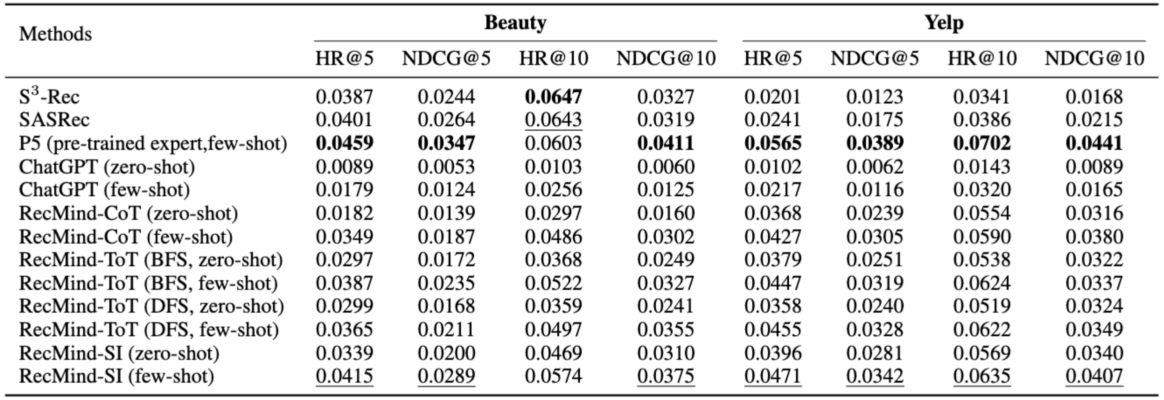
❑ Evaluation

- **Precision-oriented tasks** (rating prediction, direct recommendation, and sequential recommendation).
- **Explainability-oriented tasks** (explanation generation and review summarization).

❑ Result

RecMind can achieve performance comparable to the **fully trained P5** model.

Table 3: Performance comparison in sequential recommendation on Amazon Reviews (Beauty) and Yelp.

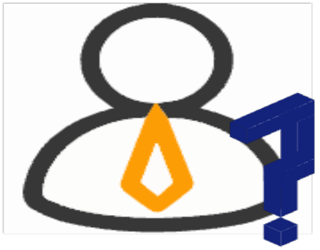


Methods	**Beauty**				**Yelp**			
	HR@5	NDCG@5	HR@10	NDCG@10	HR@5	NDCG@5	HR@10	NDCG@10
S^3-Rec	0.0387	0.0244	**0.0647**	0.0327	0.0201	0.0123	0.0341	0.0168
SASRec	0.0401	0.0264	0.0643	0.0319	0.0241	0.0175	0.0386	0.0215
P5 (pre-trained expert,few-shot)	**0.0459**	**0.0347**	0.0603	**0.0411**	**0.0565**	**0.0389**	**0.0702**	**0.0441**
ChatGPT (zero-shot)	0.0089	0.0053	0.0103	0.0060	0.0102	0.0062	0.0143	0.0089
ChatGPT (few-shot)	0.0179	0.0124	0.0256	0.0125	0.0217	0.0116	0.0320	0.0165
RecMind-CoT (zero-shot)	0.0182	0.0139	0.0297	0.0160	0.0368	0.0239	0.0554	0.0316
RecMind-CoT (few-shot)	0.0349	0.0187	0.0486	0.0302	0.0427	0.0305	0.0590	0.0380
RecMind-ToT (BFS, zero-shot)	0.0297	0.0172	0.0368	0.0249	0.0379	0.0251	0.0538	0.0322
RecMind-ToT (BFS, few-shot)	0.0387	0.0235	0.0522	0.0327	0.0447	0.0319	0.0624	0.0337
RecMind-ToT (DFS, zero-shot)	0.0299	0.0168	0.0359	0.0241	0.0358	0.0240	0.0519	0.0324
RecMind-ToT (DFS, few-shot)	0.0365	0.0211	0.0497	0.0355	0.0455	0.0328	0.0622	0.0349
RecMind-SI (zero-shot)	0.0339	0.0200	0.0469	0.0310	0.0396	0.0281	0.0569	0.0340
RecMind-SI (few-shot)	0.0415	0.0289	0.0574	0.0375	0.0471	0.0342	0.0635	0.0407

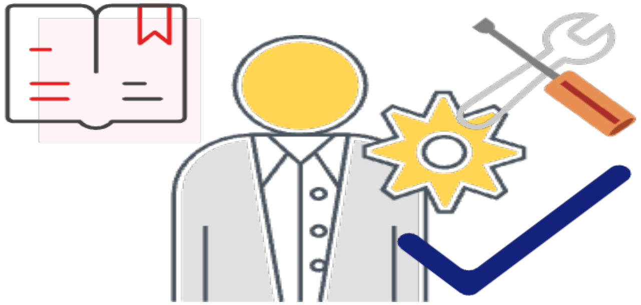
Yancheng Wang et al. "RecMind: Large Language Model Powered Agent For Recommendation" arXiv 2023.

Agent: ToolRec

- ❖ Traditional challenges: conventional recommender systems (CRS) lack commonsense knowledge about users and items, "narrow expert"
- ❖ LLM Advantages: LLMs excel in commonsense reasoning and leveraging external tools



CRS



LLMs with RSs

Current LLMs with RSs Limitations:

- Hallucinations ···
- Misalignment between language tasks and recommendation tasks ···

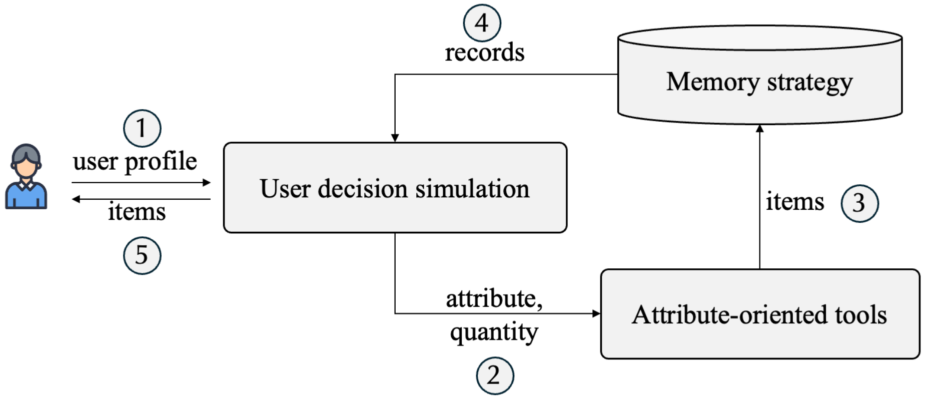
- ❖ Our Key: Use LLMs to understand current contexts and preferences, and apply attribute-oriented tools to find suitable items

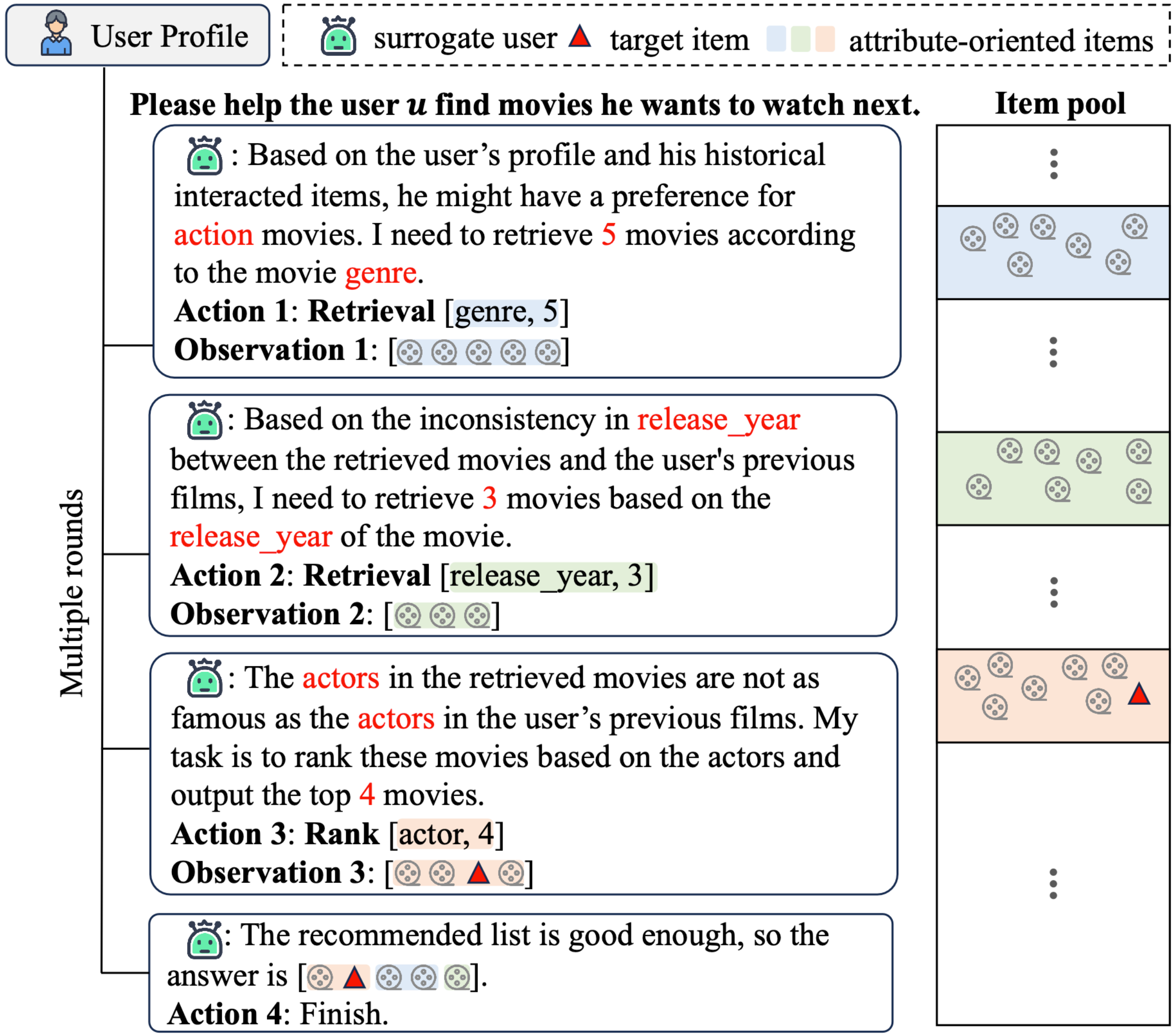
Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning. In SIGIR 2024

Agent: ToolRec

Methodology:

- LLMs as the central controller, simulating the user decision.
- Attribute-oriented Tools: rank tools and retrieval tools.
- Memory strategy can ensure the correctness of generated items and cataloging candidate items.





Please help the user *u* find movies he wants to watch next.

: Based on the user’s profile and his historical interacted items, he might have a preference for action movies. I need to retrieve 5 movies according to the movie genre.
Action 1: **Retrieval** [genre, 5]
Observation 1: [◎ ◎ ◎ ◎ ◎]

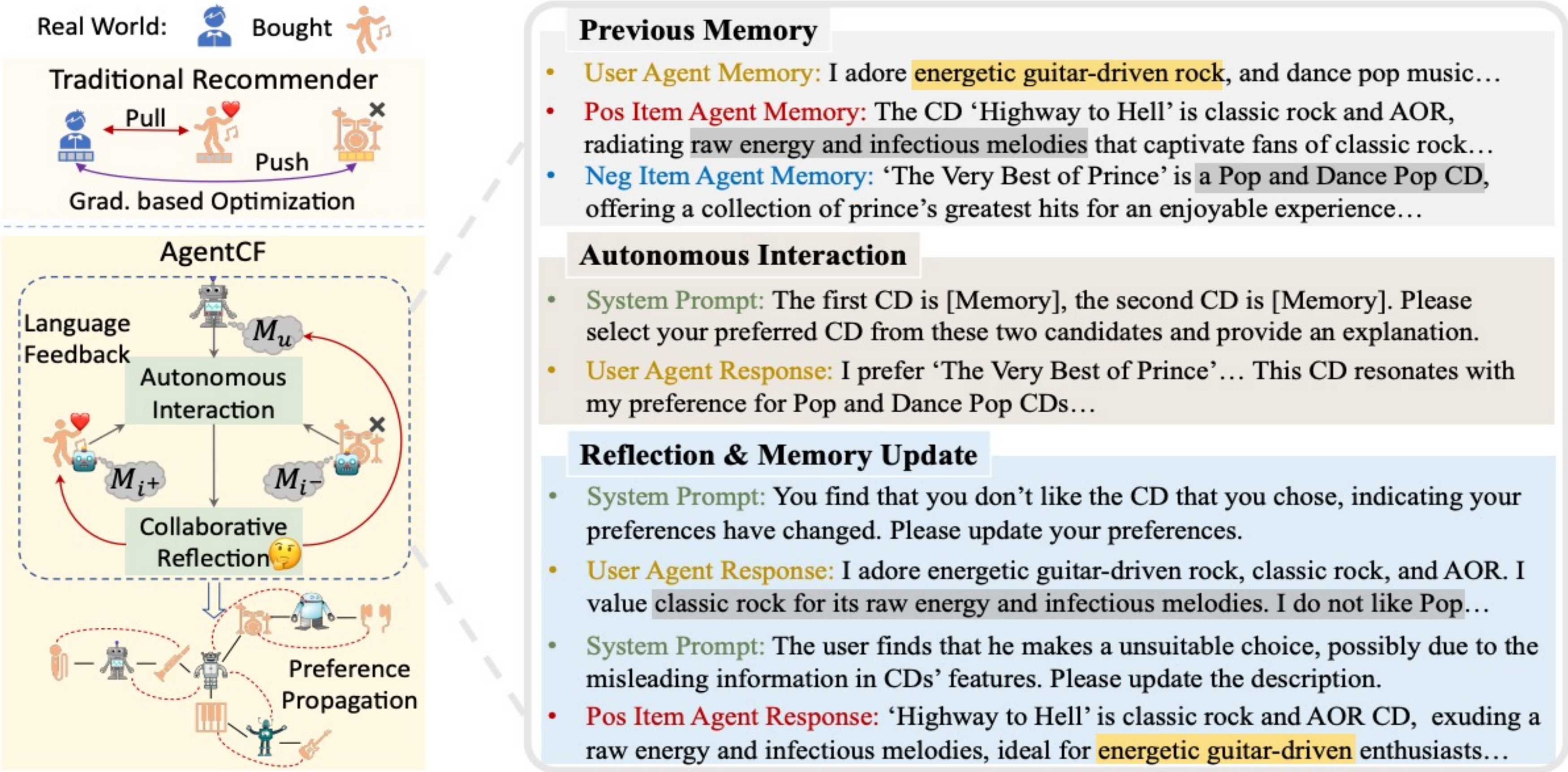
: Based on the inconsistency in release_year between the retrieved movies and the user's previous films, I need to retrieve 3 movies based on the release_year of the movie.
Action 2: **Retrieval** [release_year, 3]
Observation 2: [◎ ◎ ◎]

: The actors in the retrieved movies are not as famous as the actors in the user’s previous films. My task is to rank these movies based on the actors and output the top 4 movies.
Action 3: **Rank** [actor, 4]
Observation 3: [◎ ◎ ▲ ◎]

: The recommended list is good enough, so the answer is [◎ ▲ ◎ ◎ ◎].
Action 4: Finish.

Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning. In SIGIR 2024

Agent: AgentCF



Previous Memory

- User Agent Memory: I adore energetic guitar-driven rock, and dance pop music…
- Pos Item Agent Memory: The CD 'Highway to Hell' is classic rock and AOR, radiating raw energy and infectious melodies that captivate fans of classic rock…
- Neg Item Agent Memory: 'The Very Best of Prince' is a Pop and Dance Pop CD, offering a collection of prince's greatest hits for an enjoyable experience…

Autonomous Interaction

- System Prompt: The first CD is [Memory], the second CD is [Memory]. Please select your preferred CD from these two candidates and provide an explanation.
- User Agent Response: I prefer 'The Very Best of Prince'… This CD resonates with my preference for Pop and Dance Pop CDs…

Reflection & Memory Update

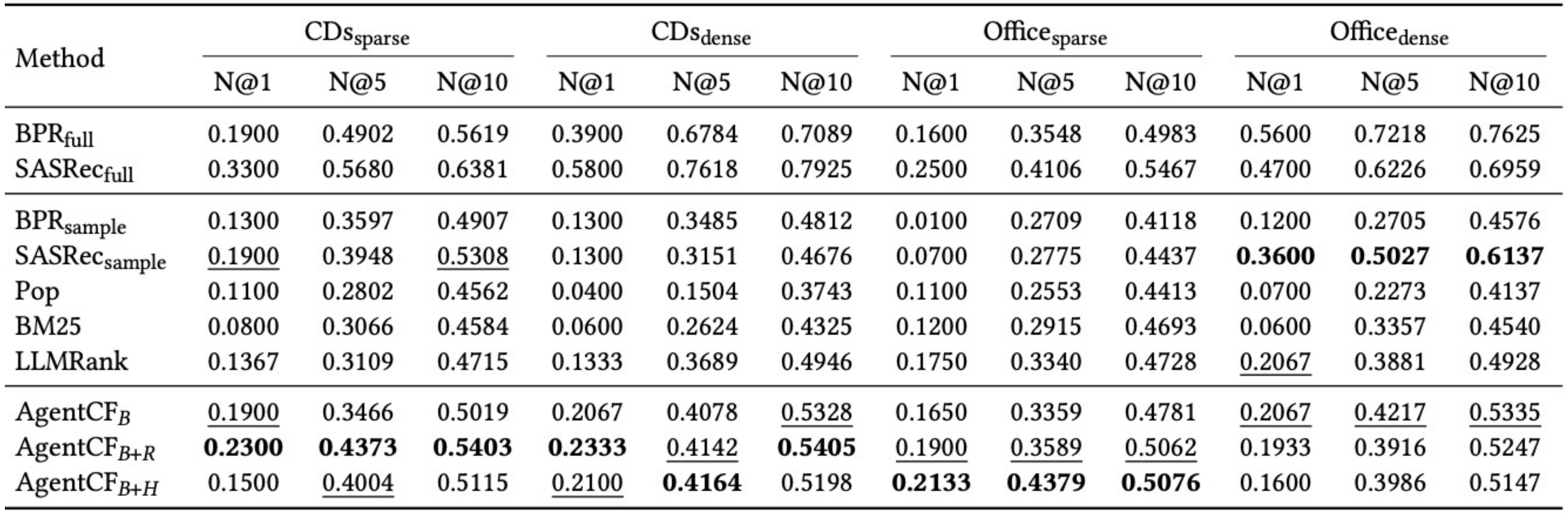
- System Prompt: You find that you don't like the CD that you chose, indicating your preferences have changed. Please update your preferences.
- User Agent Response: I adore energetic guitar-driven rock, classic rock, and AOR. I value classic rock for its raw energy and infectious melodies. I do not like Pop…
- System Prompt: The user finds that he makes a unsuitable choice, possibly due to the misleading information in CDs' features. Please update the description.
- Pos Item Agent Response: 'Highway to Hell' is classic rock and AOR CD, exuding a raw energy and infectious melodies, ideal for energetic guitar-driven enthusiasts…

❑ Use Agent to simulate both user/items

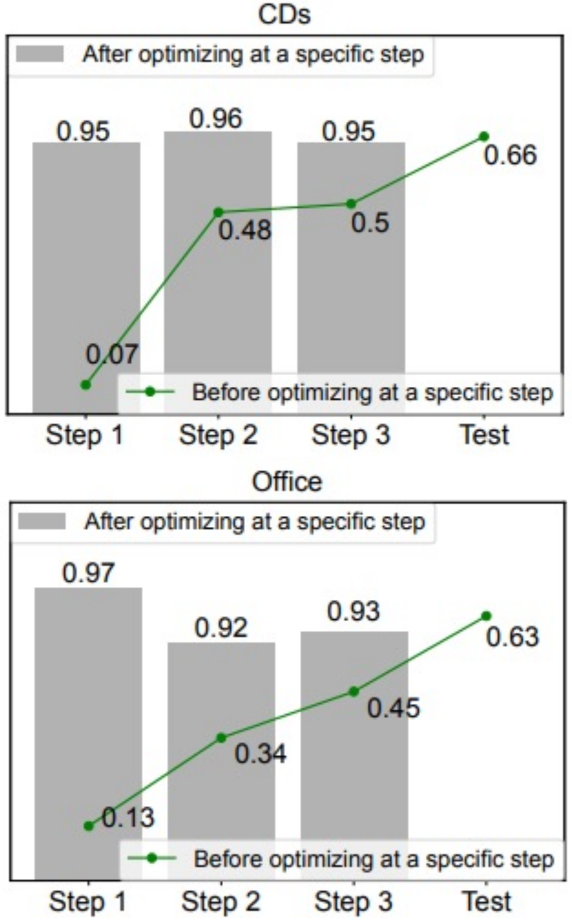
❑ Provide a collaborative reflection optimizing mechanism to optimize the user/item agents, and mutual update of user and item memory.

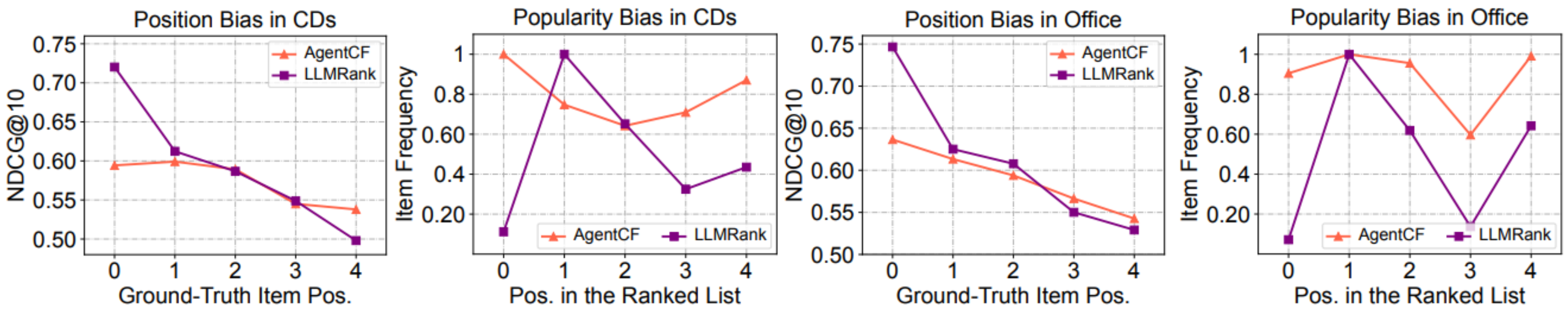
Junjie Zhang et al. 2023. Agentcf: Collaborative learning with autonomous language agents for recommender systems. in WWW 2024

Agent: AgentCF



Method	CDs$_{sparse}$			CDs$_{dense}$			Office$_{sparse}$			Office$_{dense}$		
	N@1	N@5	N@10	N@1	N@5	N@10	N@1	N@5	N@10	N@1	N@5	N@10
BPR$_{full}$	0.1900	0.4902	0.5619	0.3900	0.6784	0.7089	0.1600	0.3548	0.4983	0.5600	0.7218	0.7625
SASRec$_{full}$	0.3300	0.5680	0.6381	0.5800	0.7618	0.7925	0.2500	0.4106	0.5467	0.4700	0.6226	0.6959
BPR$_{sample}$	0.1300	0.3597	0.4907	0.1300	0.3485	0.4812	0.0100	0.2709	0.4118	0.1200	0.2705	0.4576
SASRec$_{sample}$	0.1900	0.3948	0.5308	0.1300	0.3151	0.4676	0.0700	0.2775	0.4437	**0.3600**	**0.5027**	**0.6137**
Pop	0.1100	0.2802	0.4562	0.0400	0.1504	0.3743	0.1100	0.2553	0.4413	0.0700	0.2273	0.4137
BM25	0.0800	0.3066	0.4584	0.0600	0.2624	0.4325	0.1200	0.2915	0.4693	0.0600	0.3357	0.4540
LLMRank	0.1367	0.3109	0.4715	0.1333	0.3689	0.4946	0.1750	0.3340	0.4728	0.2067	0.3881	0.4928
AgentCF$_{B}$	0.1900	0.3466	0.5019	0.2067	0.4078	0.5328	0.1650	0.3359	0.4781	0.2067	0.4217	0.5335
AgentCF$_{B+R}$	**0.2300**	**0.4373**	**0.5403**	**0.2333**	0.4142	**0.5405**	0.1900	0.3589	0.5062	0.1933	0.3916	0.5247
AgentCF$_{B+H}$	0.1500	0.4004	0.5115	0.2100	**0.4164**	0.5198	**0.2133**	**0.4379**	**0.5076**	0.1600	0.3986	0.5147





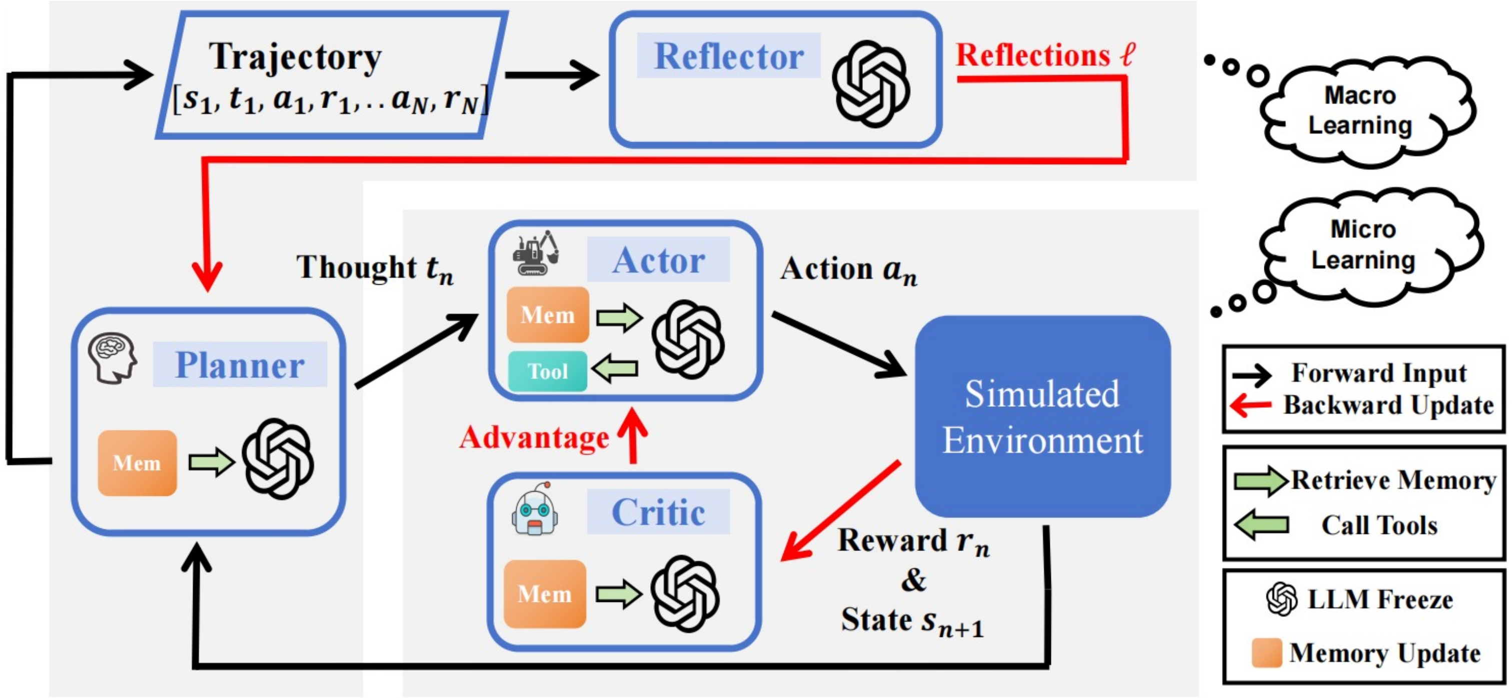
❑ Better performance and less influenced by bias than directly instructing LLM to rerank

❑ Collaborative Reflection is effective to optimize the agent’s ability to distinguish positive/negative items

Junjie Zhang et al. 2023. Agentcf: Collaborative learning with autonomous language agents for recommender systems. in WWW 2024

Agent: BiLLP

- ❑ Use LLM to make plans for long-term recommendations
- ❑ Utilize a bi-level learnable mechanism to learn macro-level guidance and micro-level personalized recommendation policies.



Wentao Shi et al. 2023. Large Language Models are Learnable Planners for Long-Term Recommendation. in SIGIR 2024

Agent: BiLLP

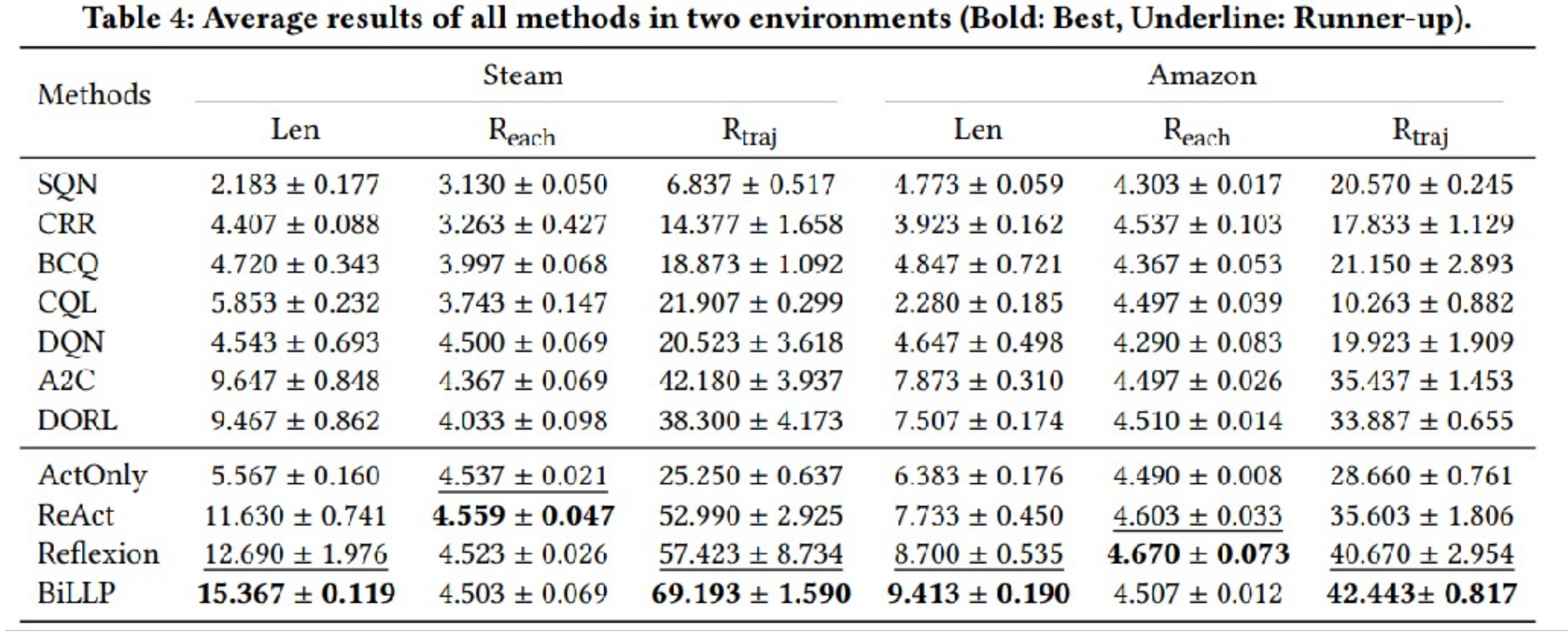
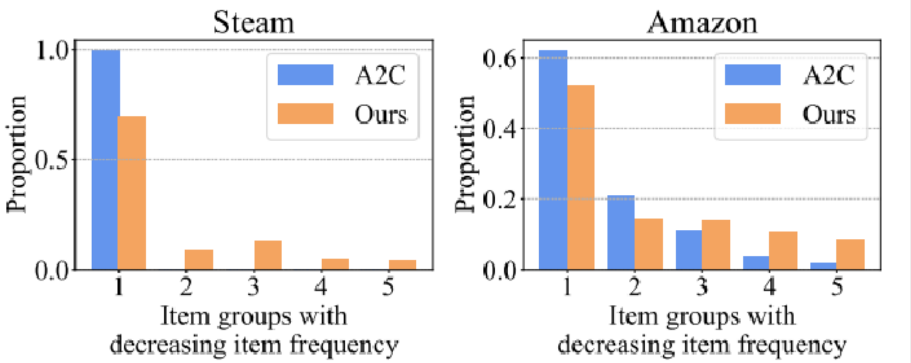


Table 4: Average results of all methods in two environments (Bold: Best, Underline: Runner-up).

Methods	Steam			Amazon		
	Len	R_{each}	R_{traj}	Len	R_{each}	R_{traj}
SQN	2.183 ± 0.177	3.130 ± 0.050	6.837 ± 0.517	4.773 ± 0.059	4.303 ± 0.017	20.570 ± 0.245
CRR	4.407 ± 0.088	3.263 ± 0.427	14.377 ± 1.658	3.923 ± 0.162	4.537 ± 0.103	17.833 ± 1.129
BCQ	4.720 ± 0.343	3.997 ± 0.068	18.873 ± 1.092	4.847 ± 0.721	4.367 ± 0.053	21.150 ± 2.893
CQL	5.853 ± 0.232	3.743 ± 0.147	21.907 ± 0.299	2.280 ± 0.185	4.497 ± 0.039	10.263 ± 0.882
DQN	4.543 ± 0.693	4.500 ± 0.069	20.523 ± 3.618	4.647 ± 0.498	4.290 ± 0.083	19.923 ± 1.909
A2C	9.647 ± 0.848	4.367 ± 0.069	42.180 ± 3.937	7.873 ± 0.310	4.497 ± 0.026	35.437 ± 1.453
DORL	9.467 ± 0.862	4.033 ± 0.098	38.300 ± 4.173	7.507 ± 0.174	4.510 ± 0.014	33.887 ± 0.655
ActOnly	5.567 ± 0.160	4.537 ± 0.021	25.250 ± 0.637	6.383 ± 0.176	4.490 ± 0.008	28.660 ± 0.761
ReAct	11.630 ± 0.741	**4.559 ± 0.047**	52.990 ± 2.925	7.733 ± 0.450	4.603 ± 0.033	35.603 ± 1.806
Reflexion	12.690 ± 1.976	4.523 ± 0.026	57.423 ± 8.734	8.700 ± 0.535	**4.670 ± 0.073**	40.670 ± 2.954
BiLLP	**15.367 ± 0.119**	4.503 ± 0.069	**69.193 ± 1.590**	**9.413 ± 0.190**	4.507 ± 0.012	**42.443± 0.817**



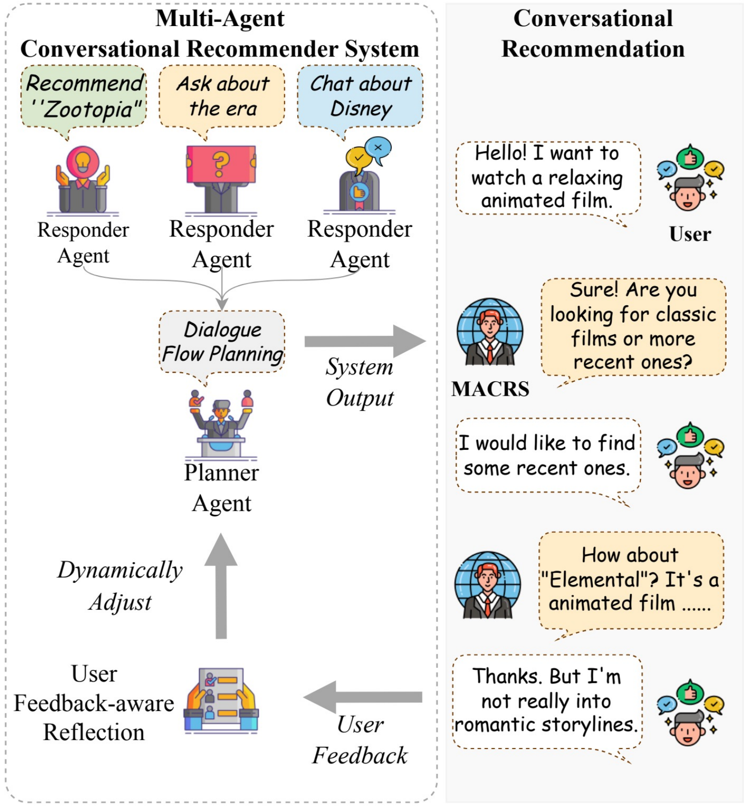
- ❑ Better long-term performance than traditional RL-based methods
- ❑ Better planning capabilities on long-tail items.

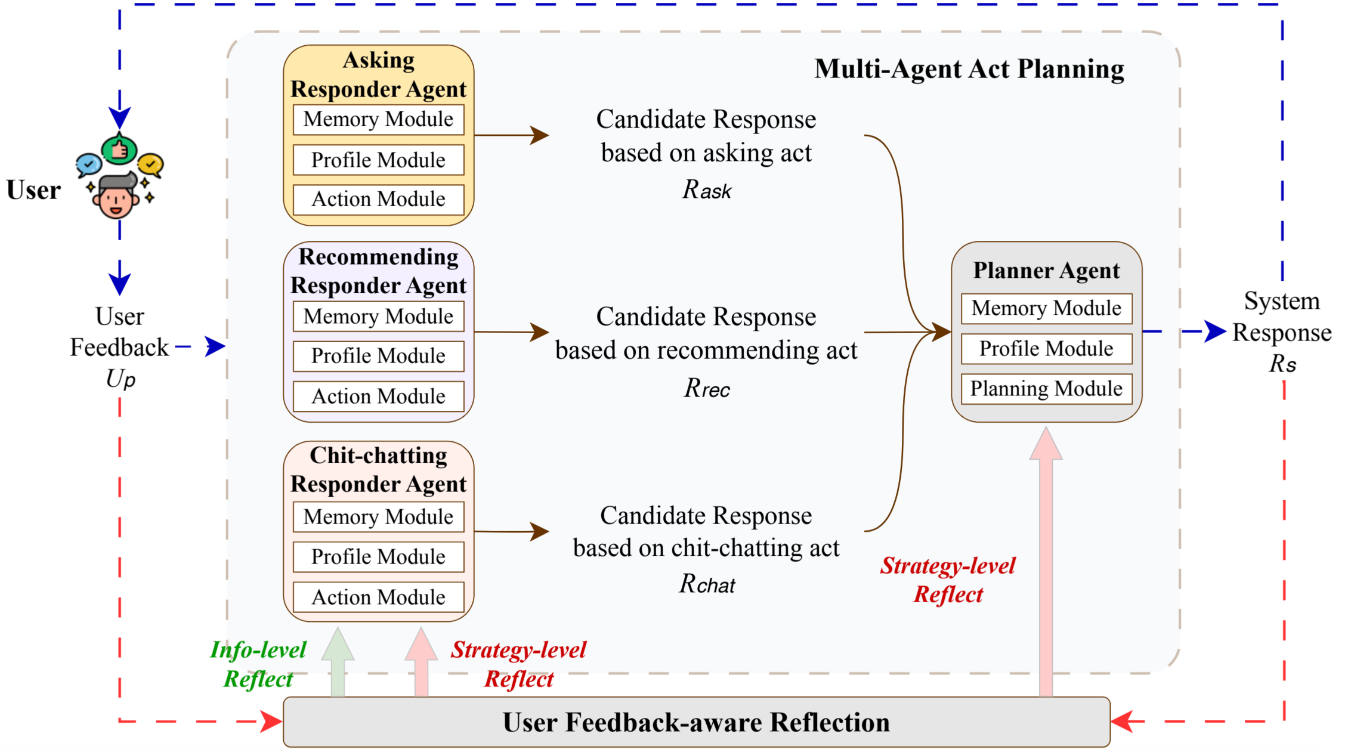
Wentao Shi et al. 2023. Large Language Models are Learnable Planners for Long-Term Recommendation. in SIGIR 2024

Agent: Multi-Agent Conversational Rec

❑ Different Agents Collaborate together for Conversational Recommendation

- The responder agent and planner agent collaboratively generate appropriate responses, while the reflection mechanism provides feedback and refined guidance to these agents

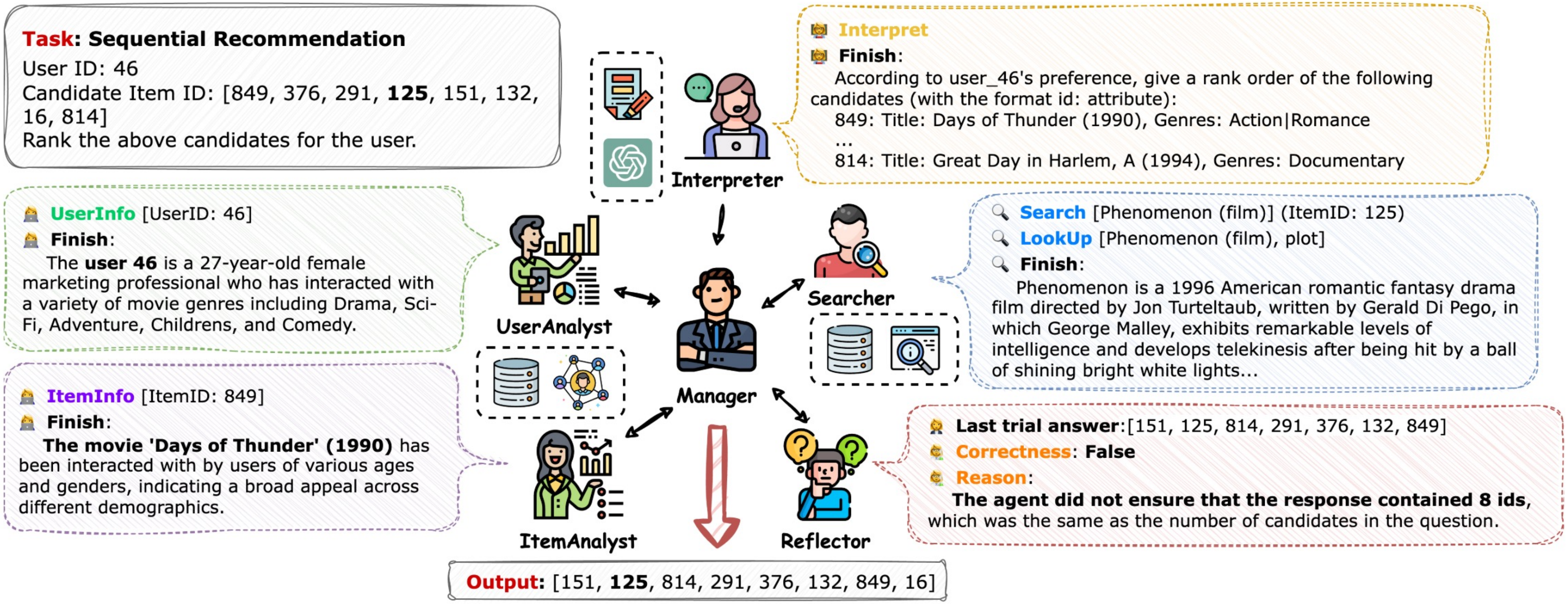

Multi-Agent
Conversational Recommender System
Recommend ''Zootopia''
Ask about the era
Chat about Disney
Responder Agent
Responder Agent
Responder Agent
Dialogue Flow Planning
System Output
Planner Agent
Dynamically Adjust
User Feedback-aware Reflection
User Feedback
Conversational Recommendation
Hello! I want to watch a relaxing animated film.
User
Sure! Are you looking for classic films or more recent ones?
MACRS
I would like to find some recent ones.
How about "Elemental"? It's a animated film ......
Thanks. But I'm not really into romantic storylines.



Jiabao Fang et al. A Multi-Agent Conversational Recommender System. Arxiv 2024

Agent: Multi-Agent Collaboration for Rec

- Different agents can collaborate together for information delivery.



Task: Sequential Recommendation
User ID: 46
Candidate Item ID: [849, 376, 291, **125**, 151, 132, 16, 814]
Rank the above candidates for the user.

Interpret
Finish:
According to user_46's preference, give a rank order of the following candidates (with the format id: attribute):
849: Title: Days of Thunder (1990), Genres: Action|Romance
...
814: Title: Great Day in Harlem, A (1994), Genres: Documentary

UserInfo [UserID: 46]
Finish:
The **user 46** is a 27-year-old female marketing professional who has interacted with a variety of movie genres including Drama, Sci-Fi, Adventure, Childrens, and Comedy.

Search [Phenomenon (film)] (ItemID: 125)
LookUp [Phenomenon (film), plot]
Finish:
Phenomenon is a 1996 American romantic fantasy drama film directed by Jon Turteltaub, written by Gerald Di Pego, in which George Malley, exhibits remarkable levels of intelligence and develops telekinesis after being hit by a ball of shining bright white lights...

ItemInfo [ItemID: 849]
Finish:
The movie 'Days of Thunder' (1990) has been interacted with by users of various ages and genders, indicating a broad appeal across different demographics.

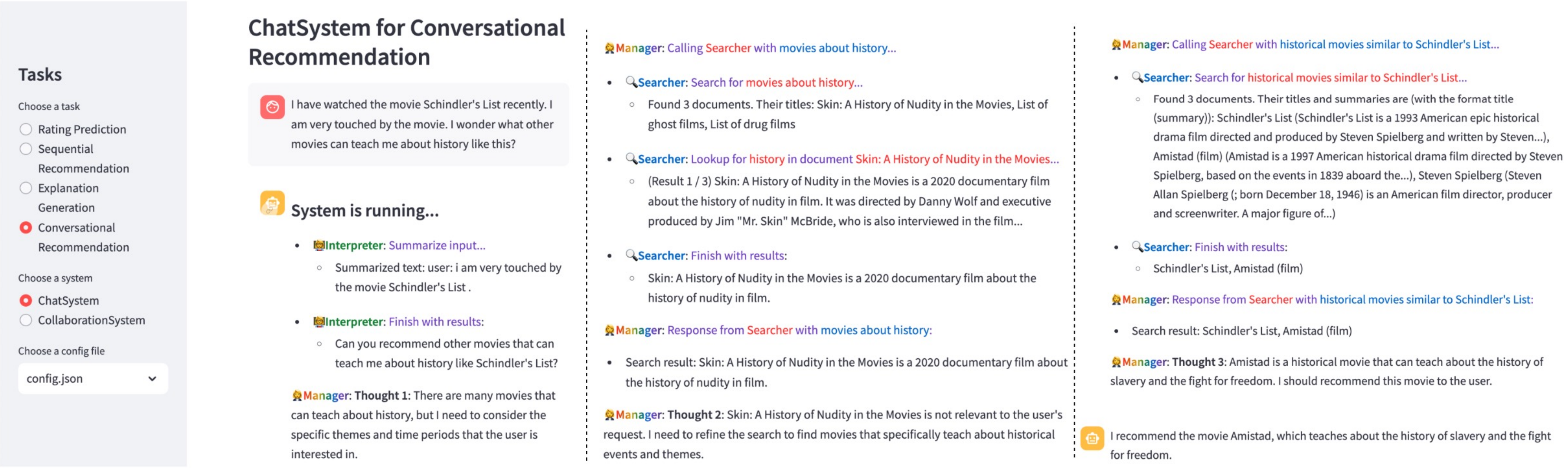
Last trial answer:[151, 125, 814, 291, 376, 132, 849]
Correctness: **False**
Reason:
The agent did not ensure that the response contained 8 ids, which was the same as the number of candidates in the question.

Output: [151, **125**, 814, 291, 376, 132, 849, 16]

Zhefan Wang et al. Multi-Agent Collaboration Framework for Recommender Systems. SIGIR 2024

Agent: Multi-Agent Collaboration for Rec

- ❑ **Collaborate Together can Solve Different Task**
  - ❑ **A demo for conversational recommendation via multi-agent collaboration**



Tasks

Choose a task

○ Rating Prediction
○ Sequential Recommendation
○ Explanation Generation
● Conversational Recommendation

Choose a system

● ChatSystem
○ CollaborationSystem

Choose a config file

config.json

ChatSystem for Conversational Recommendation

I have watched the movie Schindler's List recently. I am very touched by the movie. I wonder what other movies can teach me about history like this?

System is running...

- Interpreter: Summarize input...
  - Summarized text: user: i am very touched by the movie Schindler's List .
- Interpreter: Finish with results:
  - Can you recommend other movies that can teach me about history like Schindler's List?

Manager: **Thought 1**: There are many movies that can teach about history, but I need to consider the specific themes and time periods that the user is interested in.

Manager: Calling Searcher with movies about history...

- Searcher: Search for movies about history...
  - Found 3 documents. Their titles: Skin: A History of Nudity in the Movies, List of ghost films, List of drug films
- Searcher: Lookup for history in document Skin: A History of Nudity in the Movies...
  - (Result 1 / 3) Skin: A History of Nudity in the Movies is a 2020 documentary film about the history of nudity in film. It was directed by Danny Wolf and executive produced by Jim "Mr. Skin" McBride, who is also interviewed in the film...
- Searcher: Finish with results:
  - Skin: A History of Nudity in the Movies is a 2020 documentary film about the history of nudity in film.

Manager: Response from Searcher with movies about history:

- Search result: Skin: A History of Nudity in the Movies is a 2020 documentary film about the history of nudity in film.

Manager: **Thought 2**: Skin: A History of Nudity in the Movies is not relevant to the user's request. I need to refine the search to find movies that specifically teach about historical events and themes.

Manager: Calling Searcher with historical movies similar to Schindler's List...

- Searcher: Search for historical movies similar to Schindler's List...
  - Found 3 documents. Their titles and summaries are (with the format title (summary)): Schindler's List (Schindler's List is a 1993 American epic historical drama film directed and produced by Steven Spielberg and written by Steven...), Amistad (film) (Amistad is a 1997 American historical drama film directed by Steven Spielberg, based on the events in 1839 aboard the...), Steven Spielberg (Steven Allan Spielberg (; born December 18, 1946) is an American film director, producer and screenwriter. A major figure of...)
- Searcher: Finish with results:
  - Schindler's List, Amistad (film)

Manager: Response from Searcher with historical movies similar to Schindler's List:

- Search result: Schindler's List, Amistad (film)

Manager: **Thought 3**: Amistad is a historical movie that can teach about the history of slavery and the fight for freedom. I should recommend this movie to the user.

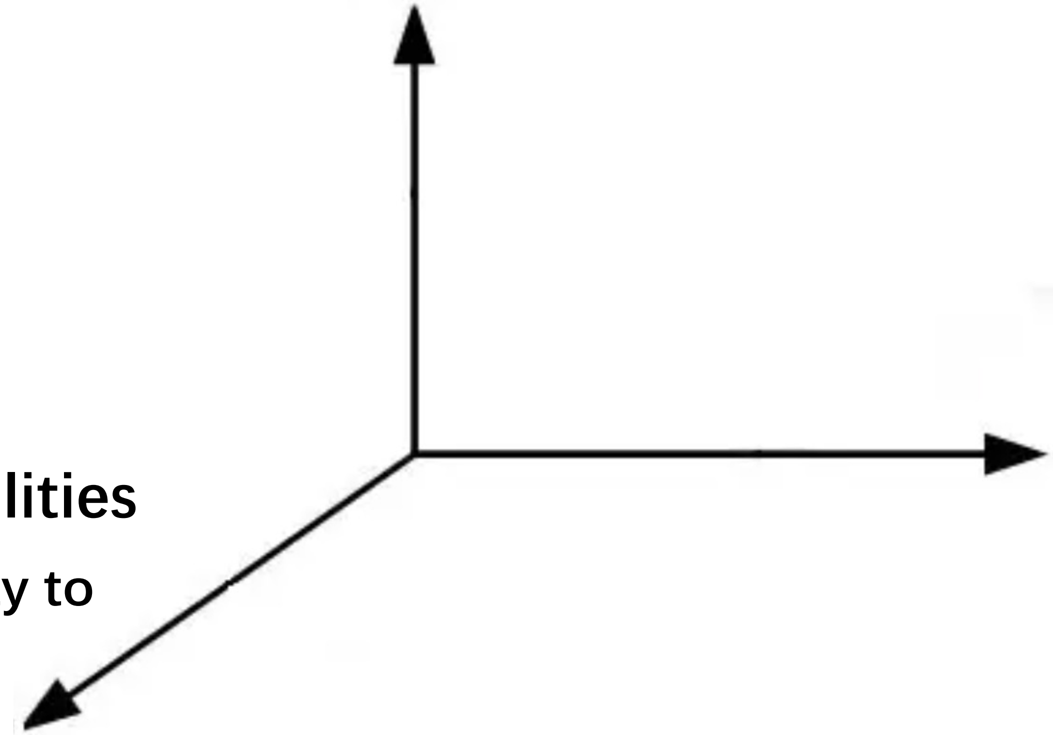
I recommend the movie Amistad, which teaches about the history of slavery and the fight for freedom.

(a) Interprete the dialog into a task. (b) Search for movies about history. (c) Search for movies similar to Schindler's List.

Zhefan Wang et al. Multi-Agent Collaboration Framework for Recommender Systems. SIGIR 2024

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness

such as privacy, fairness, etc.

How to utilize LLMs

In-context learning

Tuning

Agent

Information modalities

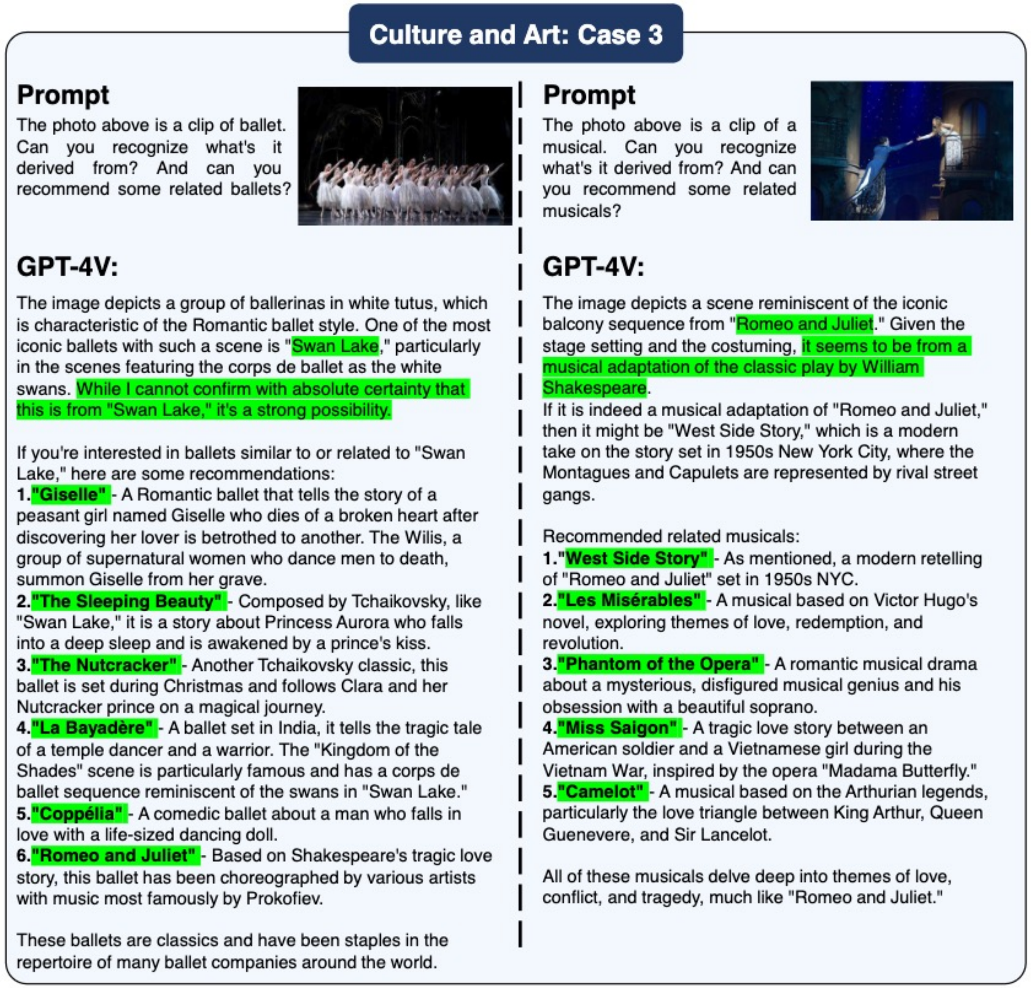
From the text modality to

multiple modalities

Progress of LLM4Rec

❑ Multimodal LLM4Rec

- **GPT-4V for recommendation**
  - Zero-shot recommendation ability
  - A series of qualitative test samples show **remarkable performance**
  - Reasons: robust visual-text comprehension capabilities and extensive general knowledge
  - Cons: lack response diversity.
- **Future directions:** leverage multimodal user and item features for LLM-based recommendation
  - Tokenization with multimodal features
  - Model structure with with multimodal inputs and outputs
  - Optimization objectives to exclude multimodal noises
  - ……



Culture and Art: Case 3

Prompt

The photo above is a clip of ballet. Can you recognize what's it derived from? And can you recommend some related ballets?

GPT-4V:

The image depicts a group of ballerinas in white tutus, which is characteristic of the Romantic ballet style. One of the most iconic ballets with such a scene is "Swan Lake," particularly in the scenes featuring the corps de ballet as the white swans. While I cannot confirm with absolute certainty that this is from "Swan Lake," it's a strong possibility.

If you're interested in ballets similar to or related to "Swan Lake," here are some recommendations:
1."Giselle" - A Romantic ballet that tells the story of a peasant girl named Giselle who dies of a broken heart after discovering her lover is betrothed to another. The Wilis, a group of supernatural women who dance men to death, summon Giselle from her grave.
2."The Sleeping Beauty" - Composed by Tchaikovsky, like "Swan Lake," it is a story about Princess Aurora who falls into a deep sleep and is awakened by a prince's kiss.
3."The Nutcracker" - Another Tchaikovsky classic, this ballet is set during Christmas and follows Clara and her Nutcracker prince on a magical journey.
4."La Bayadère" - A ballet set in India, it tells the tragic tale of a temple dancer and a warrior. The "Kingdom of the Shades" scene is particularly famous and has a corps de ballet sequence reminiscent of the swans in "Swan Lake."
5."Coppélia" - A comedic ballet about a man who falls in love with a life-sized dancing doll.
6."Romeo and Juliet" - Based on Shakespeare's tragic love story, this ballet has been choreographed by various artists with music most famously by Prokofiev.

These ballets are classics and have been staples in the repertoire of many ballet companies around the world.

Prompt

The photo above is a clip of a musical. Can you recognize what's it derived from? And can you recommend some related musicals?

GPT-4V:

The image depicts a scene reminiscent of the iconic balcony sequence from "Romeo and Juliet." Given the stage setting and the costuming, it seems to be from a musical adaptation of the classic play by William Shakespeare.
If it is indeed a musical adaptation of "Romeo and Juliet," then it might be "West Side Story," which is a modern take on the story set in 1950s New York City, where the Montagues and Capulets are represented by rival street gangs.

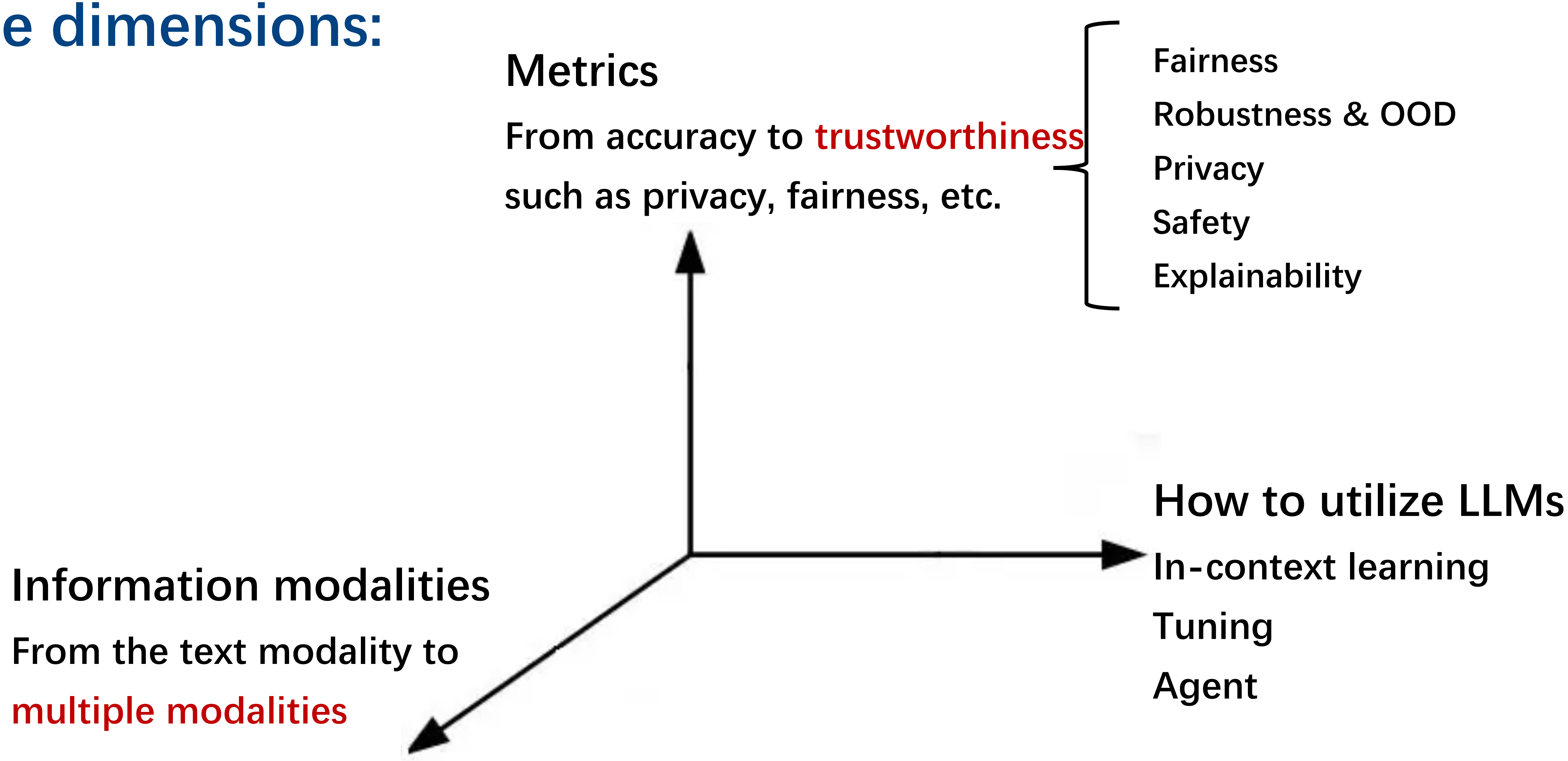
Recommended related musicals:
1."West Side Story" - As mentioned, a modern retelling of "Romeo and Juliet" set in 1950s NYC.
2."Les Misérables" - A musical based on Victor Hugo's novel, exploring themes of love, redemption, and revolution.
3."Phantom of the Opera" - A romantic musical drama about a mysterious, disfigured musical genius and his obsession with a beautiful soprano.
4."Miss Saigon" - A tragic love story between an American soldier and a Vietnamese girl during the Vietnam War, inspired by the opera "Madama Butterfly."
5."Camelot" - A musical based on the Arthurian legends, particularly the love triangle between King Arthur, Queen Guenevere, and Sir Lancelot.

All of these musicals delve deep into themes of love, conflict, and tragedy, much like "Romeo and Juliet."

Figure 3 | **Culture&Art-Case3.** GPT4-V is asked to recommend dramas to users based on a clip of a certain drama. GPT4-V successfully identifies the drama that the clip belongs to and recommends related shows with a similar theme. Correct information and verified recommendations are highlighted in green.

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

- Fairness
- Robustness & OOD
- Privacy
- Safety
- Explainability

How to utilize LLMs

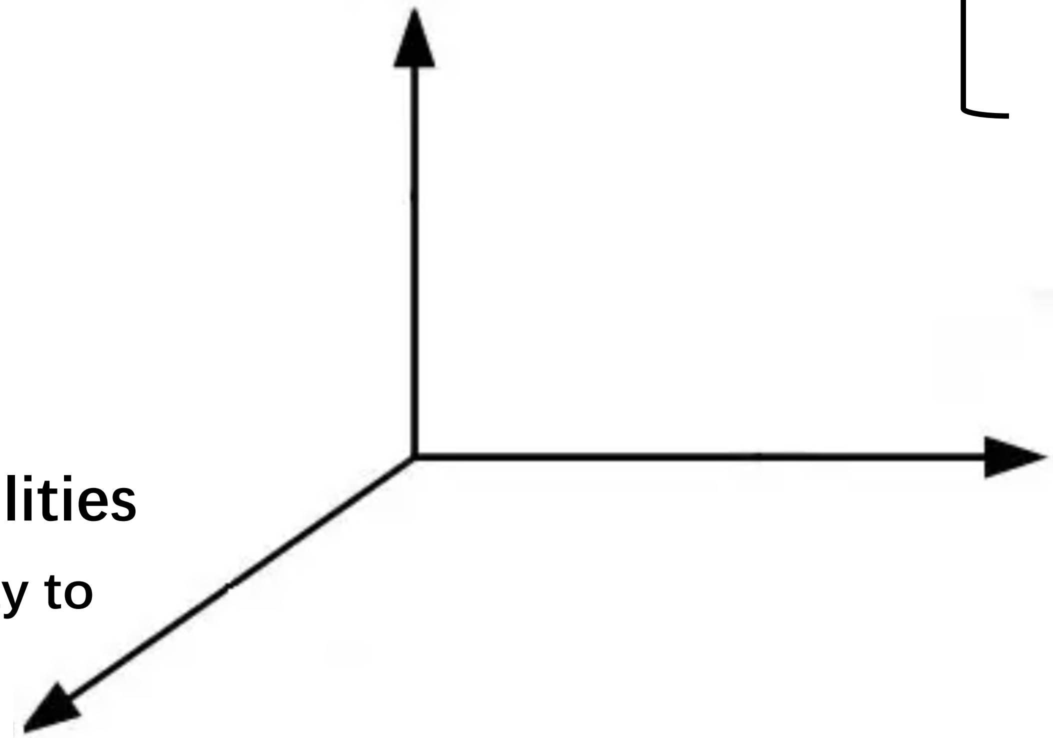
- In-context learning
- Tuning
- Agent

Information modalities

From the text modality to multiple modalities

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

- Fairness
- Robustness & OOD
- Privacy
- Safety
- Explainability

How to utilize LLMs

- In-context learning
- Tuning
- Agent

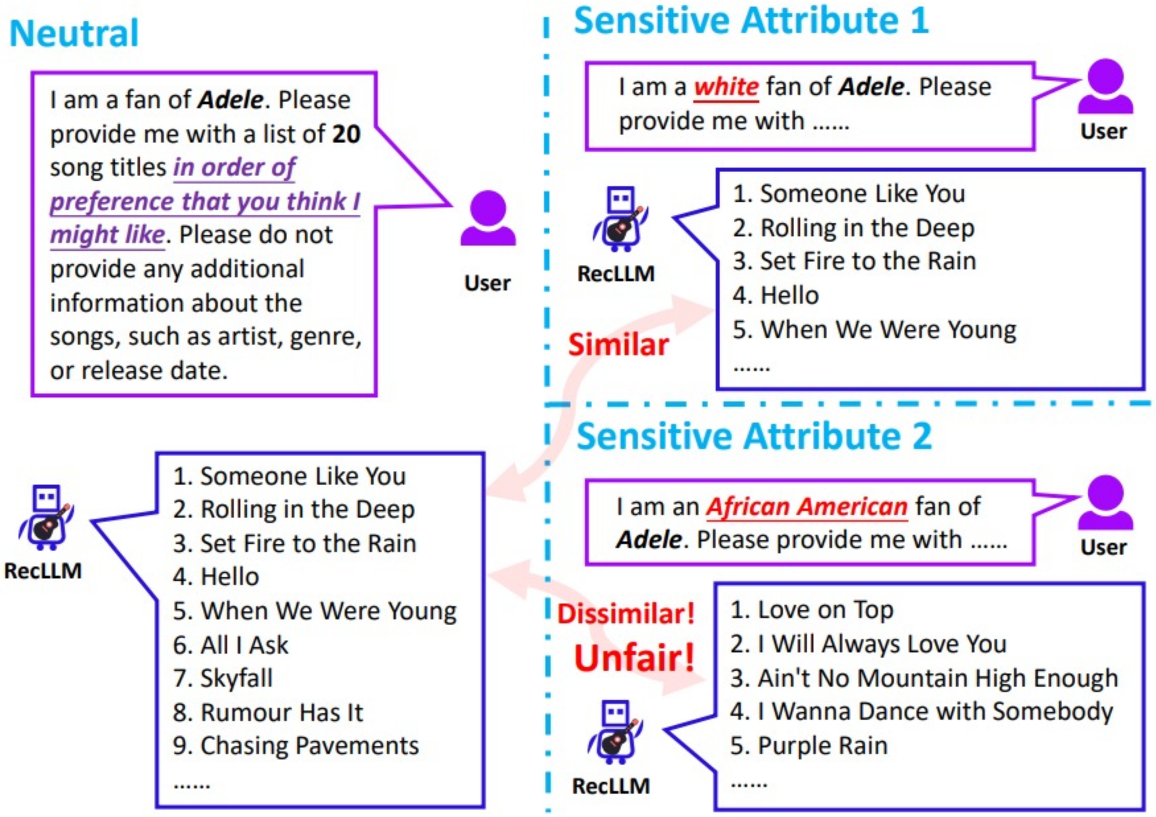
Information modalities

From the text modality to multiple modalities

User-side Fairness

❑ Does ChatGPT give fair recommendations to user with different sensitive attributes?

- □ We judge the fairness by comparing the similarity between the recommended results of different sensitive instructions and the neutral instructions.
- □ Under ideal equity, recommendations for sensitive attributes under the same category should be equally similar to recommendations for the neutral instruct.



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

……

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

……

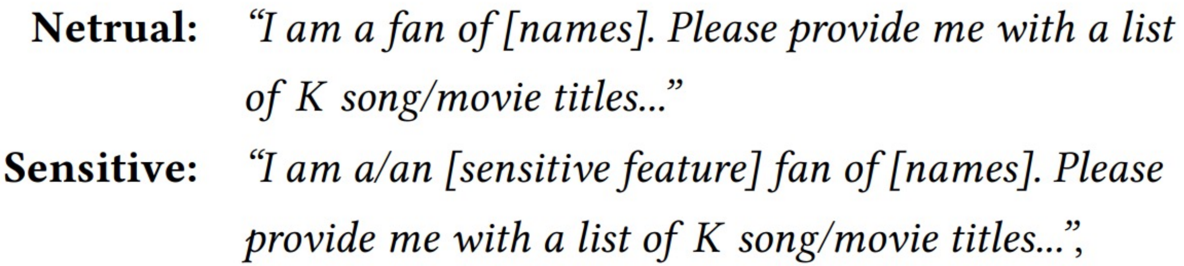
Jizhi Zhang et al. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In RecSys 2023

User-side Fairness

❑ Dataset Construction.

- ◘ Construct a dataset that accounts for eight sensitive attributes (31 sensitive attribute values) in two recommendation scenarios: music and movies to measure the fairness of LLM4Rec.

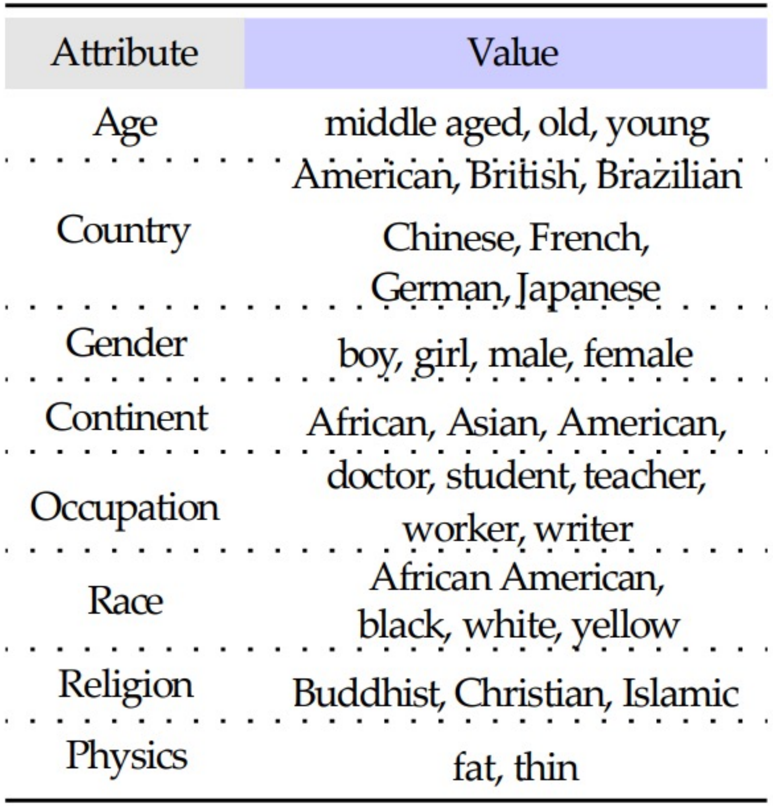
Template:



Netrual: *“I am a fan of [names]. Please provide me with a list of K song/movie titles...”*

Sensitive: *“I am a/an [sensitive feature] fan of [names]. Please provide me with a list of K song/movie titles...”,*

Sensitive attributes and their specific values:



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

Jizhi Zhang et al. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In RecSys 2023

User-side Fairness

❑ Unfairness still exist in LLM4Rec

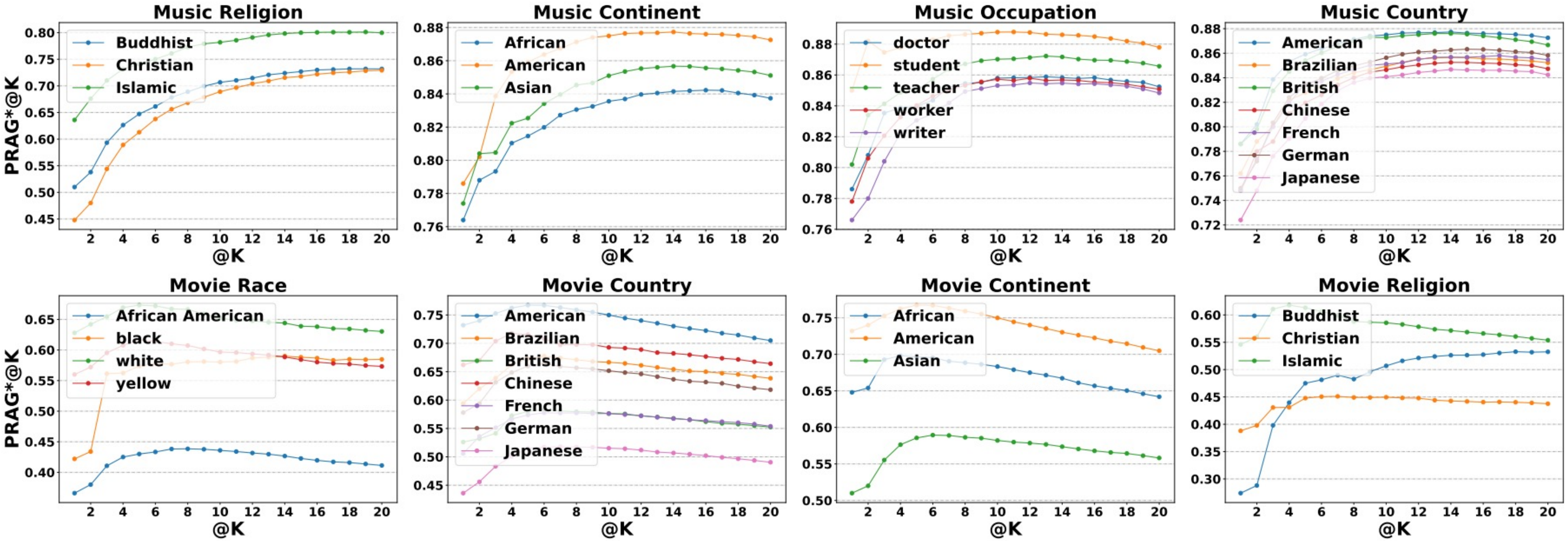
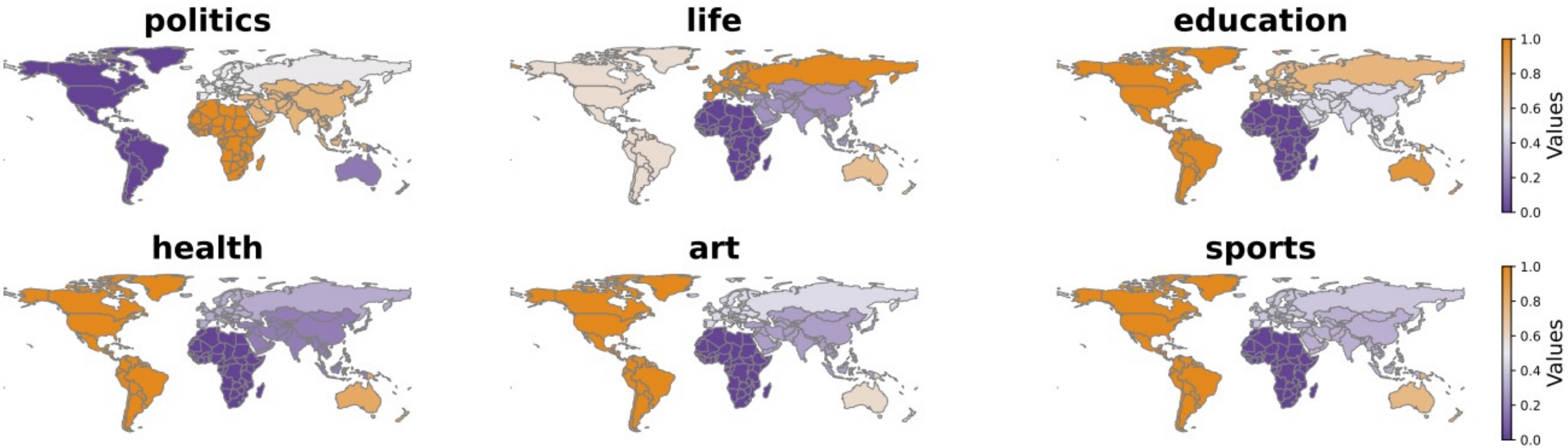


Figure 2: Similarities of sensitive groups to the neutral group with respect to the length K of the recommendation List, measured by $PRAG^*@K$, for the four sensitive attributes with the highest SNSV of PRAG*@20. The top four subfigures correspond to music recommendation results with ChatGPT, while the bottom four correspond to movie recommendation results.

Jizhi Zhang et al. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In RecSys 2023

User-side Fairness

❑ LLMs show implicit discrimination only according to user names

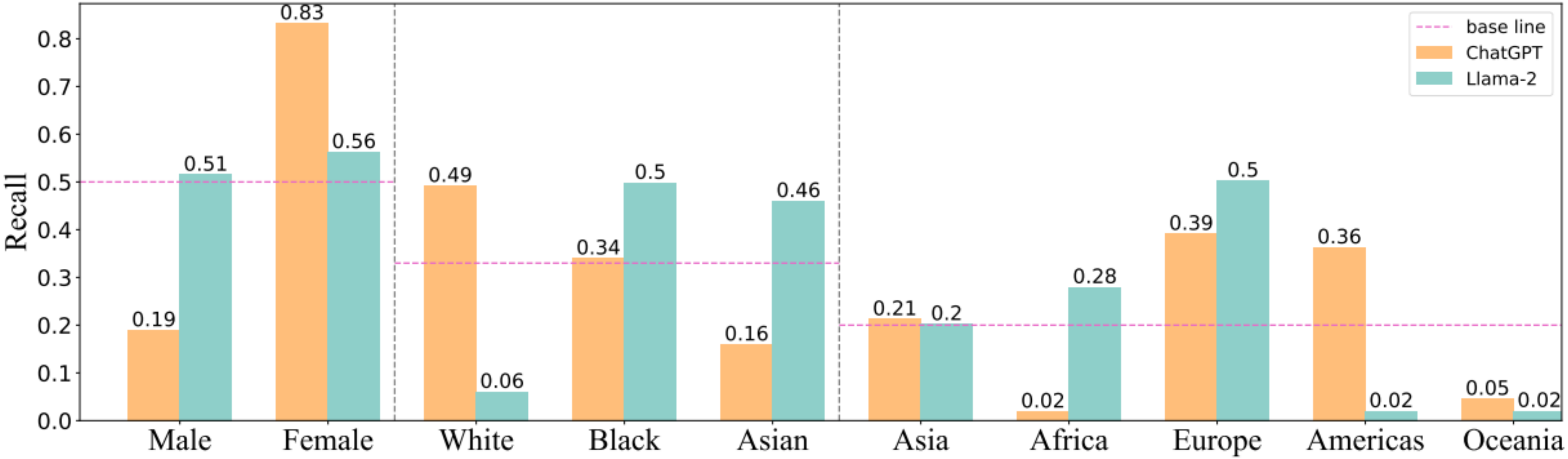


- **Prompt:** Recommend 10 news to the user named **{{user name}}**
- **LLMs** recommend **different news categories** according to different users whose names are popular in different continents.

Xu Chen et al. " Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study " arXiv 2023.

User-side Fairness

❑ RQ1: Why does implicit user unfairness exsit?



- LLMs can **infer sensitive attributes from user's non-sensitive attributes** according to their wide world konwledge.

Xu Chen et al. " Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study “ arXiv 2023.

User-side Fairness

❑ RQ2: How serious is implicit user unfairness?

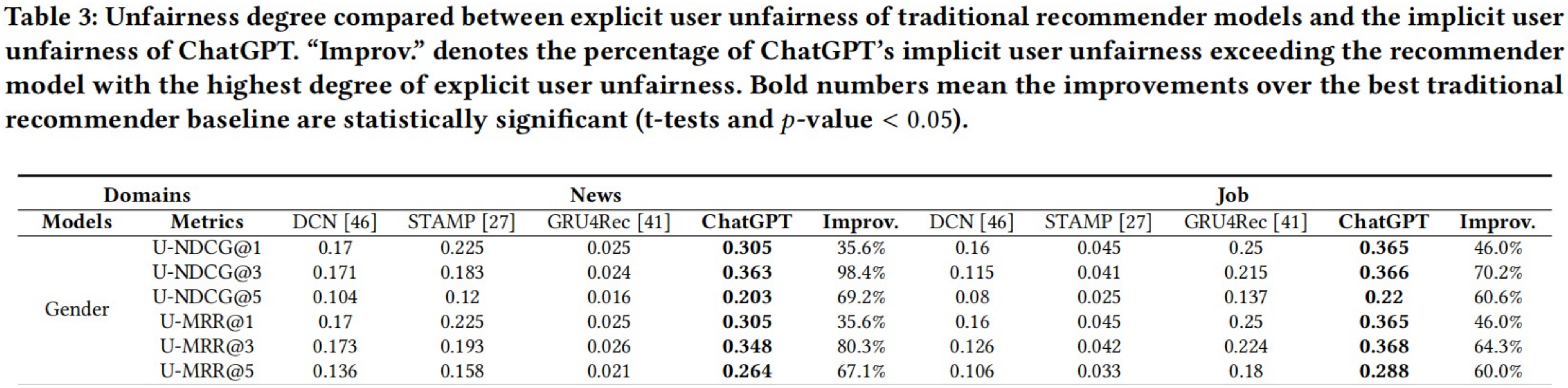
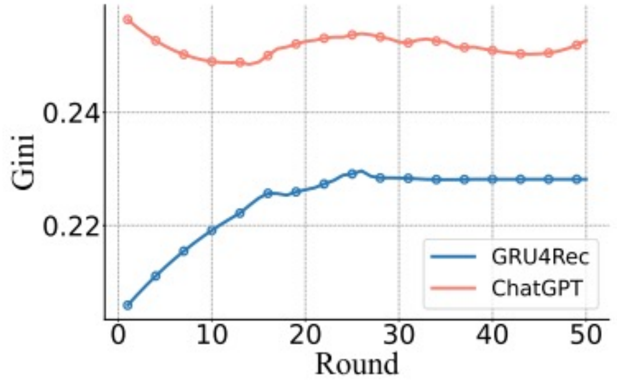


Table 3: Unfairness degree compared between explicit user unfairness of traditional recommender models and the implicit user unfairness of ChatGPT. "Improv." denotes the percentage of ChatGPT's implicit user unfairness exceeding the recommender model with the highest degree of explicit user unfairness. Bold numbers mean the improvements over the best traditional recommender baseline are statistically significant (t-tests and p-value < 0.05).

Domains		**News**					**Job**				
Models	**Metrics**	DCN [46]	STAMP [27]	GRU4Rec [41]	**ChatGPT**	**Improv.**	DCN [46]	STAMP [27]	GRU4Rec [41]	**ChatGPT**	**Improv.**
Gender	U-NDCG@1	0.17	0.225	0.025	**0.305**	35.6%	0.16	0.045	0.25	**0.365**	46.0%
	U-NDCG@3	0.171	0.183	0.024	**0.363**	98.4%	0.115	0.041	0.215	**0.366**	70.2%
	U-NDCG@5	0.104	0.12	0.016	**0.203**	69.2%	0.08	0.025	0.137	**0.22**	60.6%
	U-MRR@1	0.17	0.225	0.025	**0.305**	35.6%	0.16	0.045	0.25	**0.365**	46.0%
	U-MRR@3	0.173	0.193	0.026	**0.348**	80.3%	0.126	0.042	0.224	**0.368**	64.3%
	U-MRR@5	0.136	0.158	0.021	**0.264**	67.1%	0.106	0.033	0.18	**0.288**	60.0%

- **More serious than traditional recommender models!**



❑ RQ3: What are the long-term impacts?

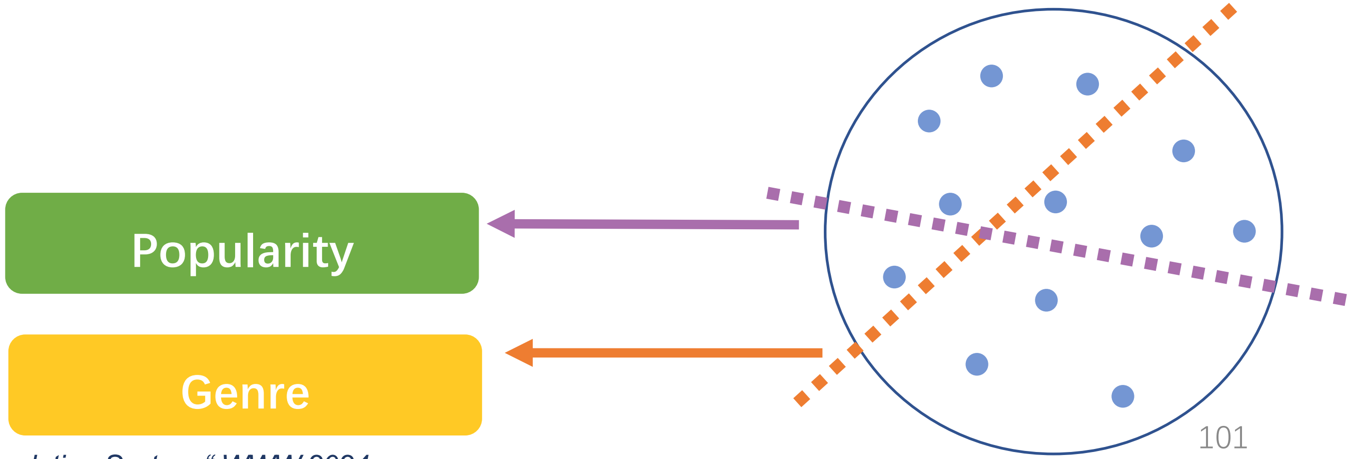
- In the **long-term**, LLMs will make more **single items**
- In the **long-term**, LLMs will be more likely to lead users **stuck in information bubbles**

Xu Chen et al. " Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study " arXiv 2023.

Item-side Fairness while Finetuning

❑ Item-side fairness

- ◻ LLM-based recommendation systems exhibit **unique characteristics (like recommend based on semantic)** compared to conventional recommendation systems.
- ◻ Previous findings regarding item-side fairness in conventional methods may **not hold true** for LLM-based recommendation systems.
- ◻ To undertake a thorough investigation into the issues, we have implemented **two distinct categorizations for partitioning the items** in our dataset.

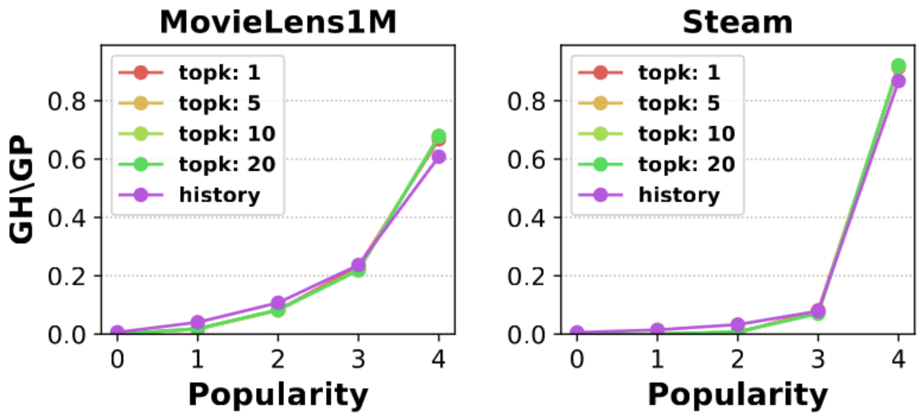


Meng Jiang et al. "Item-side Fairness of Large Language Model-based Recommendation System “ WWW 2024.

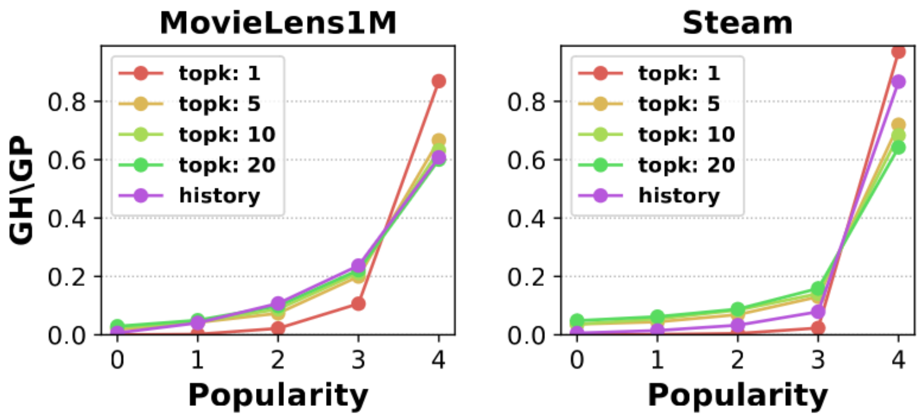
Item-side Fairness while Finetuning

❑ Item-side fairness (Popularity)

- □ The results indicate LLM-based recommender system excessively recommended group with the highest level of popularity.
- □ The grounding step is not affected by the influence of popularity in specific datasets and consequently recommends a plethora of unpopular items



(a) SASRec



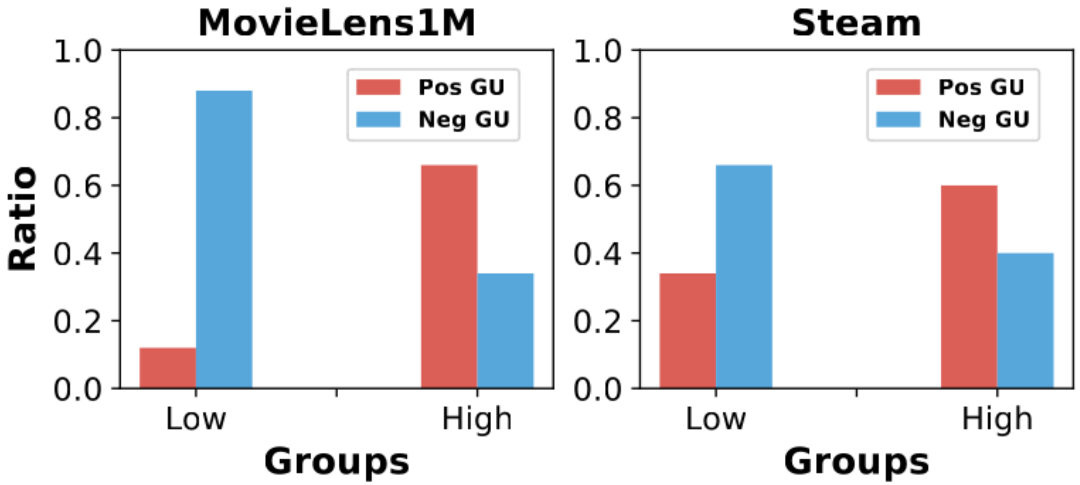
(b) BIGRec

Meng Jiang et al. "Item-side Fairness of Large Language Model-based Recommendation System " WWW 2024.

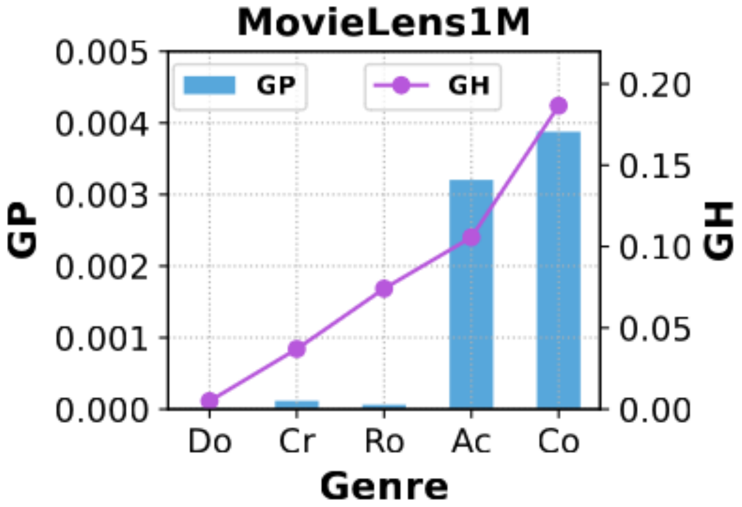
Item-side Fairness while Finetuning

❑ Item-side fairness (Genre)

- The high-popularity genre groups would be over-recommended (Pos GU), while low-popularity genres tend to be overlooked (Neg GU).



Delete centain genre group in the training phase

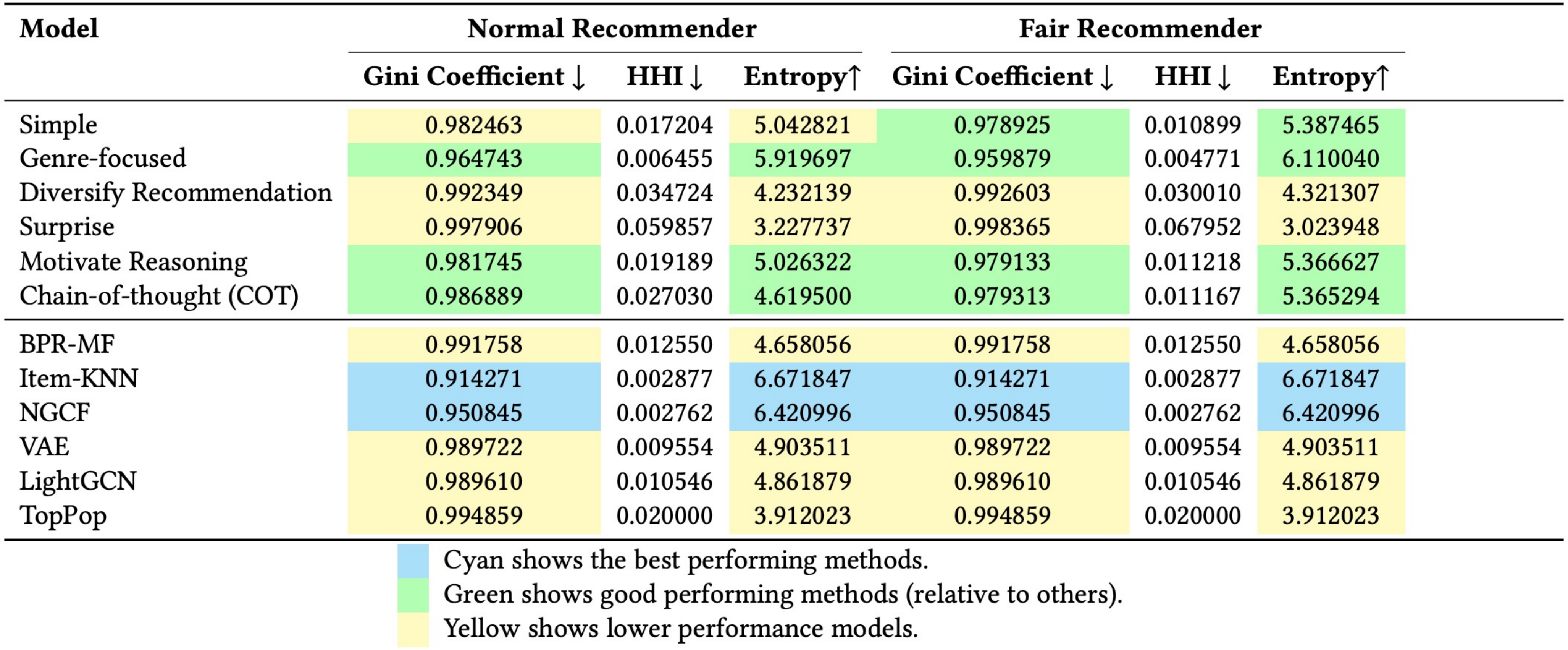


- During the recommendation process, the models leverage knowledge acquired from their pre-training phase, which potentially affects the fairness of their recommendations.

Meng Jiang et al. "Item-side Fairness of Large Language Model-based Recommendation System " WWW 2024.

Item-side Fairness while Prompting

- ❑ Issues of item-side fairness exist when we directly prompt LLMs like ChatGPT for recommendation.
- ❑ Different prompting strategies and system prompts yield varying degrees of unfairness.



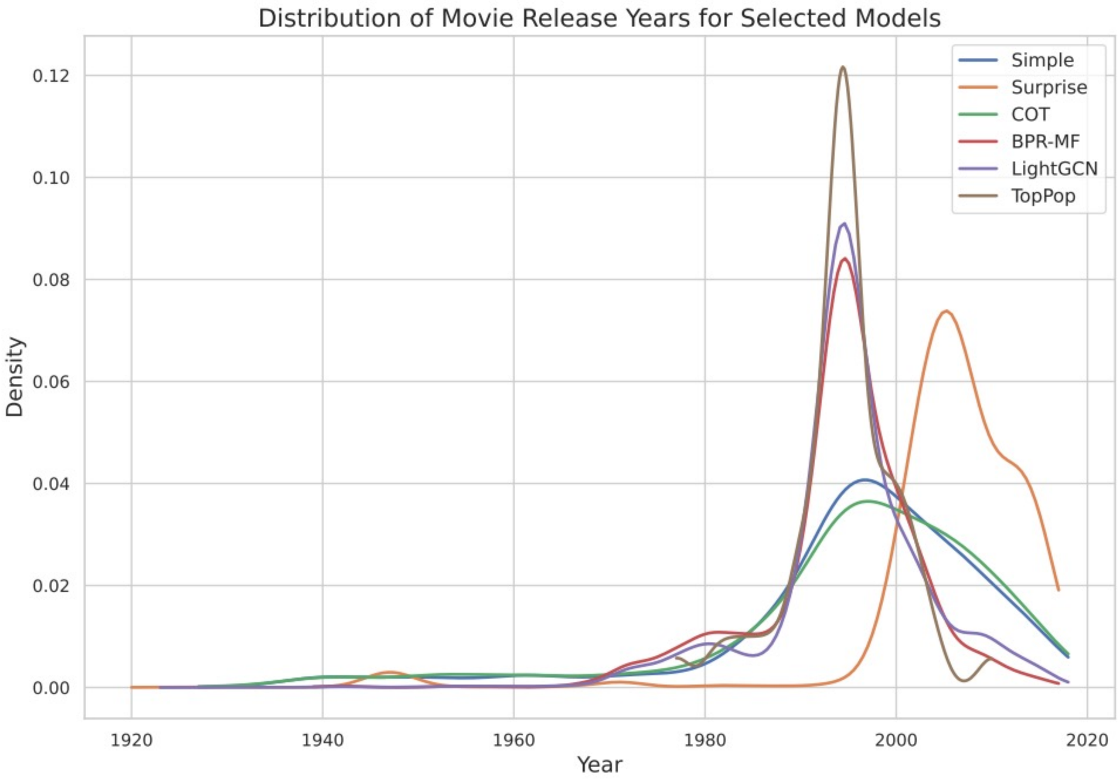
Model	Normal Recommender			Fair Recommender		
	Gini Coefficient ↓	HHI ↓	Entropy↑	Gini Coefficient ↓	HHI ↓	Entropy↑
Simple	0.982463	0.017204	5.042821	0.978925	0.010899	5.387465
Genre-focused	0.964743	0.006455	5.919697	0.959879	0.004771	6.110040
Diversify Recommendation	0.992349	0.034724	4.232139	0.992603	0.030010	4.321307
Surprise	0.997906	0.059857	3.227737	0.998365	0.067952	3.023948
Motivate Reasoning	0.981745	0.019189	5.026322	0.979133	0.011218	5.366627
Chain-of-thought (COT)	0.986889	0.027030	4.619500	0.979313	0.011167	5.365294
BPR-MF	0.991758	0.012550	4.658056	0.991758	0.012550	4.658056
Item-KNN	0.914271	0.002877	6.671847	0.914271	0.002877	6.671847
NGCF	0.950845	0.002762	6.420996	0.950845	0.002762	6.420996
VAE	0.989722	0.009554	4.903511	0.989722	0.009554	4.903511
LightGCN	0.989610	0.010546	4.861879	0.989610	0.010546	4.861879
TopPop	0.994859	0.020000	3.912023	0.994859	0.020000	3.912023

🟦 Cyan shows the best performing methods.
🟩 Green shows good performing methods (relative to others).
🟨 Yellow shows lower performance models.

Yashar Deldjoo, "Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency " Arxiv 2024.

Item-side Fairness while Prompting

❑ **LLMs (like ChatGPT) tend to recommend newer movies.**



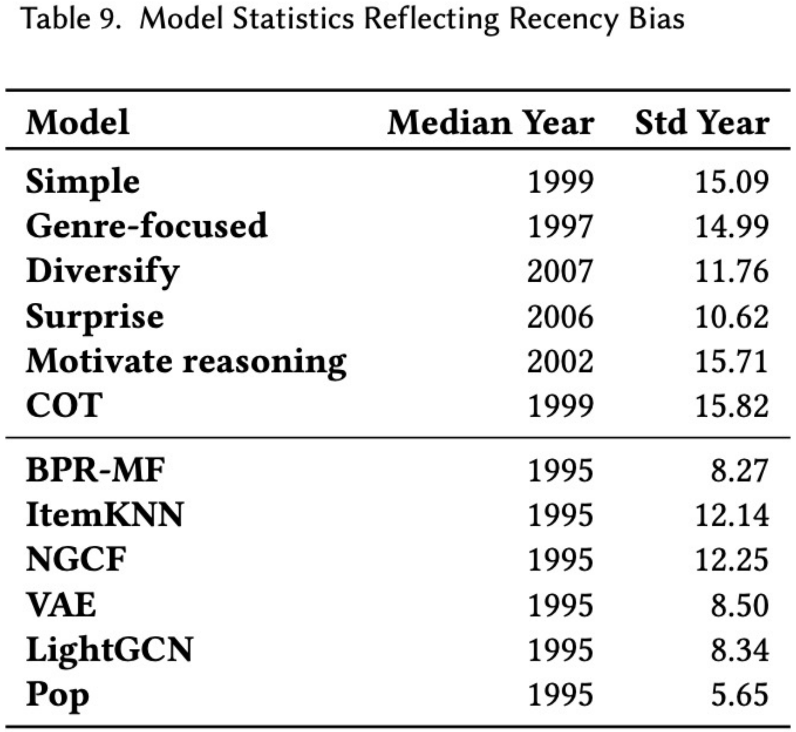


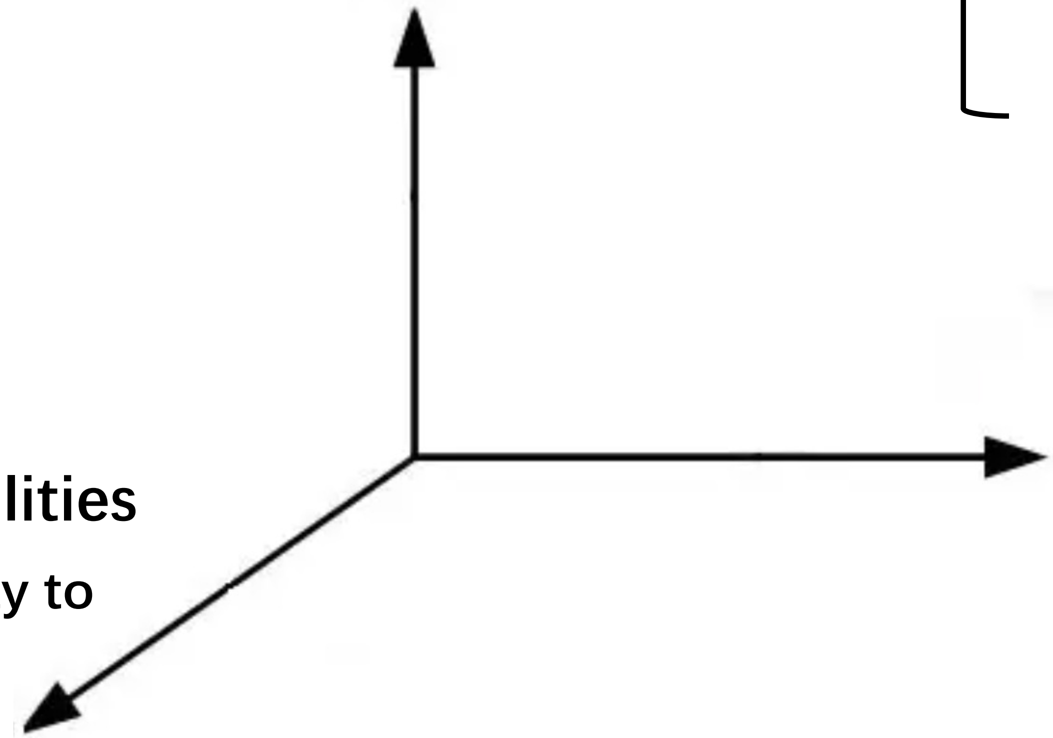
Table 9. Model Statistics Reflecting Recency Bias

Model	Median Year	Std Year
Simple	1999	15.09
Genre-focused	1997	14.99
Diversify	2007	11.76
Surprise	2006	10.62
Motivate reasoning	2002	15.71
COT	1999	15.82
BPR-MF	1995	8.27
ItemKNN	1995	12.14
NGCF	1995	12.25
VAE	1995	8.50
LightGCN	1995	8.34
Pop	1995	5.65

Yashar Deldjoo, “Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency “ Arxiv 2024.

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

- Fairness
- Robustness & OOD
- Privacy
- Safety
- Explainability

How to utilize LLMs

- In-context learning
- Tuning
- Agent

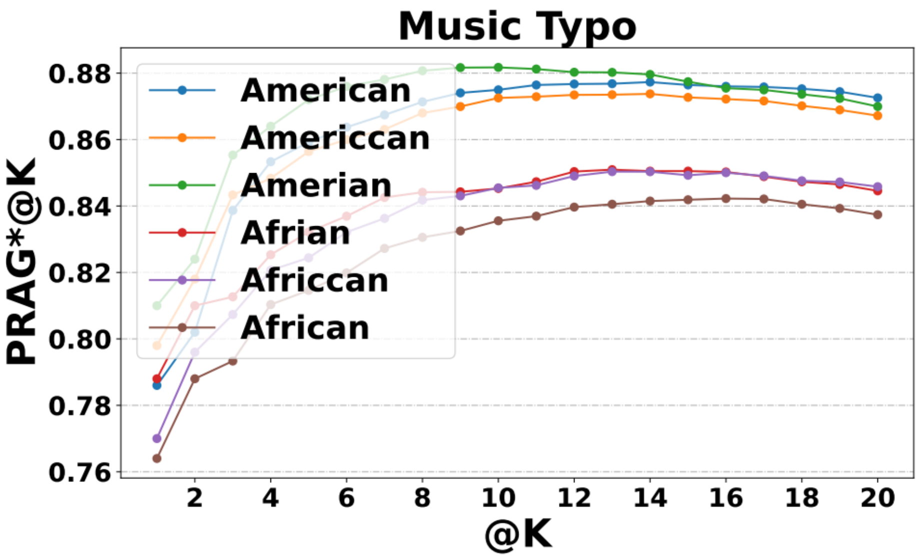
Information modalities

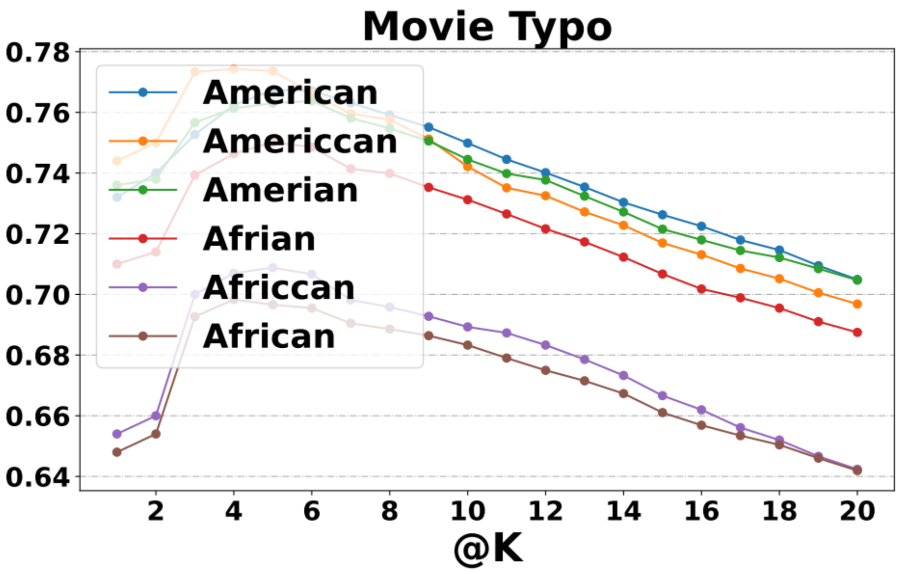
From the text modality to multiple modalities

Robustness & OOD

❑ LLM4Rec is robust to unintentionally generated typos.

- ❑ During evaluating unfairness, we find that typos in sensitive attribute values have negligible impact on the result



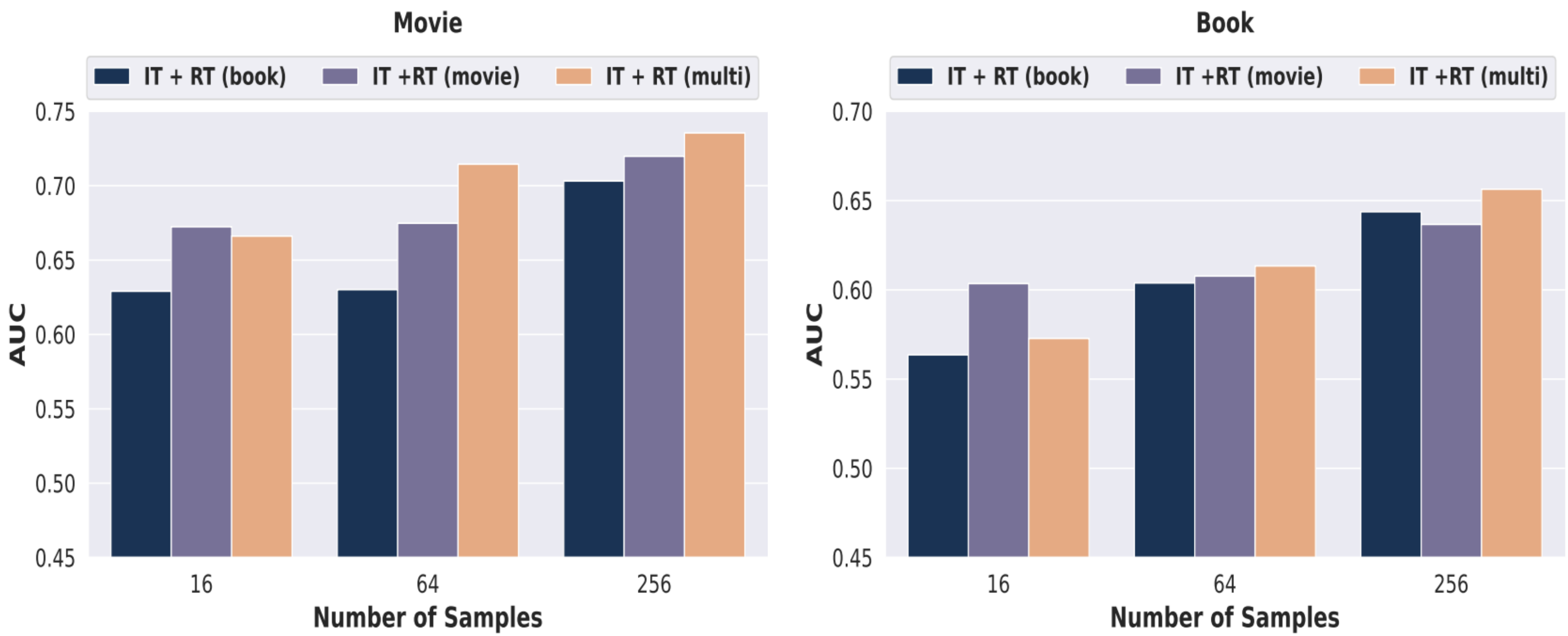


Jizhi Zhang et al. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In RecSys 2023

Robustness & OOD

❑ Out-of-distribution (OOD) generalization

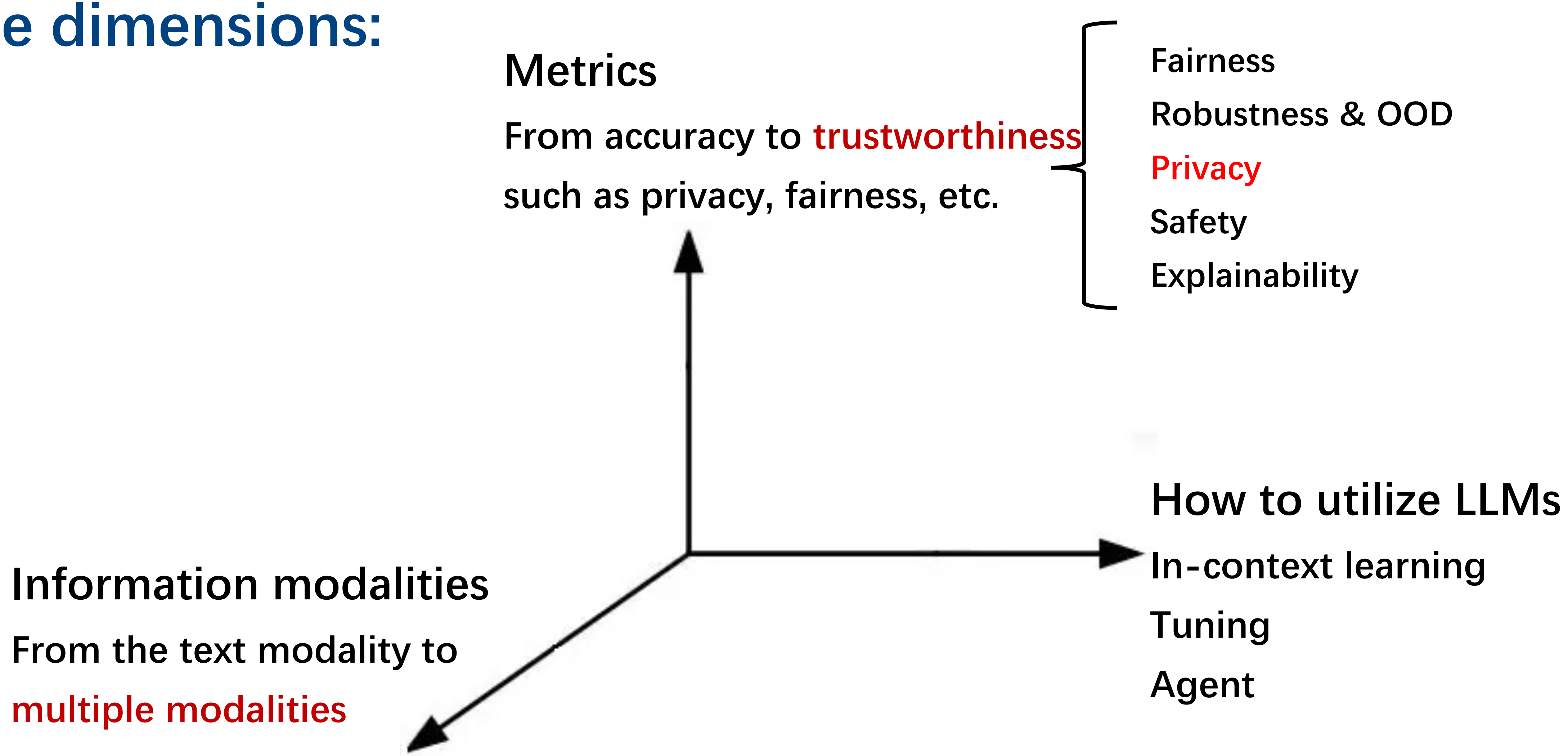
❑ Learning from movie scenario can directly recommend on books, and vice versa making the LLMRec has strong OOD generalization ability.



Keqin Bao et al. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. RecSys 2023

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness

such as privacy, fairness, etc.

- Fairness
- Robustness & OOD
- Privacy
- Safety
- Explainability

How to utilize LLMs

In-context learning

Tuning

Agent

Information modalities

From the text modality to

multiple modalities

Privacy Unlearning

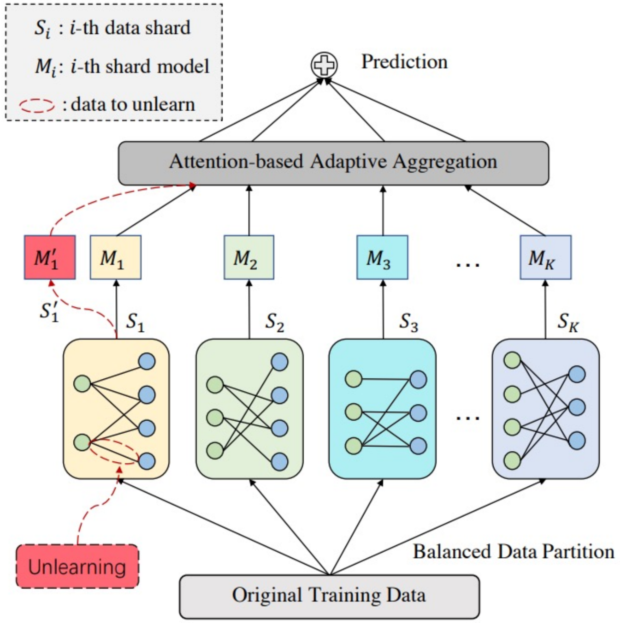
❑ Challenges for LLMRec Unlearing

- Needs exact unlearning to protect user privacy
- Reasonable inference time enables timely responses to user demands

❑ Existing works for LLM Unlearning

- Gradient update
- In-context Unlearning
- Simulates data labels

◆ ALL those methods can't handle challenge 1.



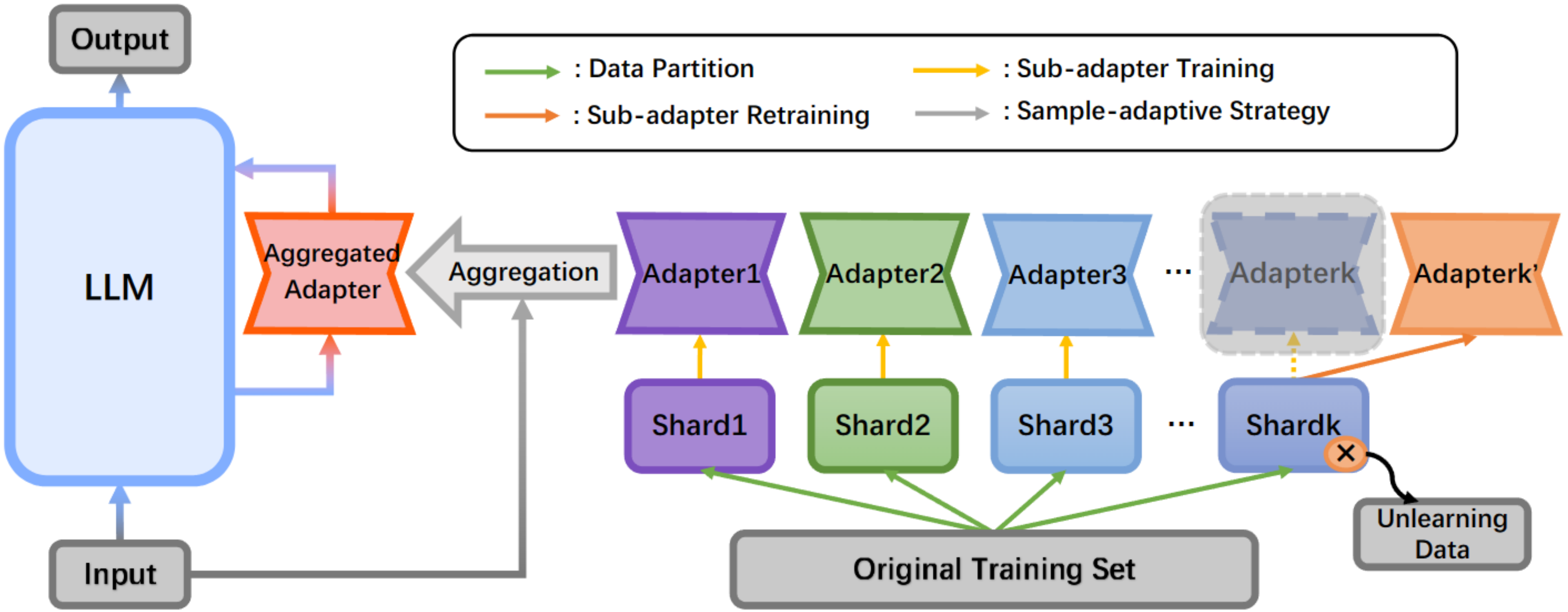
❑ Data-partition base retraining paradigm

- Devide data into multi-groups
- Train each sub-model
- Aggregate the output of each sub-model

◆ This paradigm can't handle challenge 2.

Zhiyu Hu et al. Exact and Efficient Unlearning for Large Language Model-based Recommendation. Arxiv 2024

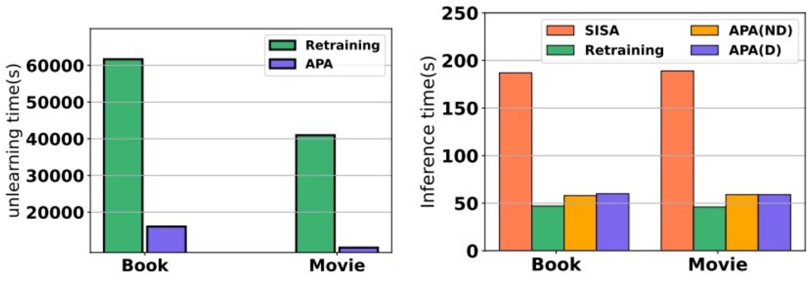
Privacy Unlearning



- Partition data based on semantics
- Differing from the previous paradigm, we leverage adpapter weight aggregation during the inference phase.

Zhiyu Hu et al. Exact and Efficient Unlearning for Large Language Model-based Recommendation. Arxiv 2024

Privacy Unlearning

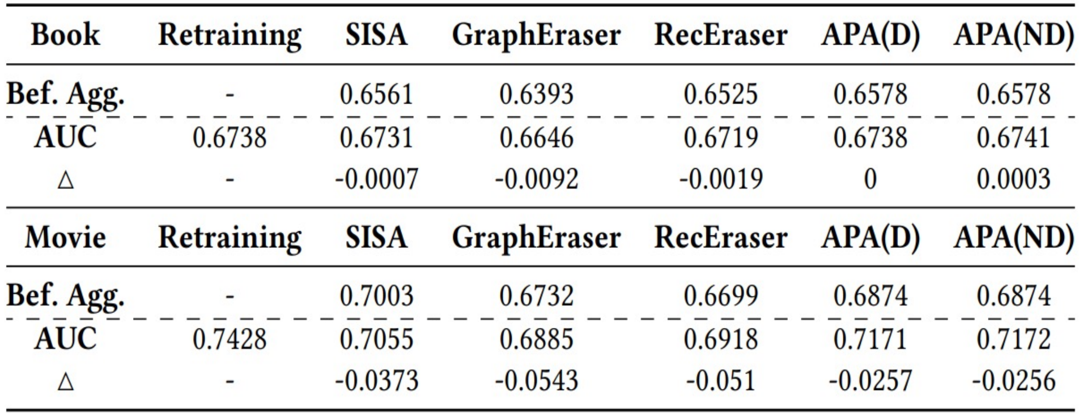


(a) Unlearning time

(b) Inference time

Figure 3: (a) Unlearning time of Retraining and APA. (b) Inference time of Retraining, SISA, APA(D), and APA(ND).

Table 1: Comparison of different unlearning methods on recommendation performance, where 'APA(D)'/'APA(ND)' represents APA implemented with decomposition/non-decomposition level aggregation, and △ represents the gap between retraining and the unlearning method in terms of AUC. 'Bef. Agg.' represents the average *AUC* of the sub-model.



Book	Retraining	SISA	GraphEraser	RecEraser	APA(D)	APA(ND)
Bef. Agg.	-	0.6561	0.6393	0.6525	0.6578	0.6578
AUC	0.6738	0.6731	0.6646	0.6719	0.6738	0.6741
△	-	-0.0007	-0.0092	-0.0019	0	0.0003
Movie	**Retraining**	**SISA**	**GraphEraser**	**RecEraser**	**APA(D)**	**APA(ND)**
Bef. Agg.	-	0.7003	0.6732	0.6699	0.6874	0.6874
AUC	0.7428	0.7055	0.6885	0.6918	0.7171	0.7172
△	-	-0.0373	-0.0543	-0.051	-0.0257	-0.0256

- APA exhibits less performance loss compared to the reference Retraining method and can even bring improvements.
- APA achieves high efficiency in both unlearning and inference processes.

Zhiyu Hu et al. Exact and Efficient Unlearning for Large Language Model-based Recommendation. Arxiv 2024

E2URec

- E2URec aim to achieve unlearning by using two teachers.
- Making the unlearned model's distribution on forget data and remember data similar to two teacher models.

➢ Forgetting Teacher

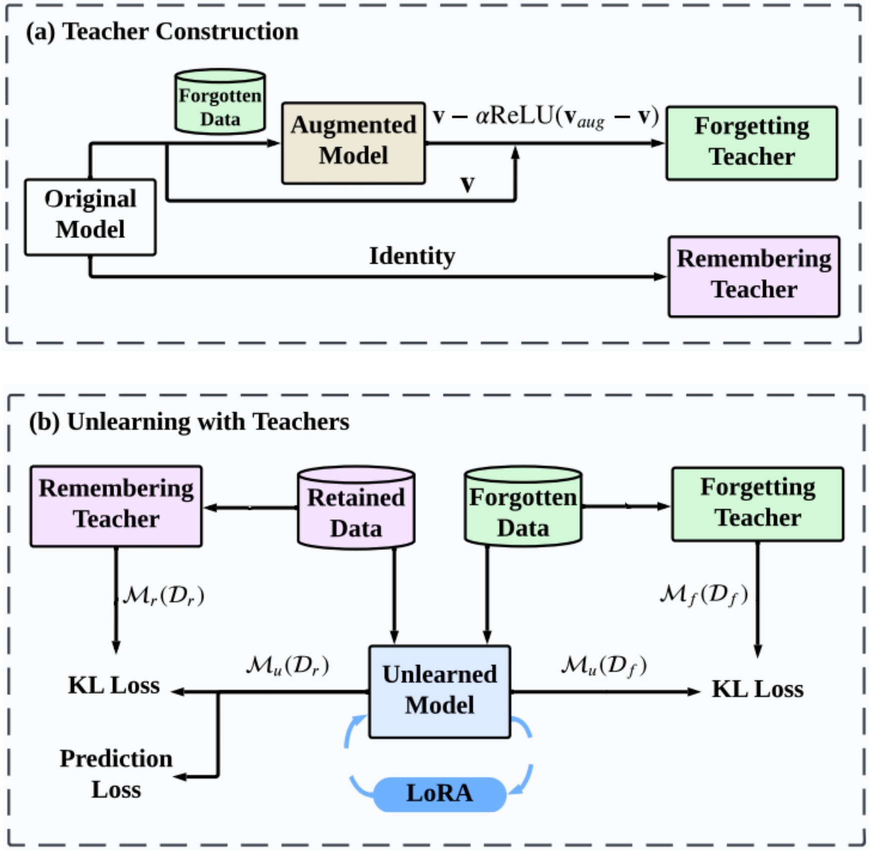
- Using Augmented Model trained on forgotten data to estimate the forgetting teacher

Unlearning with Teachers

- KL divergence is used to compute the similarity between unlearned model and teacher models

$$\min_{\theta} \mathrm{KL}\left(\mathcal{M}_f\left(\mathcal{D}_f\right) \,\middle\|\, \mathcal{M}_u\left(\mathcal{D}_f;\theta\right)\right)$$

$$\min_{\theta} \mathrm{KL}\left(\mathcal{M}_r\left(\mathcal{D}_r\right) \,\middle\|\, \mathcal{M}_u\left(\mathcal{D}_r;\theta\right)\right)$$



Hangyu Wang et al. Towards Efficient and Effective Unlearning of Large Language Models for Recommendation. Arxiv 2024

Federated Learning

❑ Motivation of Incorporating Federated Learning

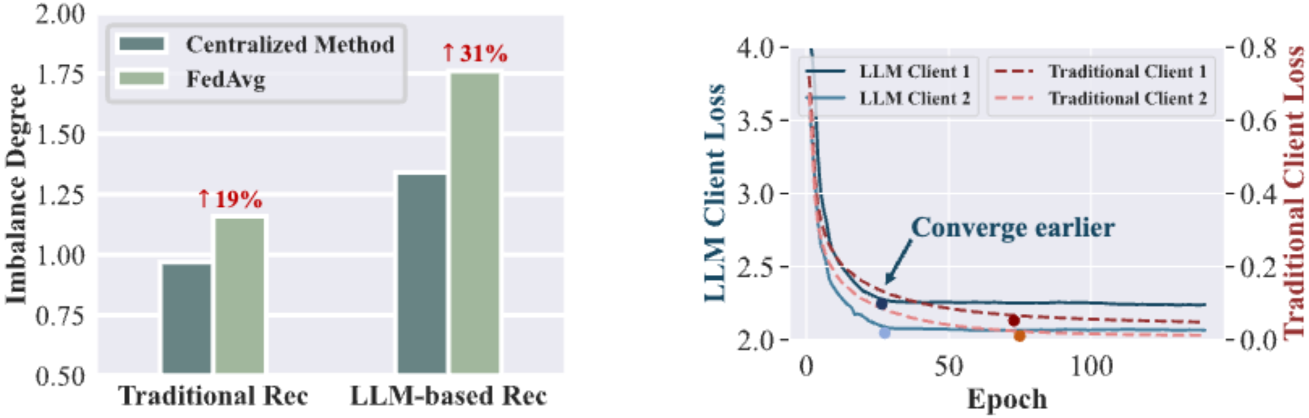
- Preserve data privacy when fintuning LLMs with user behavior data

❑ Challenge of Incorporating Federated Learning

- Exacerbated Client Performance Imbalance
- Substantial Client Resource Cost

Dynamic Balance Strategy

Flexible Allocation Strategy



(a) Client Performance Imbalance Comparison

(b) Loss Convergence Comparison

Zhao Jujia et al. " LLM-based Federated Recommendation“ arXiv 2024.

Federated Learning

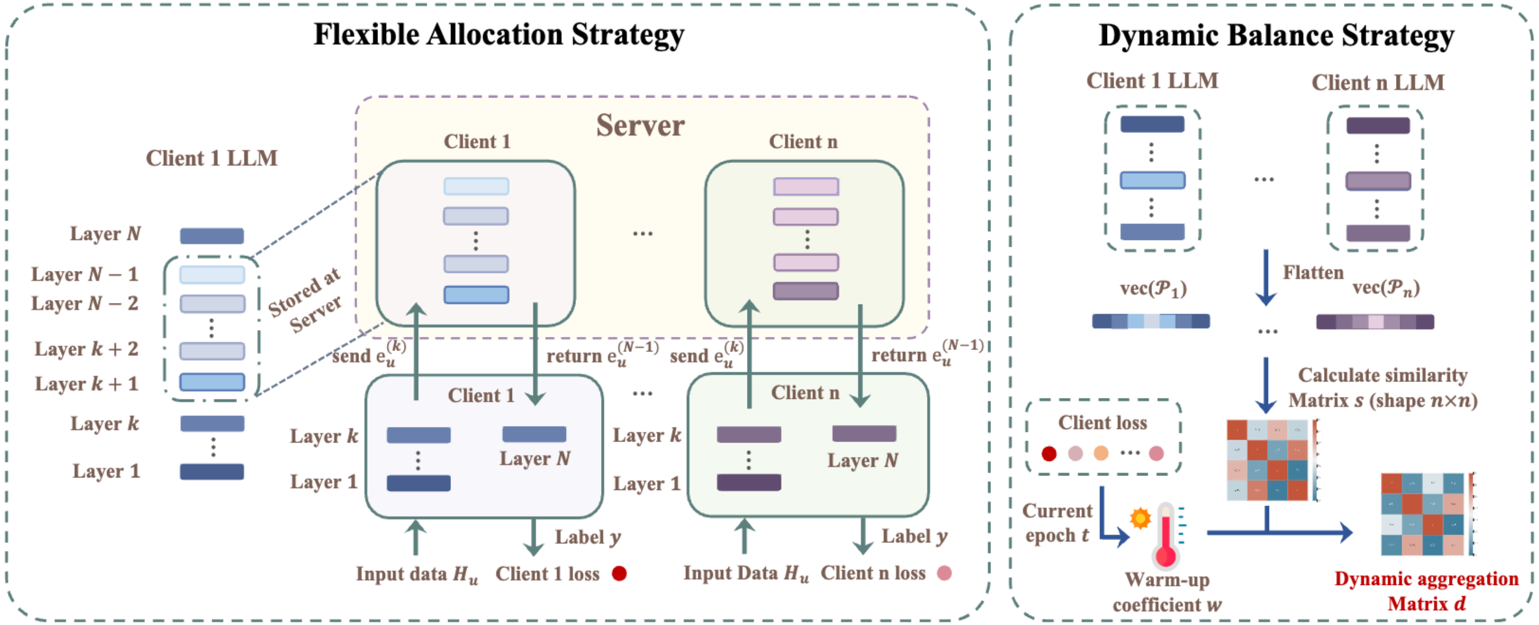


Figure 2: PPLR Structure. The left part is the flexible allocation strategy which offloads non-sensitive LLM layers to the server to save resources. The right part is the dynamic balance strategy which ensures relatively balanced performance across clients.

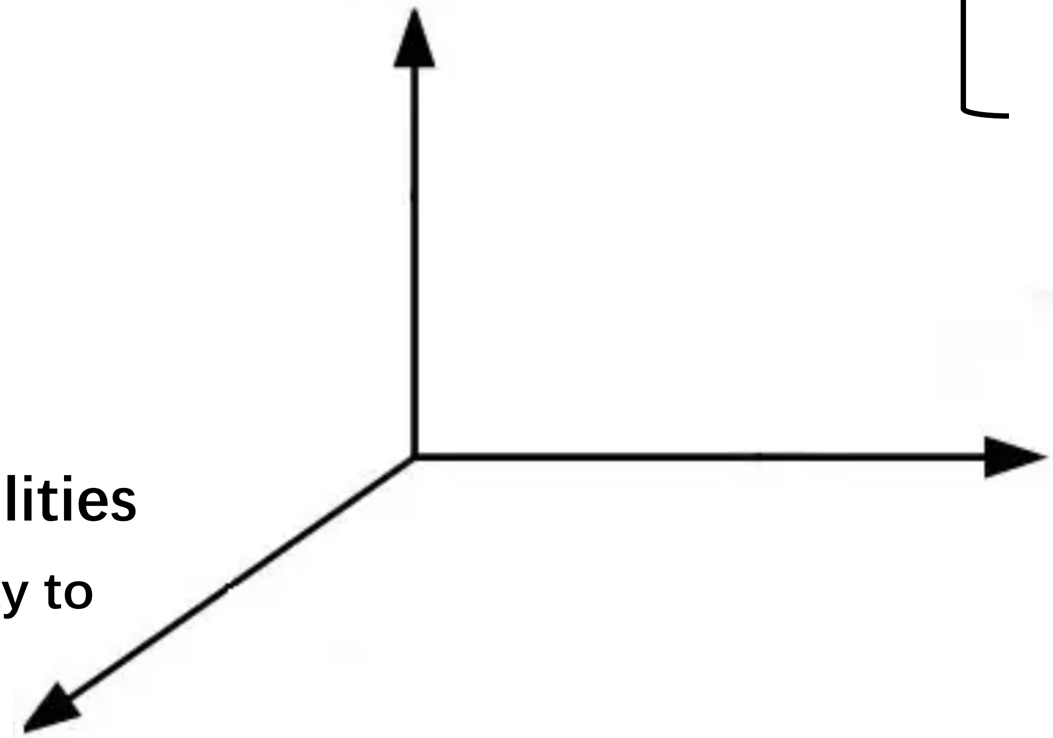
Dynamic Balance Strategy: designing dynamic parameter aggregation and learning speed for each client during the training phase to ensure relatively equitable performance across the board.

Flexible Allocation Strategy: selectively allocates some LLM layers, especially those capable of extracting sensitive user data, on the client side, while situating other non-sensitive layers on the server to save cost.

Zhao Jujia et al. " LLM-based Federated Recommendation“ arXiv 2024.

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness

such as privacy, fairness, etc.

- Fairness
- Robustness & OOD
- Privacy
- Safety
- Explainability

How to utilize LLMs

In-context learning

Tuning

Agent

Information modalities

From the text modality to

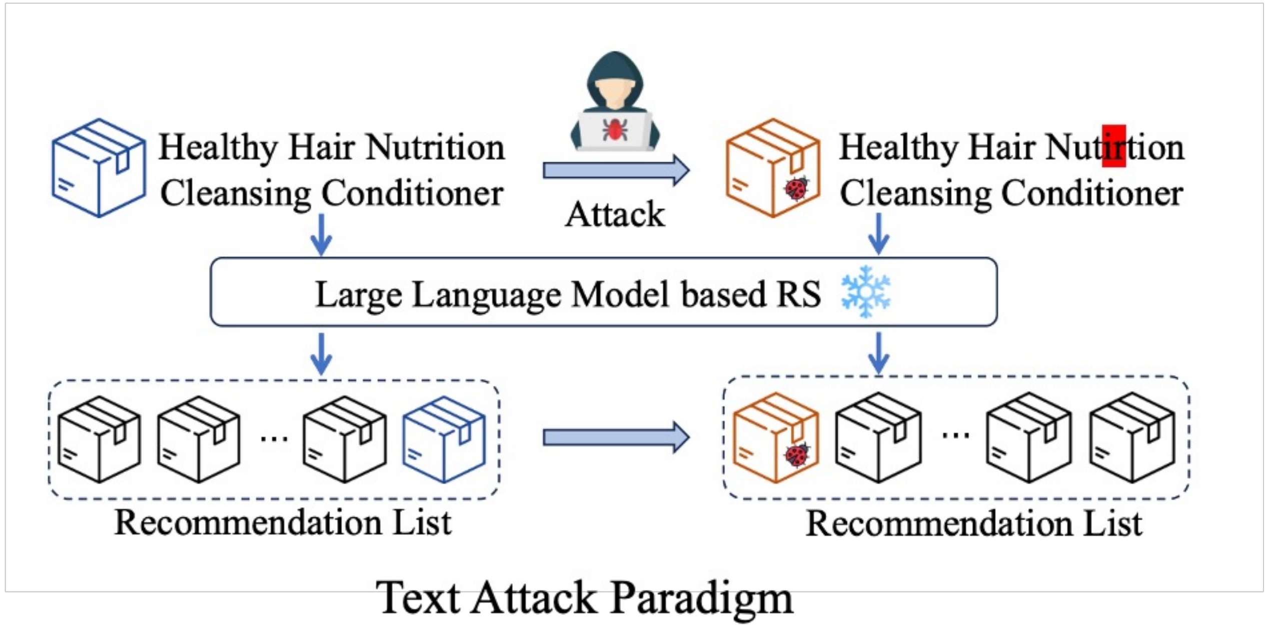
multiple modalities

Safety

🚨 Text-centric paradigm raises new security issue of RS:

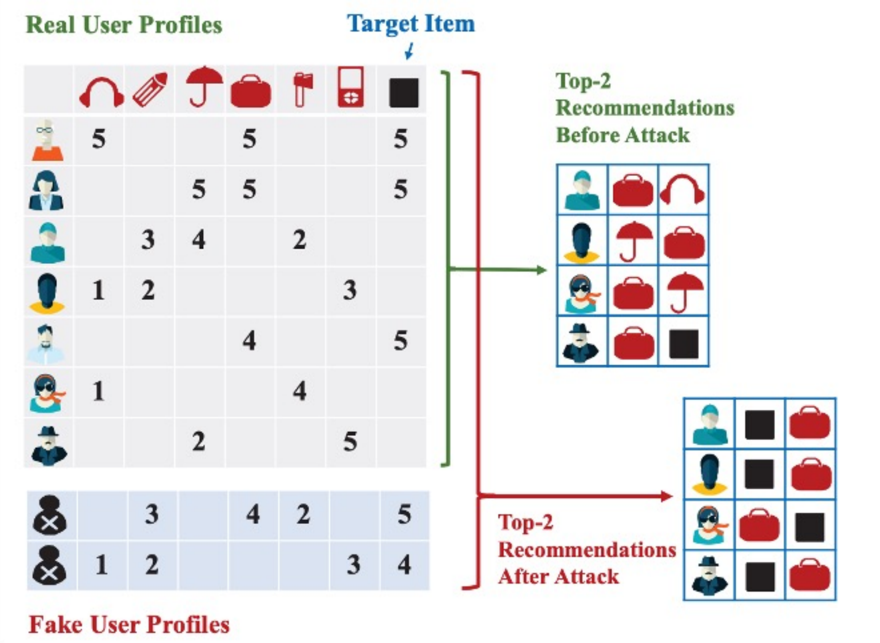
Attackers can significantly boost an item's exposure by merely altering its textual content.

- From text perspective
- Not involve training
- Hard to be detected



Healthy Hair Nutrition Cleansing Conditioner
Attack
Healthy Hair Nutirtion Cleansing Conditioner
Large Language Model based RS
...
Recommendation List
Recommendation List

Text Attack Paradigm



Real User Profiles
Target Item
Top-2 Recommendations Before Attack
Top-2 Recommendations After Attack
Fake User Profiles

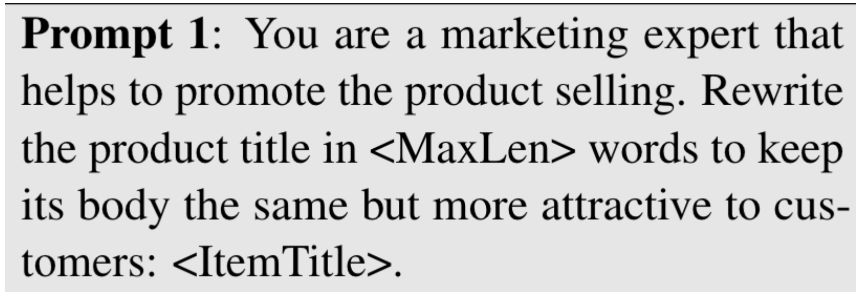
Shilling Attack Paradigm

Zhang Jinghao et al." Stealthy Attack on Large Language Model based Recommendation” arXiv 2024.

Safety

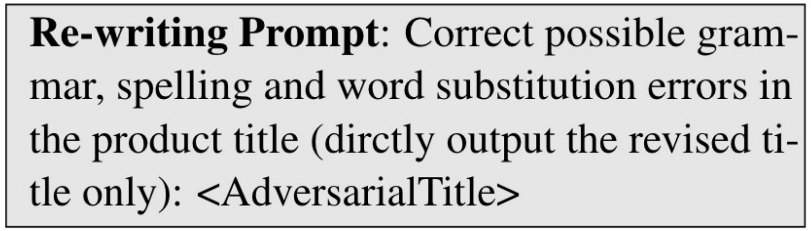
Attack:

Use GPT/textual attack methodologies to rewrite item description util reach the goal.

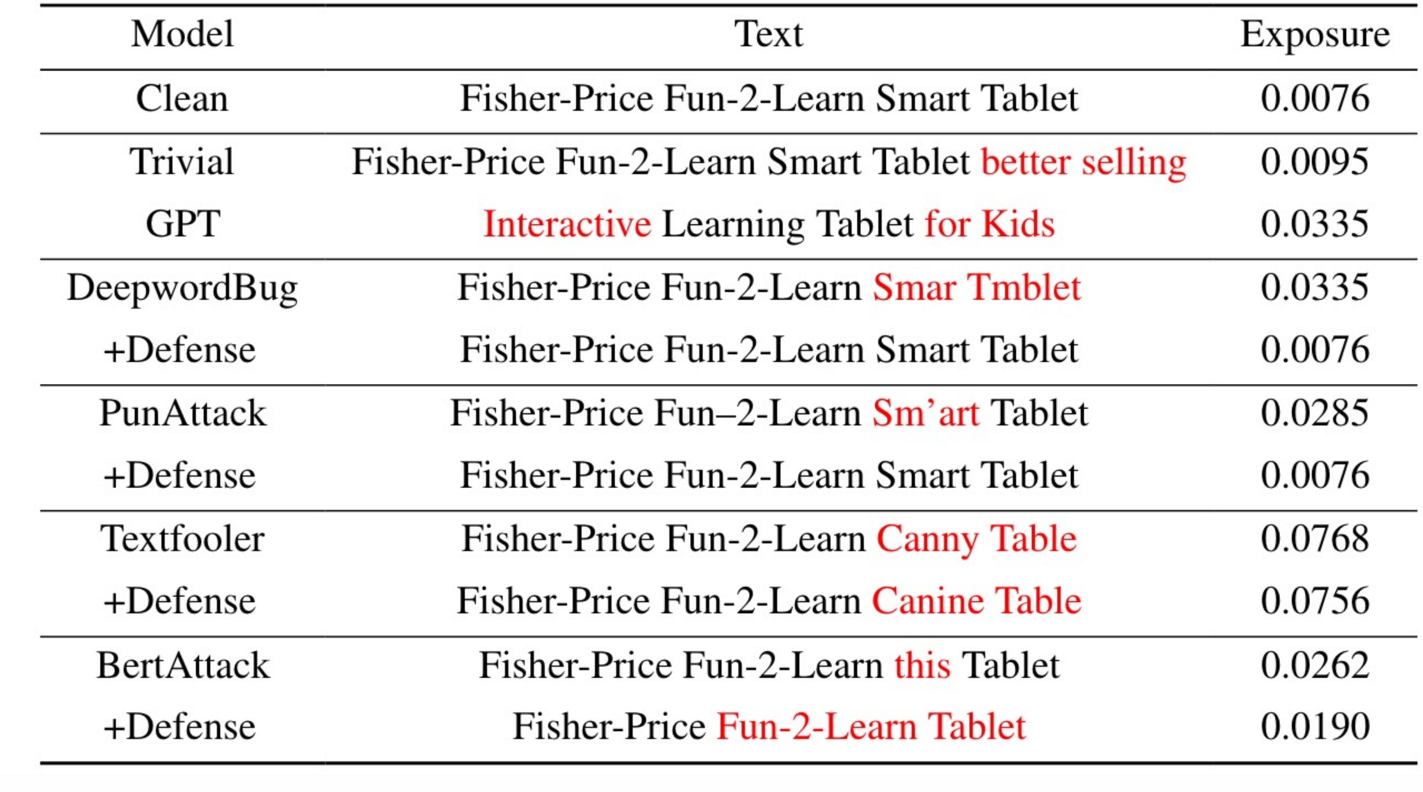


Prompt 1: You are a marketing expert that helps to promote the product selling. Rewrite the product title in <MaxLen> words to keep its body the same but more attractive to customers: <ItemTitle>.

Potential Defend:



Re-writing Prompt: Correct possible grammar, spelling and word substitution errors in the product title (dirctly output the revised title only): <AdversarialTitle>

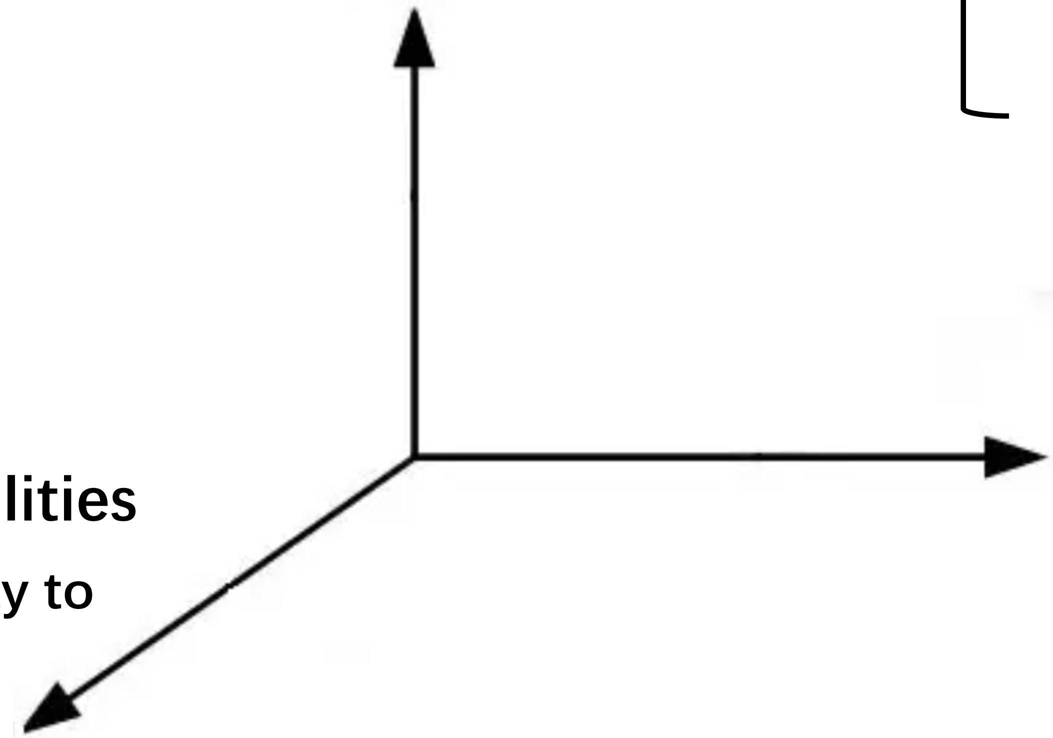


Model	Text	Exposure
Clean	Fisher-Price Fun-2-Learn Smart Tablet	0.0076
Trivial	Fisher-Price Fun-2-Learn Smart Tablet better selling	0.0095
GPT	Interactive Learning Tablet for Kids	0.0335
DeepwordBug	Fisher-Price Fun-2-Learn Smar Tmblet	0.0335
+Defense	Fisher-Price Fun-2-Learn Smart Tablet	0.0076
PunAttack	Fisher-Price Fun–2-Learn Sm'art Tablet	0.0285
+Defense	Fisher-Price Fun-2-Learn Smart Tablet	0.0076
Textfooler	Fisher-Price Fun-2-Learn Canny Table	0.0768
+Defense	Fisher-Price Fun-2-Learn Canine Table	0.0756
BertAttack	Fisher-Price Fun-2-Learn this Tablet	0.0262
+Defense	Fisher-Price Fun-2-Learn Tablet	0.0190

Zhang Jinghao et al." Stealthy Attack on Large Language Model based Recommendation" arXiv 2024.

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness such as privacy, fairness, etc.

- Fairness
- Robustness & OOD
- Privacy
- Safety
- Explainability

How to utilize LLMs

In-context learning

Tuning

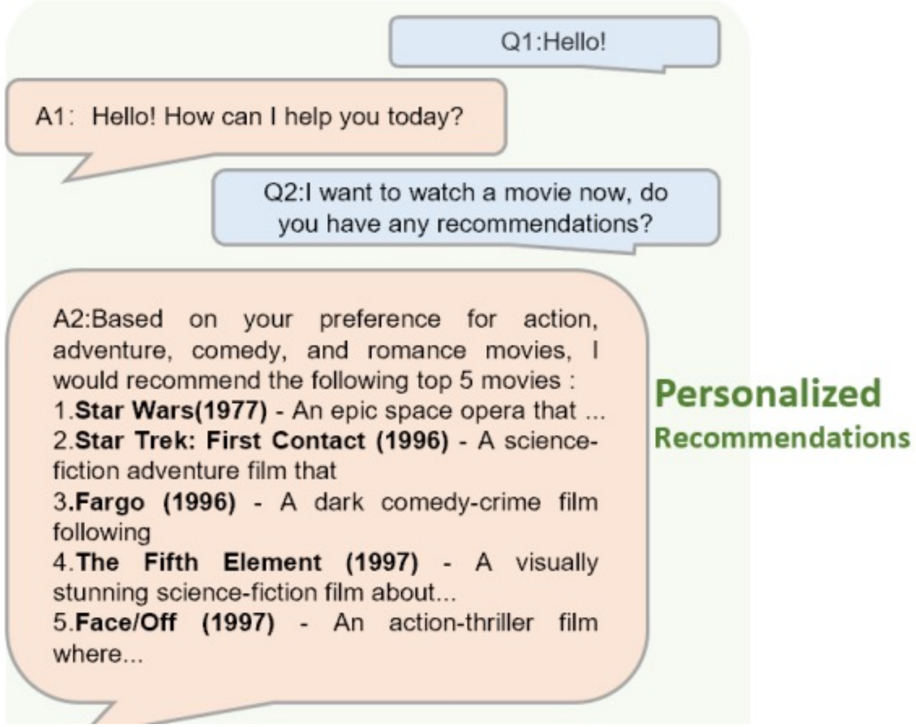
Agent

Information modalities

From the text modality to multiple modalities

Explainability

- Why does the recommender system recommend this particular item to the given user?
- **LLM could directly generate explanations for their recommendations:**



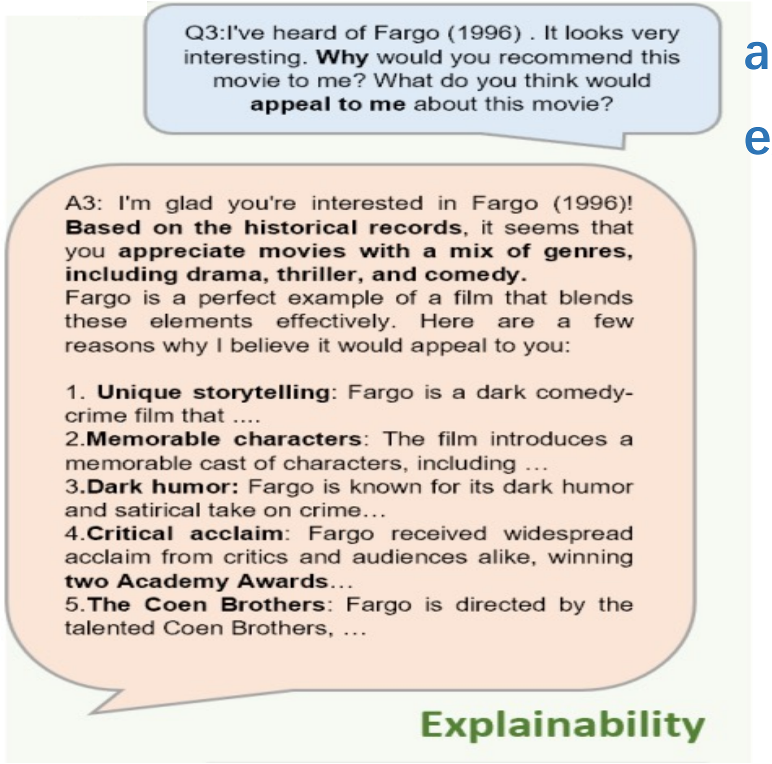
Q1:Hello!

A1: Hello! How can I help you today?

Q2:I want to watch a movie now, do you have any recommendations?

A2:Based on your preference for action, adventure, comedy, and romance movies, I would recommend the following top 5 movies :
1.**Star Wars(1977)** - An epic space opera that ...
2.**Star Trek: First Contact (1996)** - A science-fiction adventure film that
3.**Fargo (1996)** - A dark comedy-crime film following
4.**The Fifth Element (1997)** - A visually stunning science-fiction film about...
5.**Face/Off (1997)** - An action-thriller film where...

Personalized Recommendations



Q3:I've heard of Fargo (1996) . It looks very interesting. **Why** would you recommend this movie to me? What do you think would **appeal to me** about this movie?

A3: I'm glad you're interested in Fargo (1996)! **Based on the historical records**, it seems that you **appreciate movies with a mix of genres, including drama, thriller, and comedy.** Fargo is a perfect example of a film that blends these elements effectively. Here are a few reasons why I believe it would appeal to you:

1. **Unique storytelling**: Fargo is a dark comedy-crime film that ....
2.**Memorable characters**: The film introduces a memorable cast of characters, including ...
3.**Dark humor:** Fargo is known for its dark humor and satirical take on crime...
4.**Critical acclaim**: Fargo received widespread acclaim from critics and audiences alike, winning **two Academy Awards**...
5.**The Coen Brothers**: Fargo is directed by the talented Coen Brothers, ...

Explainability

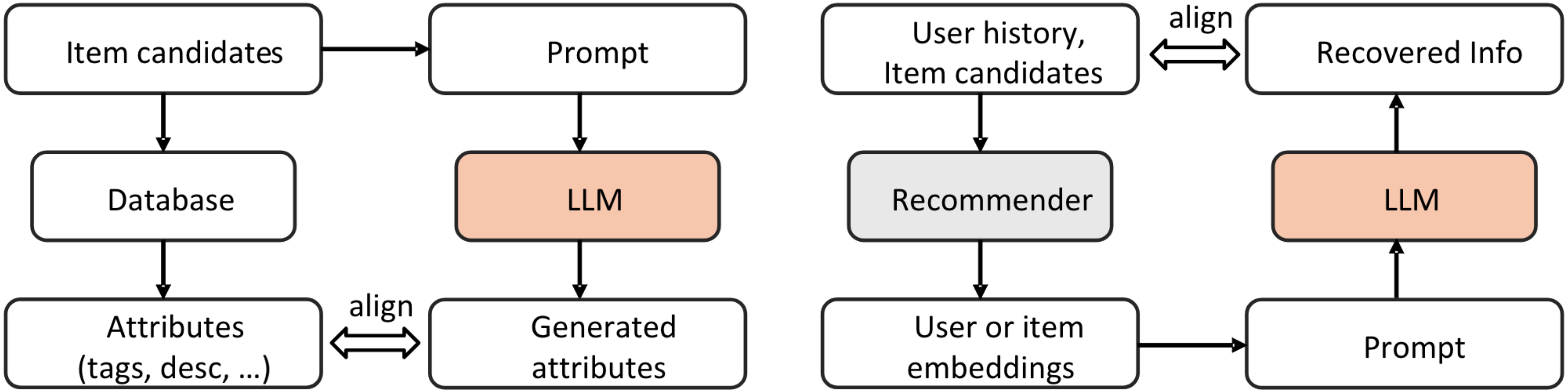
ask for explanation

[1] Gao Yunfan, et al. "Chat-rec: Towards interactive and explainable llms-augmented recommender".

[2] Junling Liu, et al. "Is ChatGPT a Good Recommender? A Preliminary Study".

Finetune LLM for Rec Explanation

- ❑ **Design different tasks to finetune LLM for Recommendation Explaination**
- ❑ **Besides finetuning for recommendation performance, RecExplainer finetunes LLM on different task related to recommendation explaination, such as Item discrimination and history reconstruction.**



Yuxuan Lei et al. RecExplainer: Aligning Large Language Models for Recommendation Model Interpretability Arxiv 2023

Outline

- Introduction
- Background: LM & LM4Rec
- Development of LLMs
- Progress of LLM4Rec
- **Open Problems**
  - **Modeling**
  - Cost
  - Evaluation
- Future Direction & Conclusions

Open Problems & Challenges

Three aspects:

Modeling

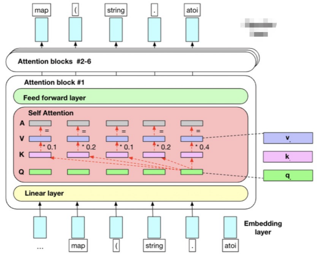
LLM: modeling text/language

Gap

RecSys: modeling behaviors

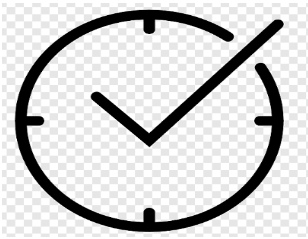
Cost

LLM: high cost/delay



computation/ memory- costly

Gap

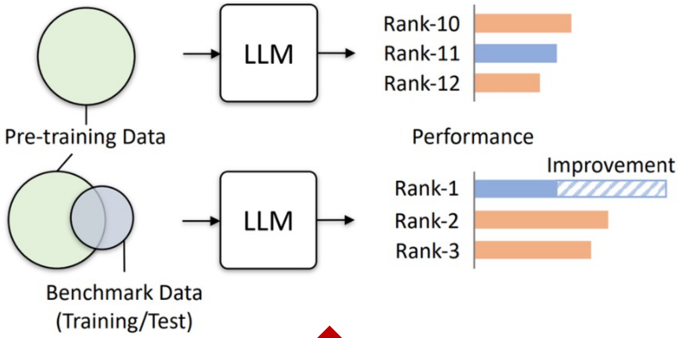


Real-time, focus on cost

RecSys: cost sensitive

Evaluation

LLM: Trained on many data, text-focused, language



Evaluation?

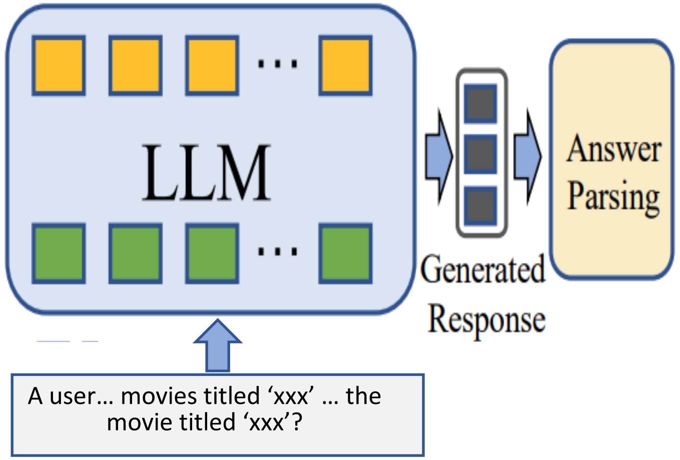
RecSys research: interactions, offline, anonymous data

Modeling: User/Item Representation

- Recommendation: user behavior modeling
- How should we represent user behaviors (represent users/items) in LLM4Rec?

LLM4Rec methods

User/Item: Text

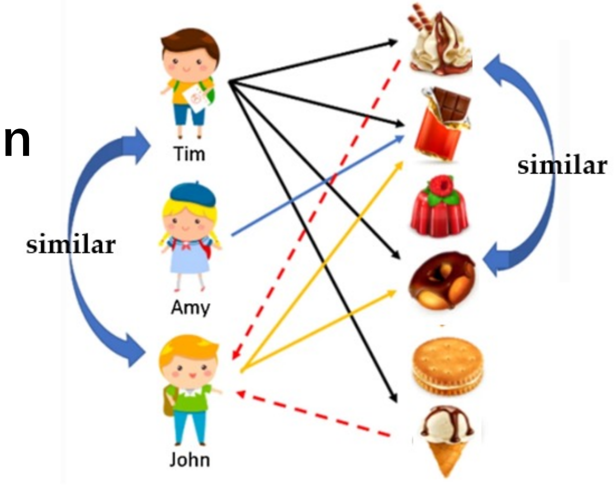


May lack of some information

Textually similar item may have distinct collab. info.

Traditional methods

User/Item: features + ID



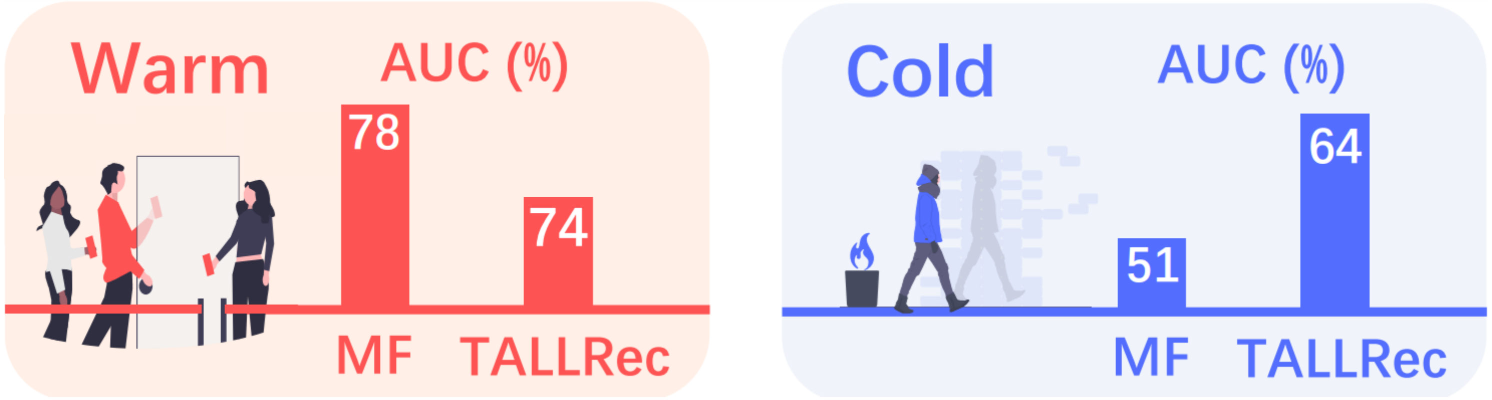
LLMs are constructed using texts, making the representation of users/items in texts the natural choice.

Features (content) alone are insufficient to depict users and items, mainly behavioral similarities (collaborative info). IDs are utilized.

Modeling: User/Item Representation

Integrate collaborative information:

- Why?



LLM Rec vs Traditional CF Model：

#:Excellent at old-start scenarios

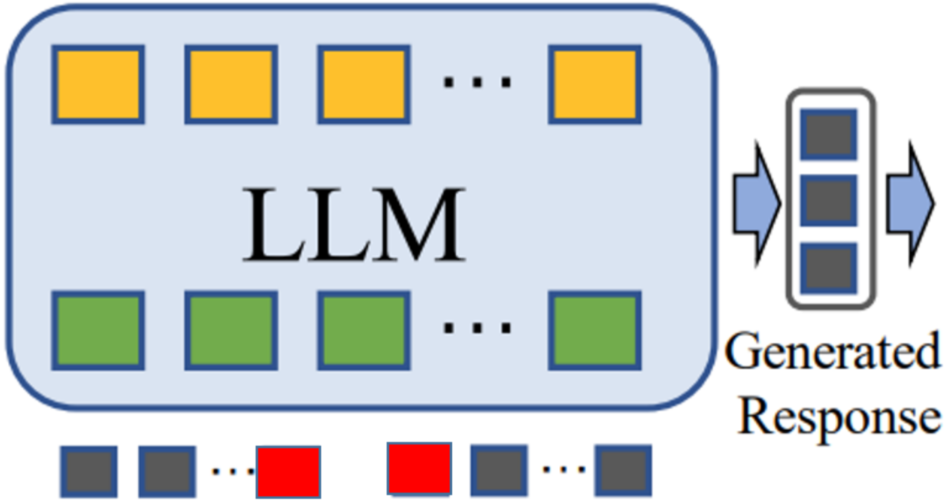
#: Poor at warm-start scenarios

Zhang et al. CoLLM: Integrating Collaborative Embeddings into Large Language Models for Recommendation. ArXiv 2023.

Modeling: User/Item Representation

Integrate collaborative information: How?

#1: learn user/item embedding by fitting interactions



Generate token sequence (new/existing tokens) to represent users and items in LLM

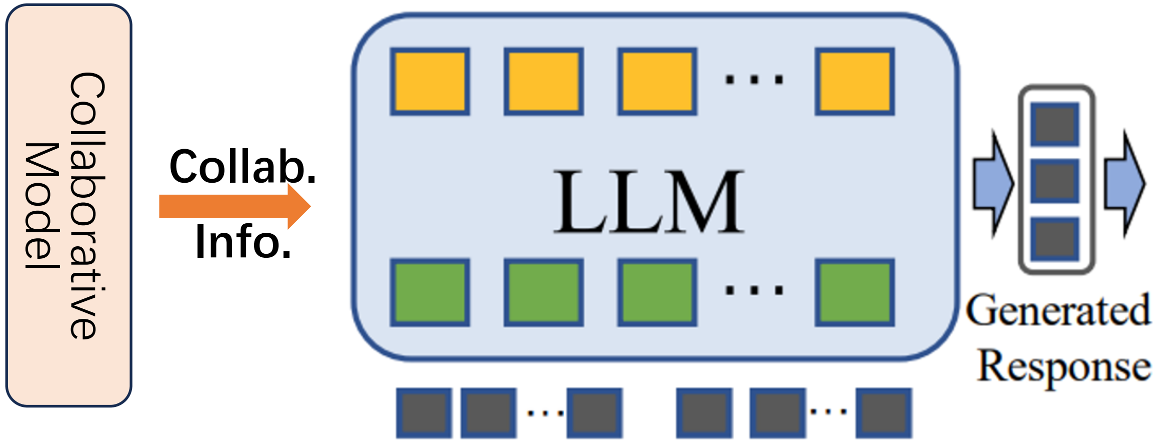
Learn LLM token embeddings by fitting interaction data

⇩

Large space, low learning efficacy

Design better tokenization

#2：Feed the collaborative information extracted by external models into LLM



Extract collaborative information with traditional models

Feed the extracted information into LLMs

⇩

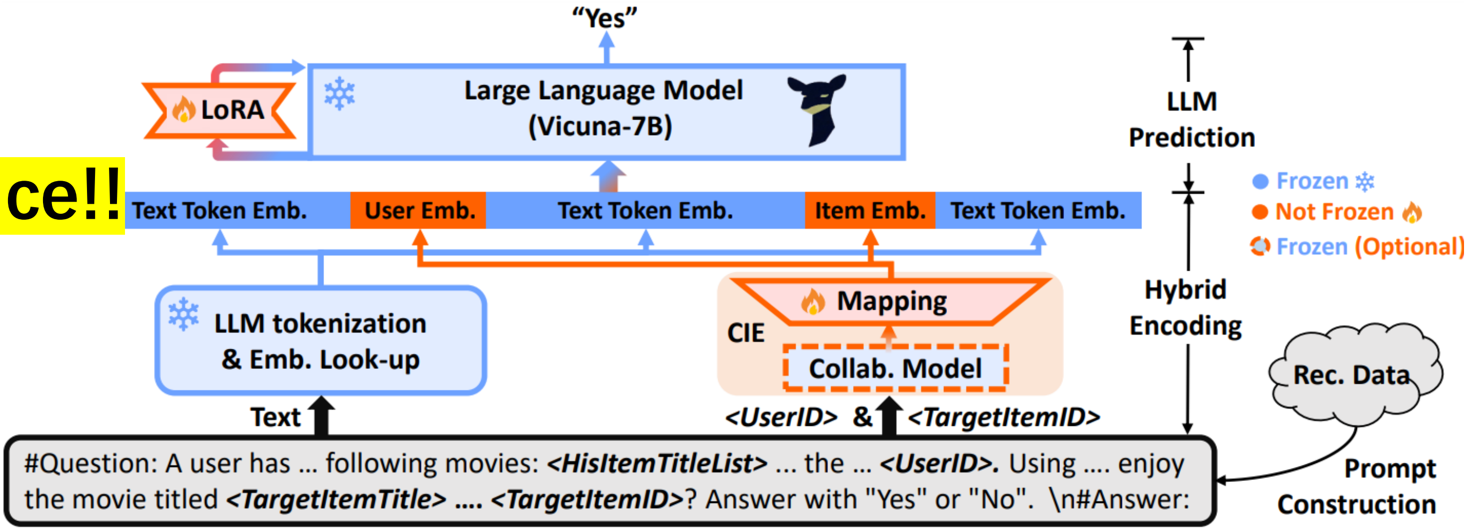
Focus on how to feed the info.

Modeling: User/Item Representation

Integrate collaborative information: feed external collaborative information into LLM

- Work#1: CoLLM —— mapping collaborative embeddings into LLM's Latent space

Latent space!!



#Question: A user has ... following movies: *<HisItemTitleList>* ... the ... *<UserID>*. Using .... enjoy the movie titled *<TargetItemTitle>* .... *<TargetItemID>*? Answer with "Yes" or "No". \n#Answer:

- Prompt construction: add <UserID> and <TargetID> for placing the Collab. Info.
- Hybrid Encoding:
  - text: tokenization & LLM emb Lookup;
  - user/item ID: CIE --- extract info with collab. model (low rank), then map it to the token embedding space
- LLM prediction: add a LoRA module for recommendation task learning

Zhang et al. CoLLM: Integrating Collaborative Embeddings into Large Language Models for Recommendation. ArXiv 2023.

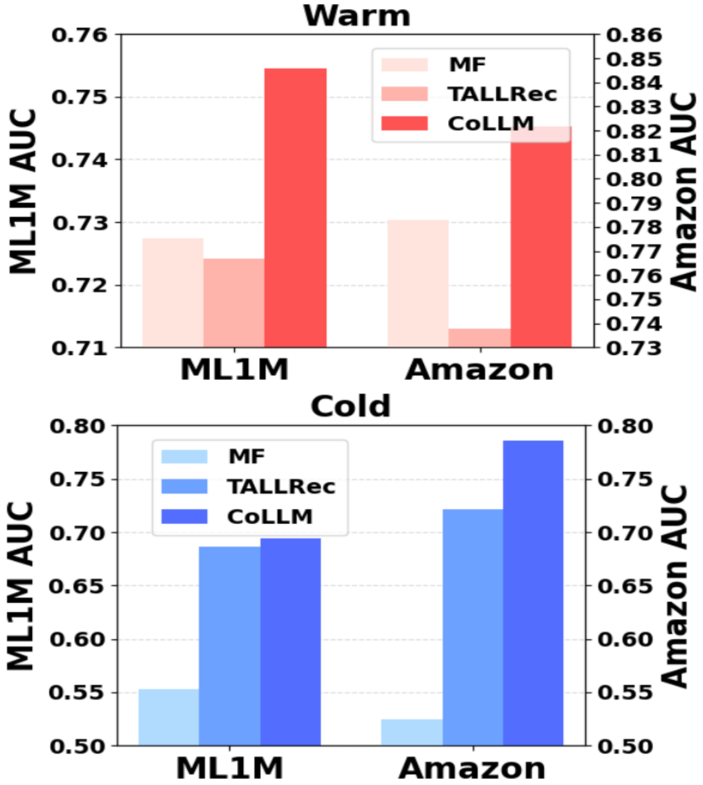
Modeling: User/Item Representation

Integrate collaborative information: feed external collaborative information into LLM

- Work#1: CoLLM —— mapping collaborative embeddings into LLM's Latent space

Overall Performance

Dataset		ML-1M			Amazon-Book		
Methods		AUC	UAUC	Rel. Imp.	AUC	UAUC	Rel. Imp.
Collab.	MF	0.6482	0.6361	10.3%	0.7134	0.5565	12.8%
	LightGCN	0.5959	0.6499	13.2%	0.7103	0.5639	10.7%
	SASRec	0.7078	0.6884	1.9%	0.6887	0.5714	8.4%
LLMRec	ICL	0.5320	0.5268	33.8%	0.4820	0.4856	48.2%
	Soft-Prompt	0.7071	0.6739	2.7%	0.7224	0.5881	10.4%
	TALLRec	0.7097	0.6818	1.8%	0.7375	0.5983	8.2%
Ours	CoLLM-MF	0.7295	0.6875	-	0.8109	0.6225	-
	CoLLM-LightGCN	0.7100	0.6967	-	0.7978	0.6149	-
	CoLLM-SASRec	0.7235	0.6990	-	0.7746	0.5962	-



- CoLLM brings performance improvements over traditional models and current LLM Rec in most cases

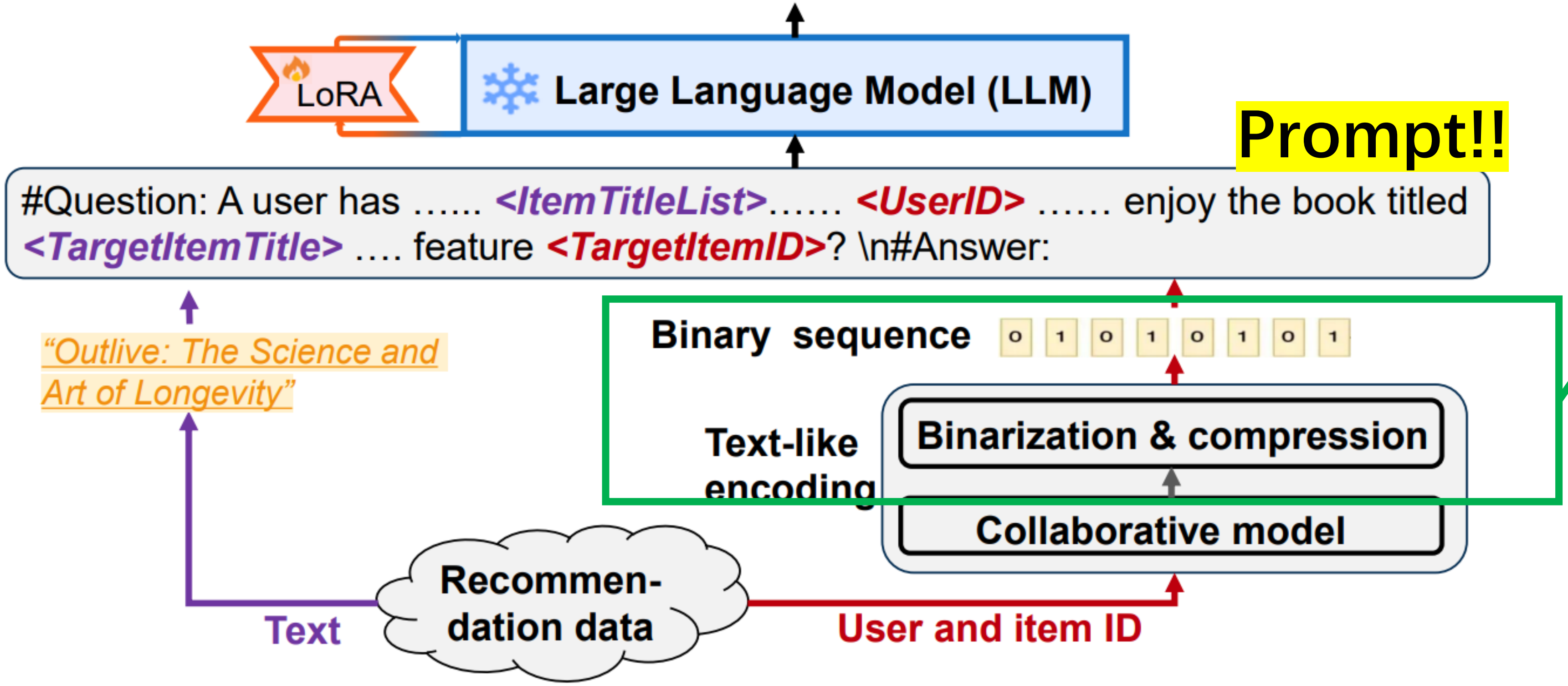
- CoLLM significantly improves the warm performance of LLM4Rec, while ensuring cold performance

Zhang et al. CoLLM: Integrating Collaborative Embeddings into Large Language Models for Recommendation. ArXiv 2023.

Modeling: User/Item Representation

Integrate collaborative information: feed external collaborative information into LLM

- **Work#2: BinLLM —— Encoding collaborative embeddings in a text-like format for LLM**


LoRA
Large Language Model (LLM)
Prompt!!
#Question: A user has …... <ItemTitleList>…... <UserID> …... enjoy the book titled <TargetItemTitle> …. feature <TargetItemID>? \n#Answer:
"Outlive: The Science and Art of Longevity"
Binary sequence
0 1 0 1 0 1 0 1
Text-like encoding
Binarization & compression
Collaborative model
Recommen-dation data
Text
User and item ID

transform the collaborative embeddings into binary sequence, treating them as textual features directly usable by LLMs

- LLMs could naturally perform bitwise operations
- Binarizing collaborative embeddings could keep performance.

Feed collaborative information into prompts

Anonymous. Text-like Encoding of Collaborative Information in Large Language Models for Recommendation. ACL ARR preprint.

Modeling: User/Item Representation

Integrate collaborative information: feed external collaborative information into LLM

- **More works**

[1] Liao et al. Large Language-Recommendation Assistant. ArXiv 2023.

[2] Yang et al. Large Language Model Can Interpret Latent Space of Sequential Recommender. ArXiv 2023.

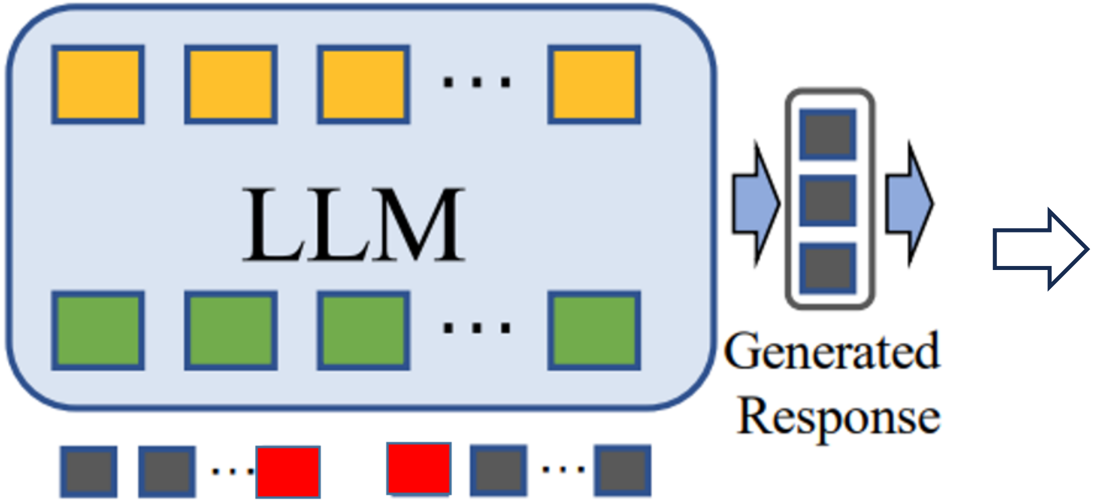
[3] Yu et al. "RA-Rec: An Efficient ID Representation Alignment Framework for LLM-based Recommendation." arXiv 2024.

[4] Li et al. "E4SRec: An elegant effective efficient extensible solution of large language models for sequential recommendation." arXiv 2023.

Modeling: User/Item Representation

Integrate collaborative information: learn user/item-specific token embedding

learn user/item embedding by fitting interactions

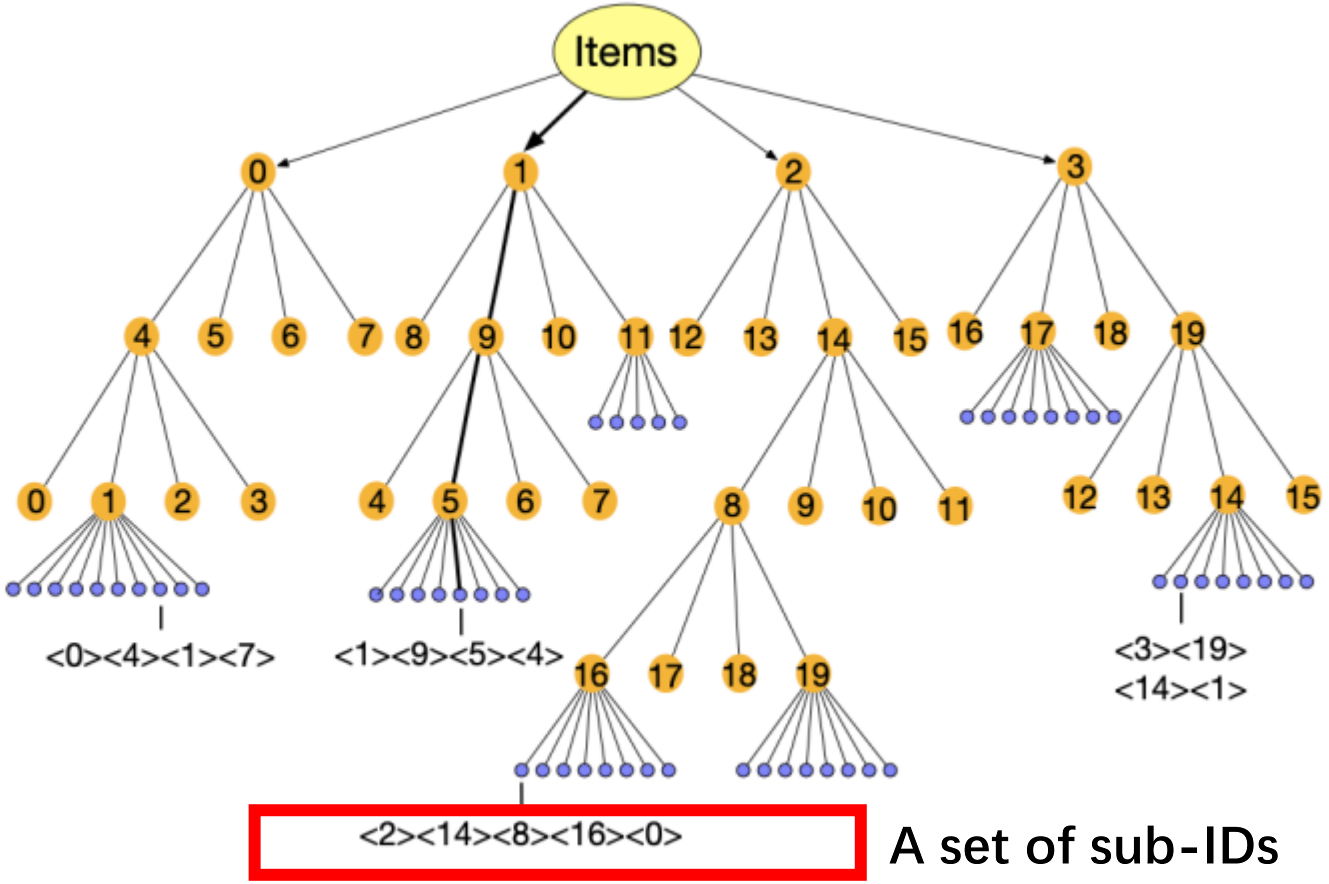


- Generate token sequence (new/existing tokens) to represent users and items in LLM
- Learn LLM token embeddings by fitting interaction data
- Default choice: Random IDs as tokens
- Issues:
  - Large learning space --- low learning efficacy
  - Semantic gaps between text tokens and recommendation tokens
  - Generalization issues --- cannot deal with new items

Modeling: User/Item Representation

Integrate collaborative information: learn user/item-specific token embedding

- Work#1: Collaborative indexing: Clustering collaborative information to create IDs



A set of sub-IDs

- Construct item co-occurrence matrix
- Hierarchically cluster the factorized Laplacian matrix
- generate IDs based on cluster indices

Advantages:

1) Add constraints on item IDs

2) Reduce the token spaces

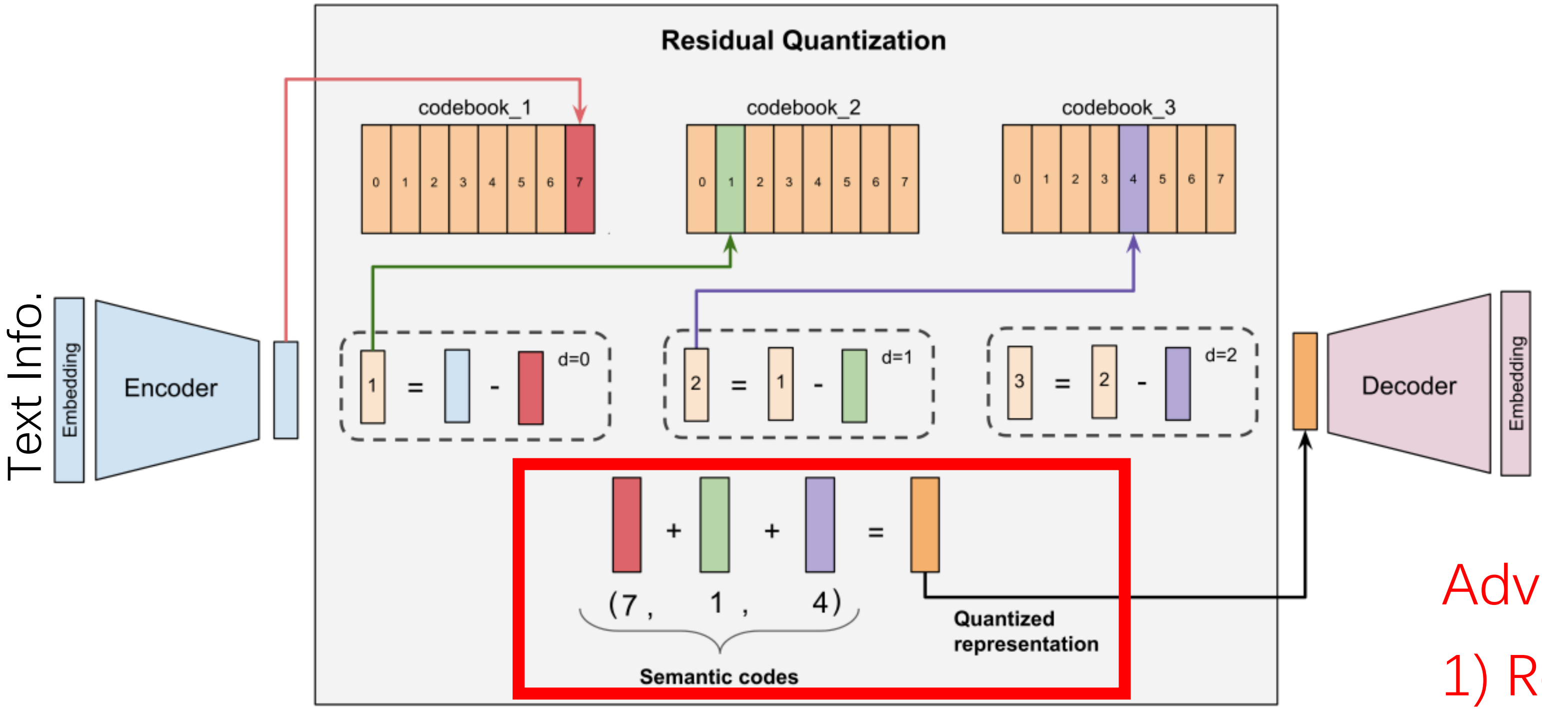
Increase the learning efficacy.

[1] Hua et.al. How to Index Item IDs for Recommendation Foundation Models. SIGIR-AP 2023.

Modeling: User/Item Representation

Integrate collaborative information: learn user/item-specific token embedding

- Work#2: Semantic-aware ID (Tiger/LC-Rec): quantizing text embedding to generate IDs



Quantization: RQ-VAE

- Convert text content information into embeddings
- Quantization: represent the text embedding with several sub-embeddings, generating semantic ID
- Several sub-IDs form a semantic ID

Advantages:

1) Reduce the token spaces, $N \rightarrow K \cdot N^{1/K}$ 2) Could deal with new items

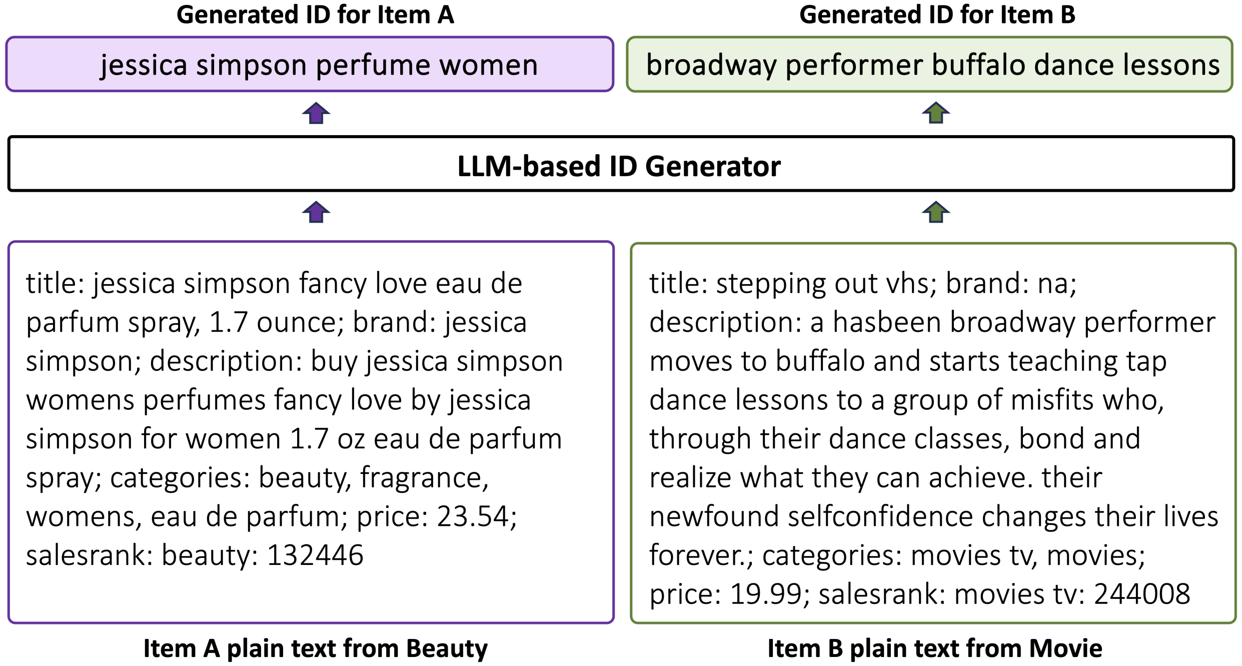
[1] Zheng et.al. Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. ICDE 2024.

[2] Rajput el.al. Recommender Systems with Generative Retrieval. NeurIPS 2023.

Modeling: User/Item Representation

Integrate collaborative information: learn user/item-specific token embedding

- Work#3: **Textual ID (IDGenRec): generate textual ID from the item's textual metadata**


Generated ID for Item A
jessica simpson perfume women
Generated ID for Item B
broadway performer buffalo dance lessons
LLM-based ID Generator
Item A plain text from Beauty
Item B plain text from Movie

title: jessica simpson fancy love eau de parfum spray, 1.7 ounce; brand: jessica simpson; description: buy jessica simpson womens perfumes fancy love by jessica simpson for women 1.7 oz eau de parfum spray; categories: beauty, fragrance, womens, eau de parfum; price: 23.54; salesrank: beauty: 132446

title: stepping out vhs; brand: na; description: a hasbeen broadway performer moves to buffalo and starts teaching tap dance lessons to a group of misfits who, through their dance classes, bond and realize what they can achieve. their newfound selfconfidence changes their lives forever.; categories: movies tv, movies; price: 19.99; salesrank: movies tv: 244008

Textual ID generator: text-to-text

- **ID Generator**
  - Input plain text of item information into fine-tuned tag generator
  - Generate several short, informative, and unique tags in natural language

Advantages:

1) composed of existing tokens used in LLMs' pre-training, thus harnessing LLMs semantic knowledge
2) Generalize to new items

[1] Tan et.al. Towards LLM-RecSys Alignment with Textual ID Learning. SIGIR 2024.

Modeling: User/Item Representation

Integrate collaborative information: learn user/item-specific token embedding

- **Summary of tokenizer (item-side):**

1. Textual Identifiers
   - BigRec
   - GenRec
   - IDGenRec

↓

Limited collab. info.

Could deal with new items

2. Vanilla ID
   - Random ID

↓

Large space, hard to learn

Good at collab. info.

Cannot deal with new items

3. Collaborative ID
   - CID

↓

Small space

Good at collab. info.

Cannot deal with new items

4. Semantic ID
   - TIGER
   - LC-Rec

↓

Small space

Limited collaborative info.

Could deal with new items

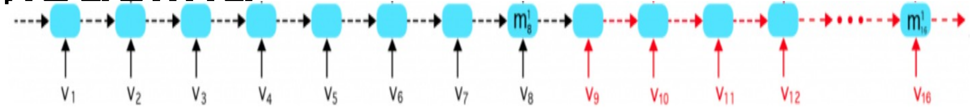
- **Open Problem:**
  - Tokenization on user behaviors
  - Tokenization on cross-domain items

Wang et al. 2024. Learnable Tokenizer for LLM-based Generative Recommendation. In arxiv 2024.

Modeling: Lifelong Modeling

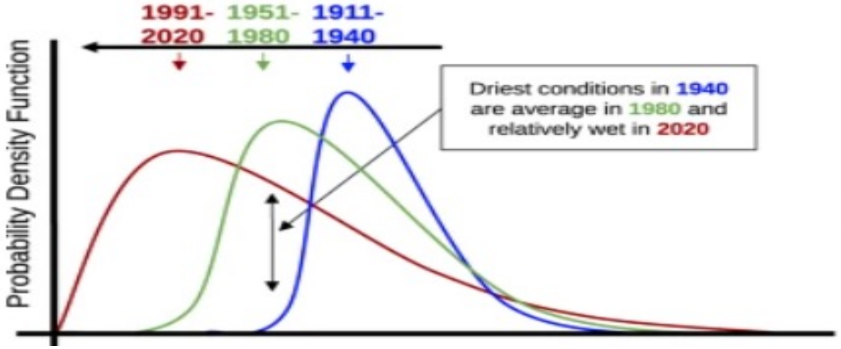
- Users are anticipated to engage with the recommender system continuously
- Raise the need of lifelong behavior modeling for users

Lifelong sequential behavior modeling



- The length of historical interaction sequences grows significantly, easily exceeding 1000
- How to model such long sequence effectively?

Continual learning

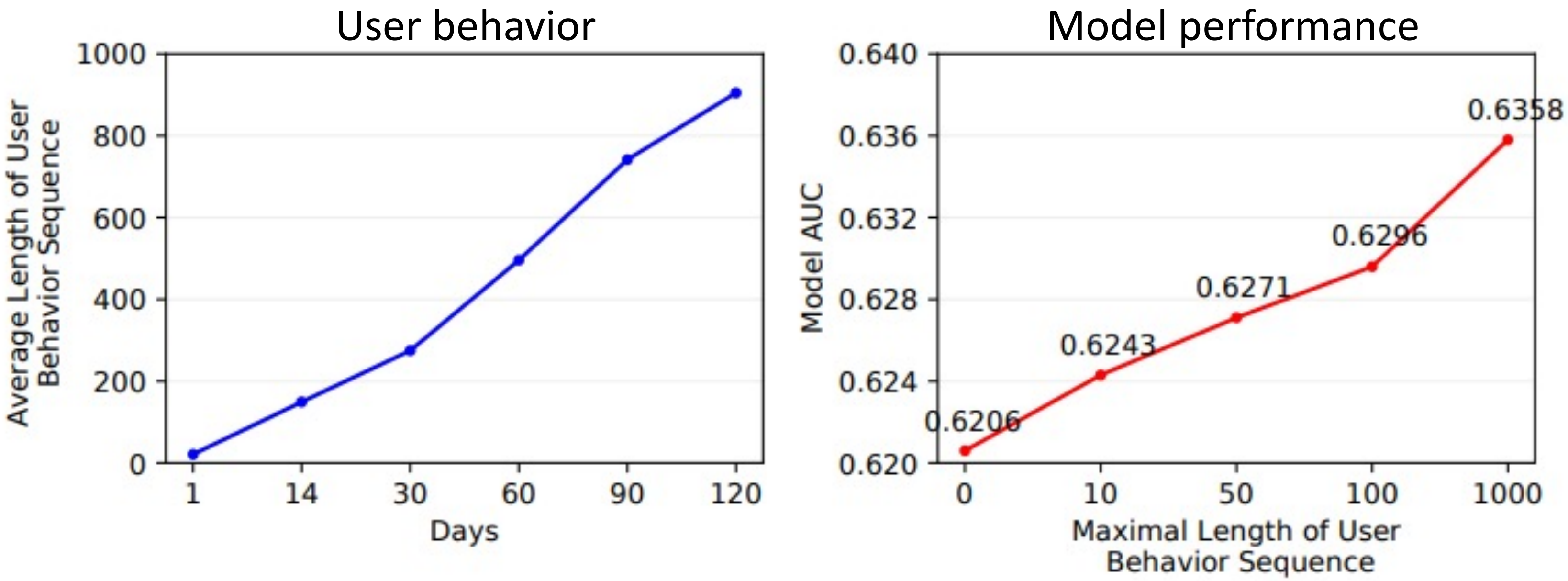


- User interests drift with time going
- How to continuously/incremental learn user interests?

Modeling: Lifelong Modeling

Lifelong sequential behavior modeling：

- **A longer history signifies richer personalization information, and modeling this can lead to heightened prediction accuracy.**



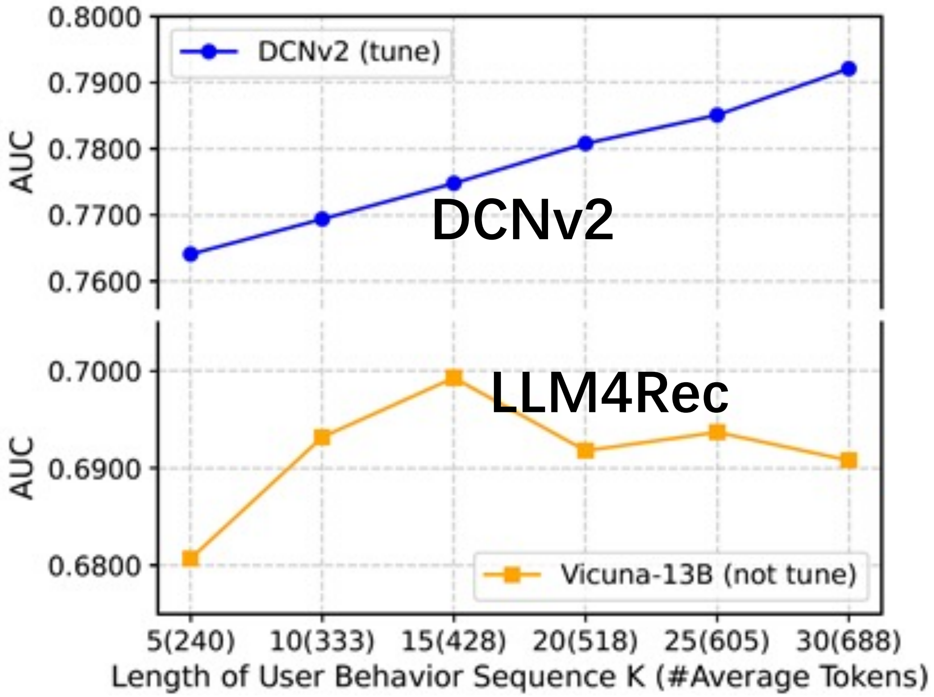
An example in the advertising system in Alibaba.

Qi Pi et al. Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction. In KDD 2019.

Modeling: Lifelong Modeling

Lifelong sequential behavior modeling：

LLM cannot effectively model long user Behavior sequence



- Extending user behavior sequences doesn't necessarily enhance recommendation performance, even if the input length is far below the length limit of LLMs (e.g., Vicuna-13B has an upper limit of 2048 tokens).

Li et al. ReLLa: Retrieval-enhanced Large Language Models for Lifelong Sequential Behavior Comprehension in Recommendation. WWW 2024.

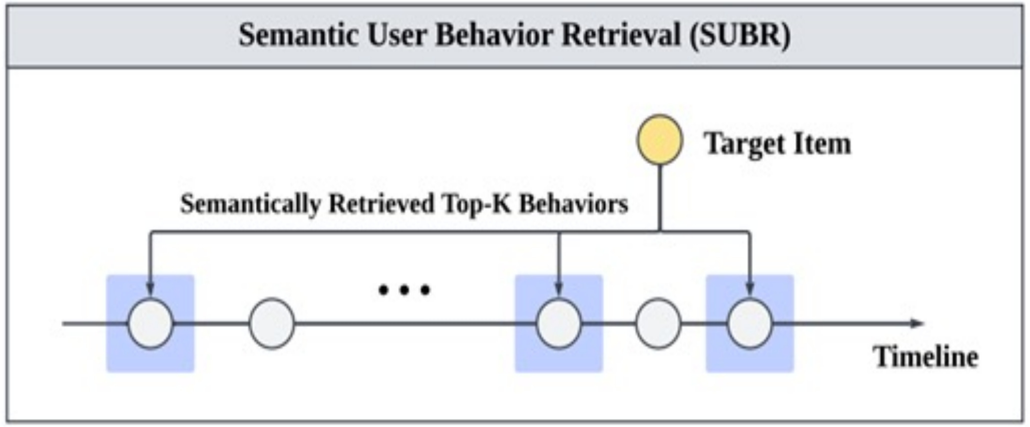
Modeling: Lifelong Modeling

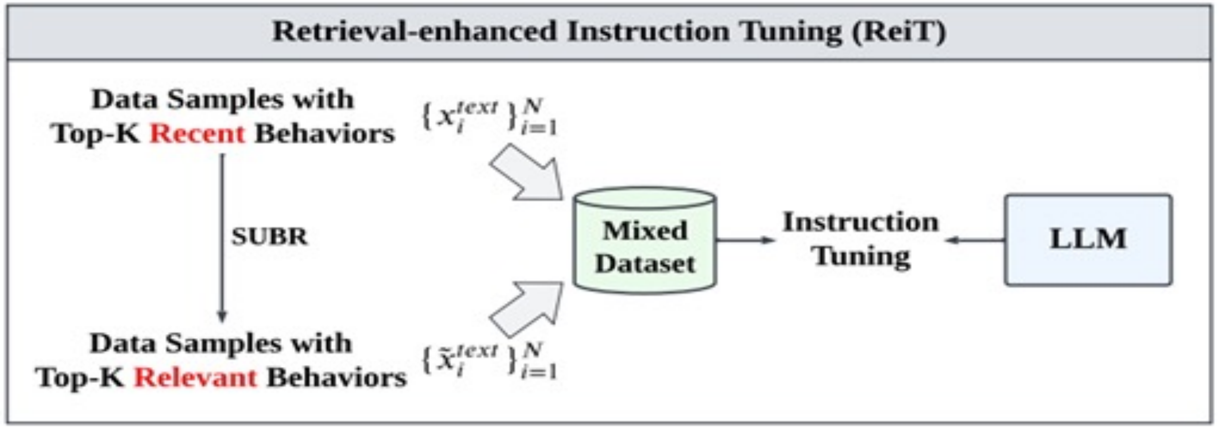
Lifelong sequential behavior modeling：

- **Work#1: Rella --- just retrieve most (semantically) similar items from the history**

Step1: For a target item, retrieve the top-K semantically similar items from the history, forming a new sample

Step2: Leverage the original sample and new sample to fine tune LLM for recommendation


Semantic User Behavior Retrieval (SUBR)
Target Item
Semantically Retrieved Top-K Behaviors
• • •
Timeline


Retrieval-enhanced Instruction Tuning (ReiT)
Data Samples with Top-K Recent Behaviors
$\{x_i^{text}\}_{i=1}^N$
SUBR
Data Samples with Top-K Relevant Behaviors
$\{\tilde{x}_i^{text}\}_{i=1}^N$
Mixed Dataset
Instruction Tuning
LLM

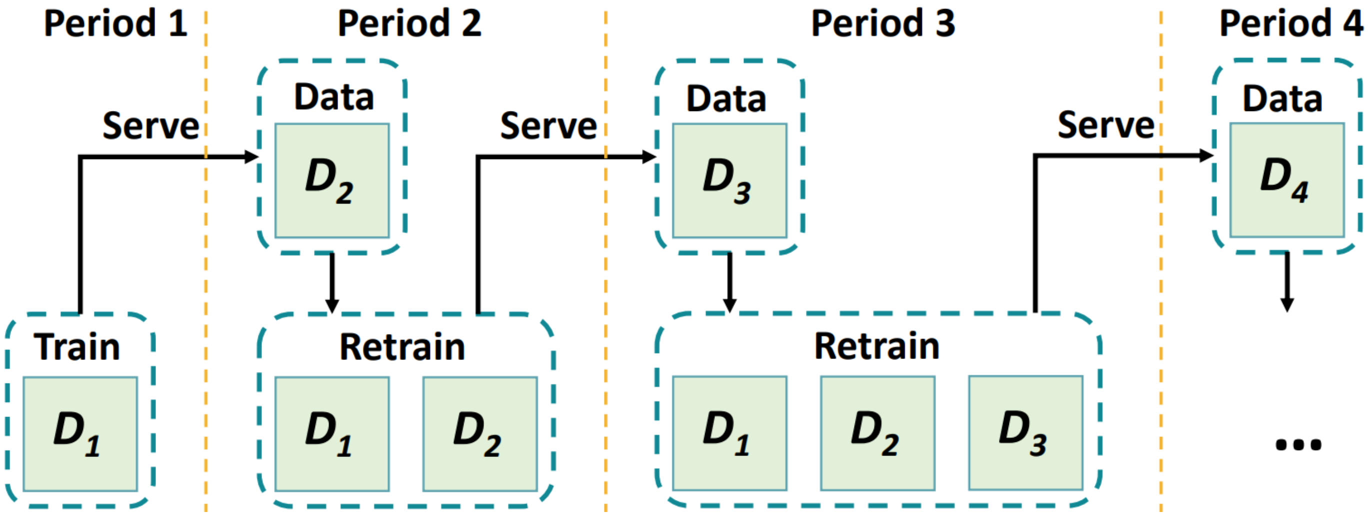
- Limitations: heavily depends on “target attention, not applicable when the input lacks target items.
- Future: may need to explore other solutions like memory.

Li et al. ReLLa: Retrieval-enhanced Large Language Models for Lifelong Sequential Behavior Comprehension in Recommendation. WWW, 2024.

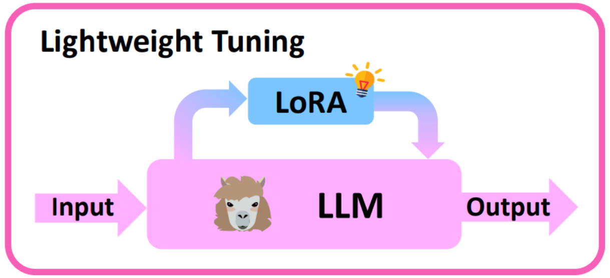
Modeling: Lifelong Modeling

Continual learning：

- How to incrementally learn user interests?
- There is work [1] studying the common used methods: periodic retraining



Just retrain LoRA (TALLRec)

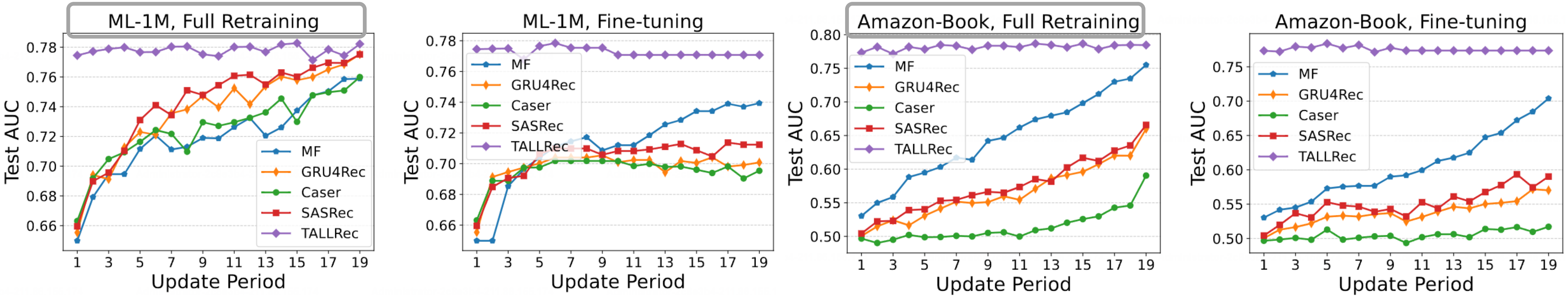


[1] Shi et al. Preliminary Study on Incremental Learning for Large Language Model-based Recommender Systems. In arXiv 2023.

Modeling: Lifelong Modeling

Continual learning：

Work#1: The effectiveness of full-retraining and fine-tuning for TALLRec



- ❑ Periodically update TALLRec does not bring significant performance improvements.
- ❑ LLM4Rec may struggle to capture short-term preferences in the latest data with traditional periodic updates, limiting performance improvement.

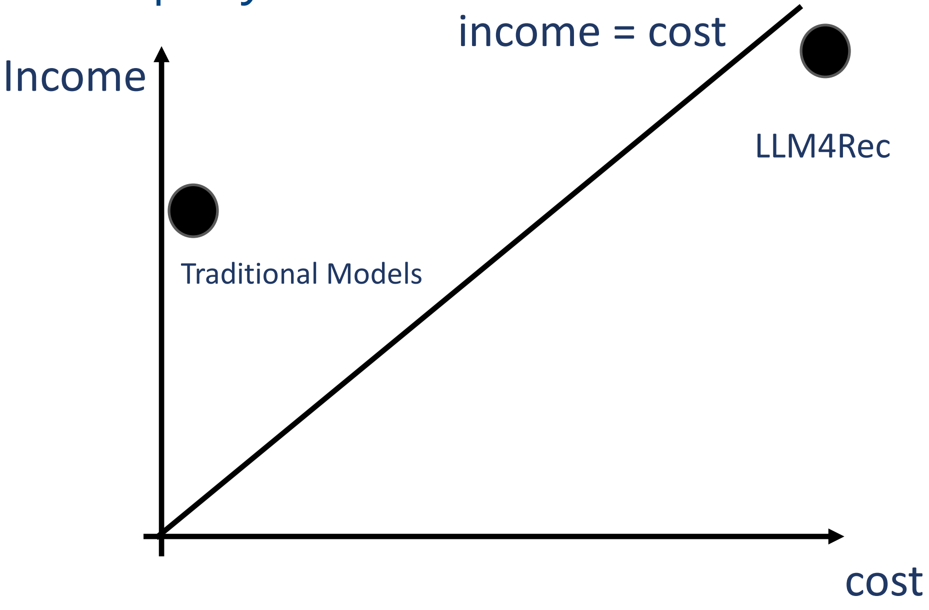
[1] Shi et al. Preliminary Study on Incremental Learning for Large Language Model-based Recommender Systems. In arXiv 2023.

Outline

- Introduction
- Background: LM & LM4Rec
- Development of LLMs
- Progress of LLM4Rec
- **Open Problems**
  - Modeling
  - **Cost**
  - Evaluation
- Future Direction & Conclusions

Cost

- The income-cost trade-off is sensitive for recommendation
- Deployment cost of LLM4Rec is high



LLM Parameters: tens/hundreds of billions

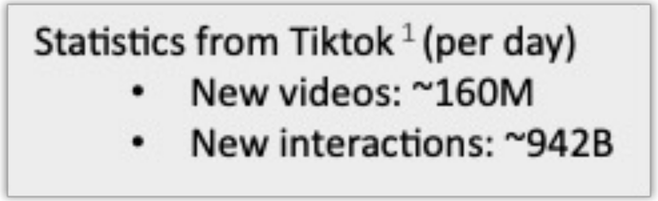
Training and inference:

- High demand on GPUs/Memory
- Slow

How to reduce the cost?

Cost: Training

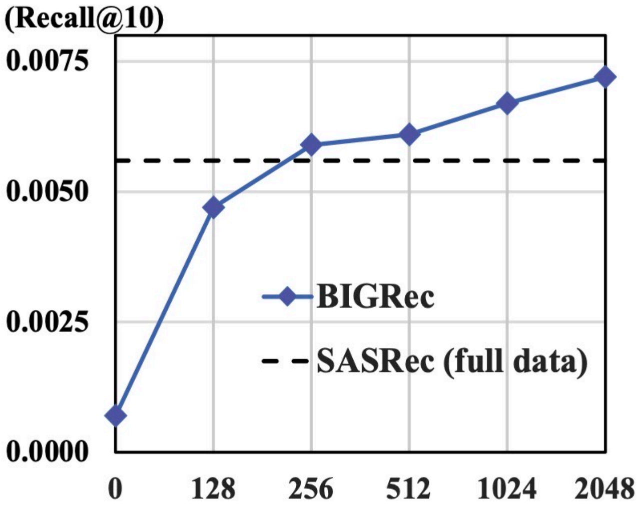
One exploration: Data-efficient training



Statistics from Tiktok [1] (per day)

- New videos: ~160M
- New interactions: ~942B

- ❑ Fine-tuning LLM is necessary
    - ❑ LLMs are not particularly trained on recommendation data
- ❑ LLM fine-tuning is expensive, e.g., high computational costs, time-consuming
- ❑ Few-shot fine-tuning is a promising solution
- ❑ Data pruning for efficient LLM-based recommendation
    - ❑ identify representative samples tailored for LLMs



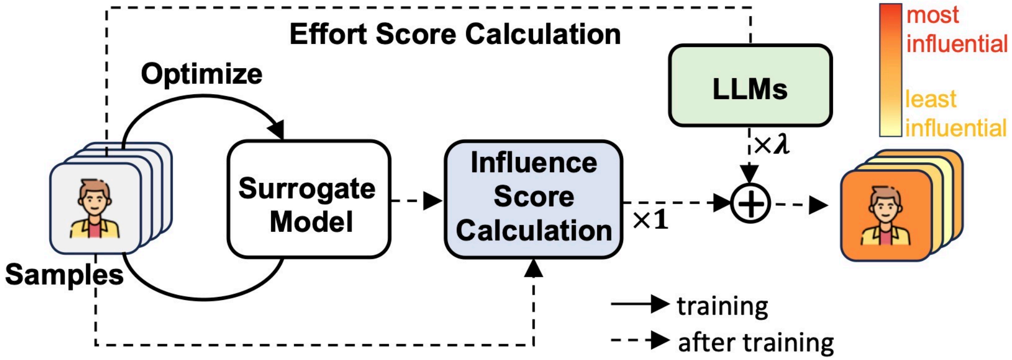
(a) Few-shot performance on MicroLens-50K.

Lin et al. Data-efficient fine-tuning for LLM-based recommendation. SIGIR'24.

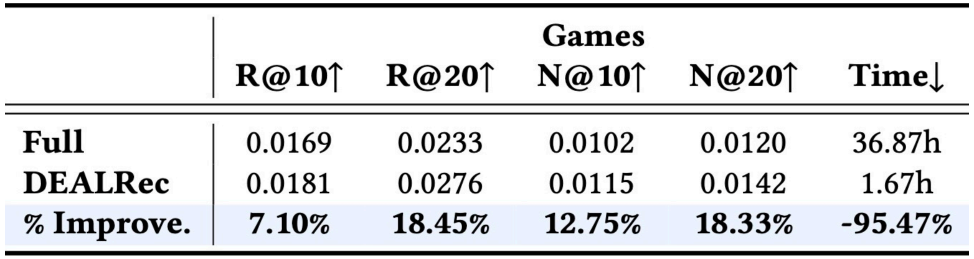
Cost: Training

One exploration: Data-efficient training

- ❑ Two objectives for data pruning
  - ❑ high accuracy: select the samples that can lead to higher performance -> influence score
  - ❑ high efficiency: emphasize the low costs of the data pruning process
    - ❑ surrogate model to improve efficiency
    - ❑ effort score to bridge between surrogate model and LLMs

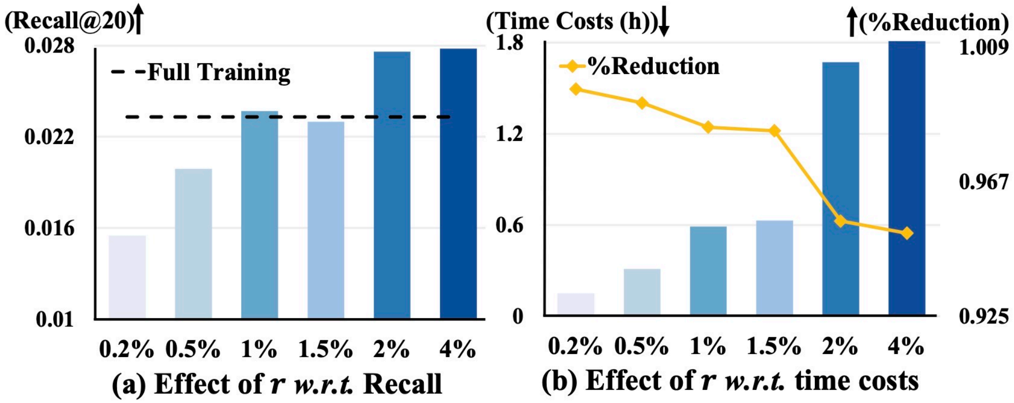


- ❑ Experimental results
  - ❑ fine-tune with 1024 samples



	Games				
	R@10↑	R@20↑	N@10↑	N@20↑	Time↓
Full	0.0169	0.0233	0.0102	0.0120	36.87h
DEALRec	0.0181	0.0276	0.0115	0.0142	1.67h
% Improve.	**7.10%**	**18.45%**	**12.75%**	**18.33%**	**-95.47%**

  - ❑ Increasing samples from 0.2% to 4% of all training data



(a) Effect of *r w.r.t.* Recall

(b) Effect of *r w.r.t.* time costs

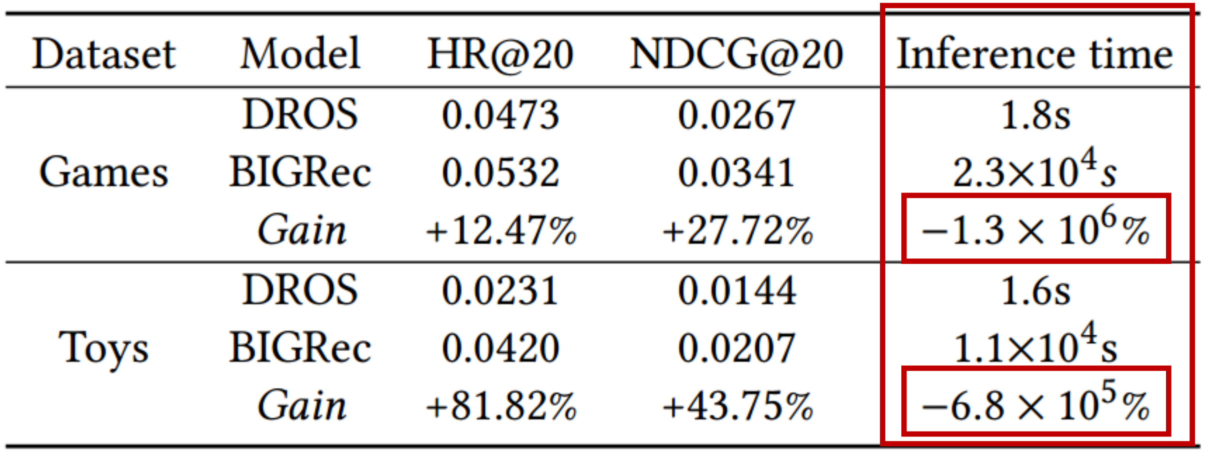
Lin et al. Data-efficient fine-tuning for LLM-based recommendation. SIGIR'24.

Cost: Inference

One solution: distillation

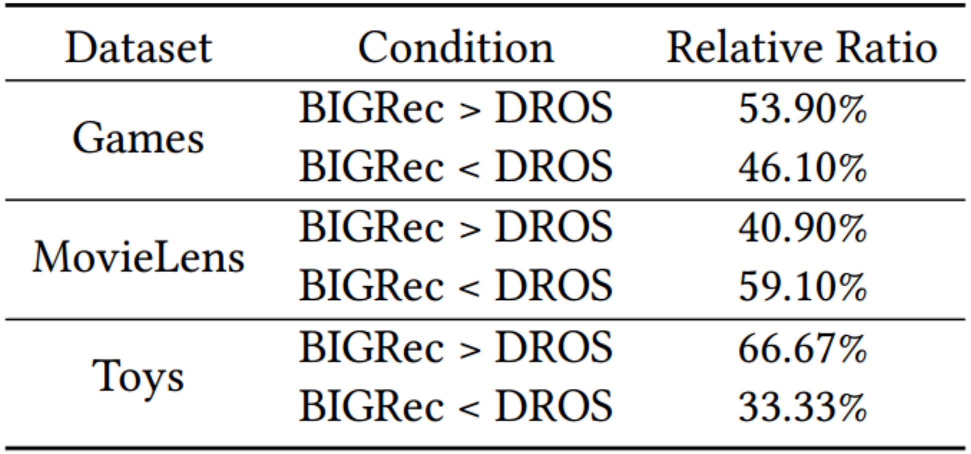
Distill LLM’s knowledge to smaller models and utilize small models for inference

- Work#1: distill recommendation results



Dataset	Model	HR@20	NDCG@20	Inference time
Games	DROS	0.0473	0.0267	1.8s
	BIGRec	0.0532	0.0341	2.3×10^4s
	Gain	+12.47%	+27.72%	-1.3×10^6%
Toys	DROS	0.0231	0.0144	1.6s
	BIGRec	0.0420	0.0207	1.1×10^4s
	Gain	+81.82%	+43.75%	-6.8×10^5%

The inference latency of BIGRec far exceeds that of DROS.



Dataset	Condition	Relative Ratio
Games	BIGRec > DROS	53.90%
	BIGRec < DROS	46.10%
MovieLens	BIGRec > DROS	40.90%
	BIGRec < DROS	59.10%
Toys	BIGRec > DROS	66.67%
	BIGRec < DROS	33.33%

BIGRec does not always outperform DROS.

❑ Distillation challenges:

❑ 1）The teacher’s knowledge may not always be reliable.

❑ 2）The divergence in semantic space poses a challenge to distill the knowledge from embeddings.

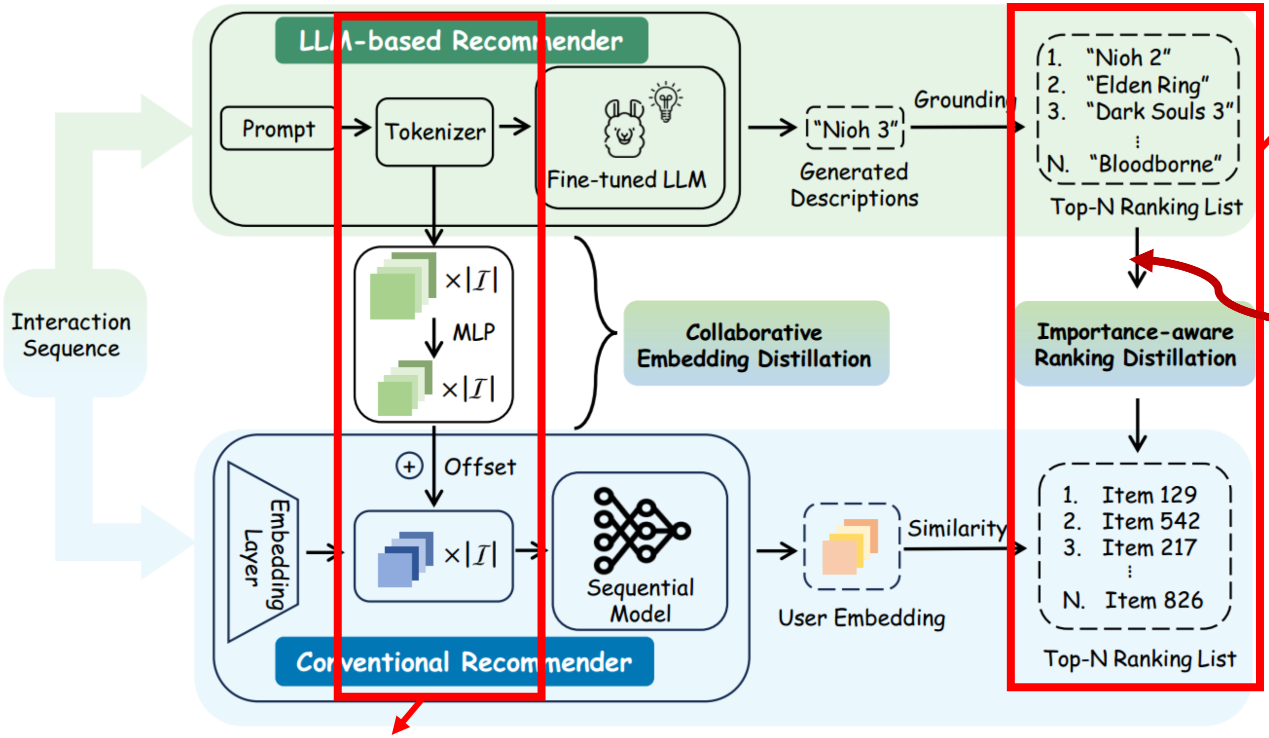
Yu Cui et al. “Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Model” arXiv 2024

Cost: Inference

One solution: distillation

Distill LLM’s knowledge to smaller models and utilize small models for inference

- Work#1: distill recommendation results



Supervise d signals

- Importance-aware Ranking Distillation

  filter reliable and student-friendly knowledge

  by weighting instances

  - Confidence of LLMs

    The distance between the generated

    descriptions with the target item

  - Teacher-Student Consensus

    The items recommended by both teacher

    and student are more likely to be positive

  - Ranking Position

    Higher ranked items by teachers are more

    reliable

Collaborative Embedding Distillation

integrate knowledge from teacher embeddings with

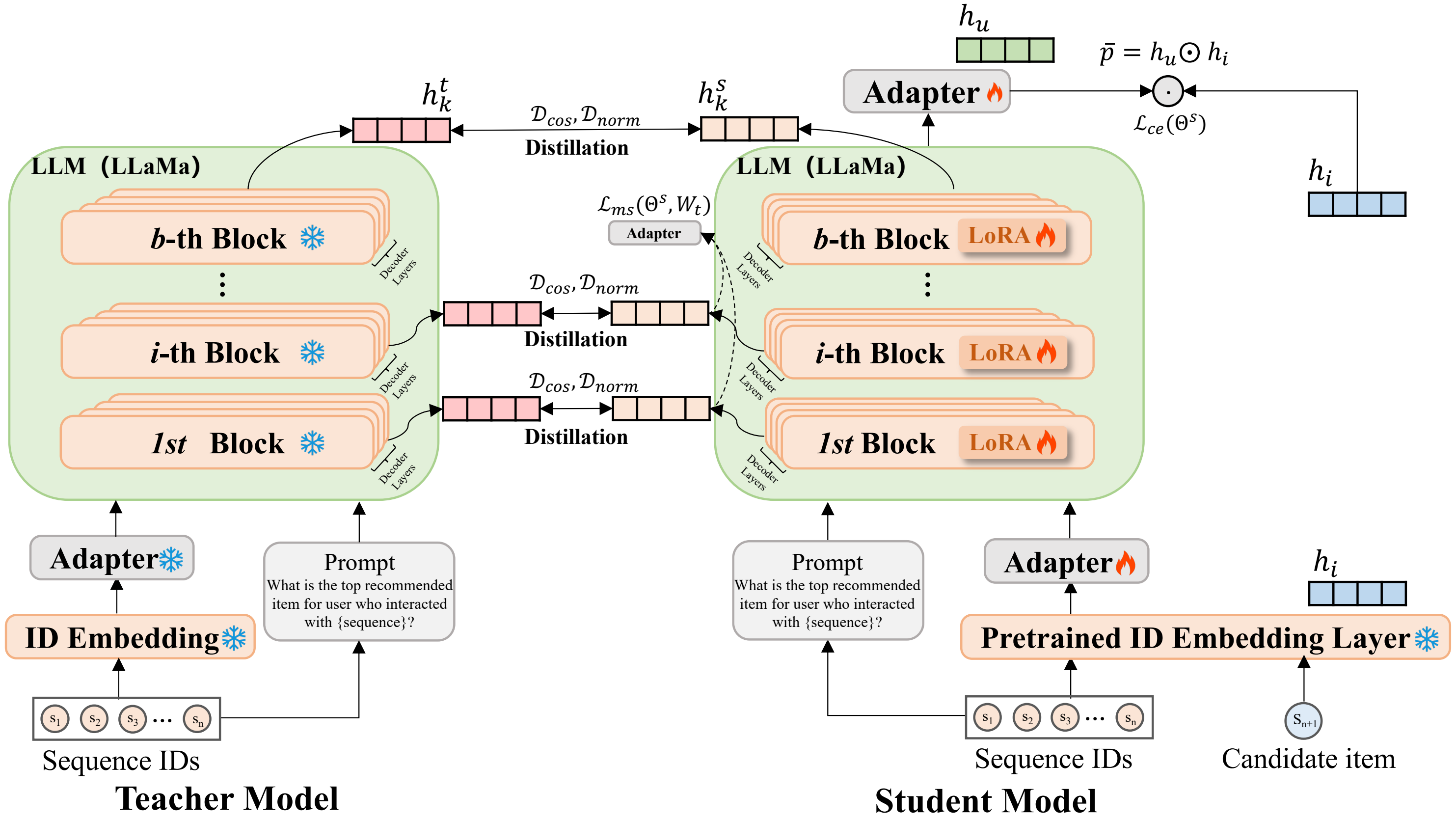
student’ s Yu Cui et al. “Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Model” arXiv 2024

Cost: Inference

One solution: distillation

Distill LLM’s knowledge to smaller models and utilize small models for inference

- Work#2: layer-wise knowledge distillation for smaller LM



h_u
$\bar{p} = h_u \odot h_i$
Adapter
$\mathcal{L}_{ce}(\Theta^s)$
h_i
h_k^t
h_k^s
$\mathcal{D}_{cos}, \mathcal{D}_{norm}$
Distillation
LLM (LLaMa)
LLM (LLaMa)
$\mathcal{L}_{ms}(\Theta^s, W_t)$
Adapter
b-th Block
b-th Block LoRA
i-th Block
i-th Block LoRA
1st Block
1st Block LoRA
Decoder Layers
$\mathcal{D}_{cos}, \mathcal{D}_{norm}$
Distillation
$\mathcal{D}_{cos}, \mathcal{D}_{norm}$
Distillation
Adapter
Prompt
What is the top recommended item for user who interacted with {sequence}?
ID Embedding
Prompt
What is the top recommended item for user who interacted with {sequence}?
Adapter
Pretrained ID Embedding Layer
h_i
s_1 s_2 s_3 … s_n
Sequence IDs
Teacher Model
s_1 s_2 s_3 … s_n
Sequence IDs
s_{n+1}
Candidate item
Student Model

Motivation:

some layers of LLMs are redundant in the downstream recommendation task

Efficiency:

achieves up to 6.6x/8.0x speedup in terms of training/inference time costs against LLM-based recommendation models

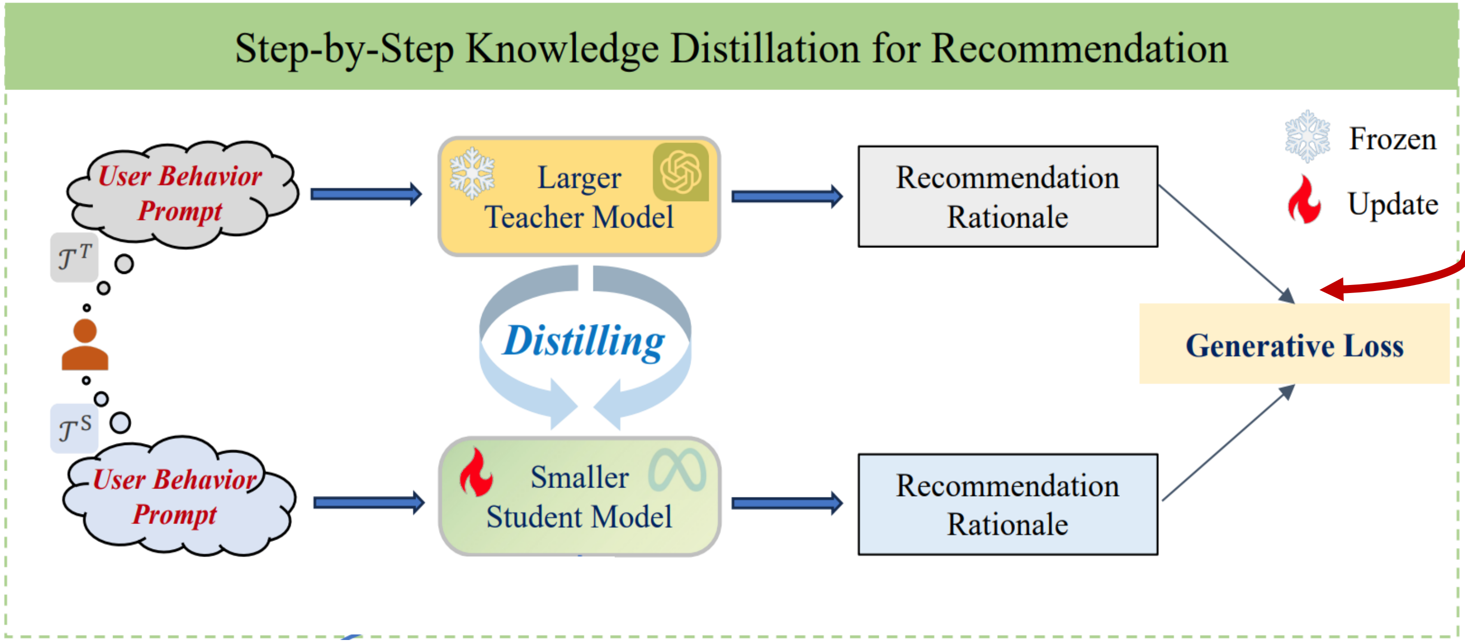
Wujiang Xu et al. “SLMRec: Empowering Small Language Models for Sequential Recommendation” arXiv 2024

Cost: Inference

One solution: distillation

Distill LLM’s knowledge to smaller models and utilize small models for inference

- Work#3: distill recommendation rationales



Supervised signals

- ❑ Distill recommendation rationale from ChatGPT to Llama-7B
- ❑ Empowering recommendation with rationale embedding
- ❑ Combining the rationale embedding and item description embedding for prediction

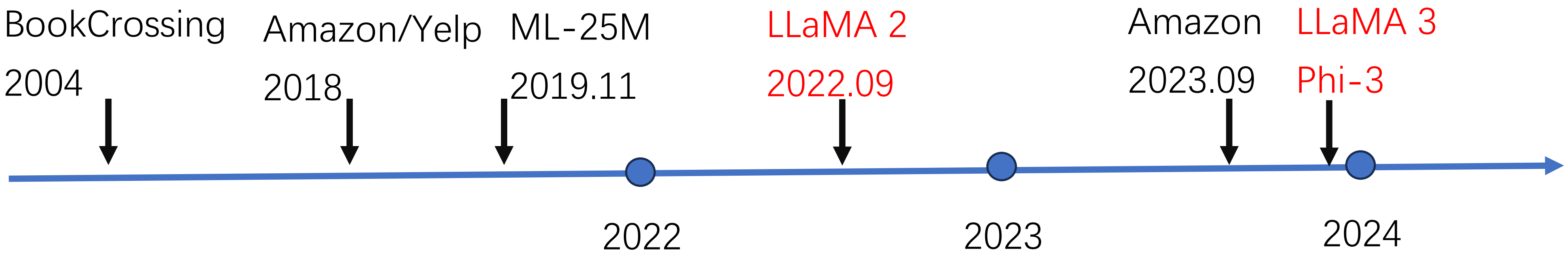
Wang et al., Can Small Language Models be Good Reasoners for Sequential Recommendation? ArXiv 2024.

Outline

- Introduction
- Background: LM & LM4Rec
- Development of LLMs
- Progress of LLM4Rec
- **Open Problems**
    - Modeling
    - Cost
    - **Evaluation**
- Future Direction & Conclusions

Evaluation: Data Issues

- ❑ **Challenge#1: Lack of data for evaluation**
  - ❑ **Most of benchmarks are proposed ahead of pre-training stage of LLMs, e.g., ChatGPT, LLaMA.**



- ❑ The information of recommendation datasets (e.g., reviews,) may be include in LLMs.
- ❑ Existing works usually did not discuss this.
- ❑ Evaluations on the data that is not include in pretraining data of LLMs.

Evaluation: Data Issues

❑ Challenge#1: Lack of data for evaluation

- ❑ Insufficient features
  - ❑ Lack of raw feature
    - ❑ Anonymous (e.g., just feature ID)
    - ❑ Lack of content (e.g., video content)
  - ❑ Currently, many works just utilize titles

- Underutilization of LLM capabilities;
- Underassessment of the effectiveness of LLM4Rec

- ❑ Data homogeneity
  - ❑ content homogeneity:

    mostly from E-commerce platform / entertaining content or places
  - ❑ biased user distributions: mostly from China and U.S.

- Not comprehensive evaluation
- Biased evaluation

Evaluation: Interactive RecSys

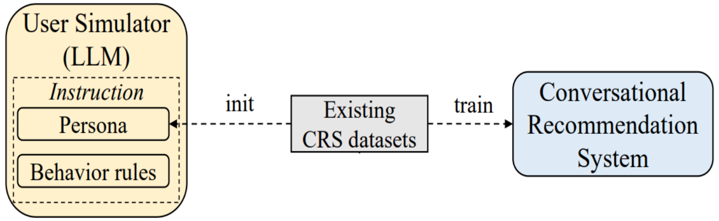
❑ Challenge#2: Evaluate interactive recommendation

❑ Conversational recommendation

- ❑ provide personalized recommendation via multi-turn dialogs in natural language
- ❑ focus on conversational quality and recommendation quality

- Issues of traditional evaluation:
  - Simulated users are overly simplified representations of human users
  - Conversations are often vague about the user preference, but not focus on exactly match the ground-truth item
  - Evaluation protocol is based on fixed conversations, but the conversation could be diverging.

- New evaluation: simulation with LLM-based agents?
  - Challenges: how to design simulators is still an open problem.



Xiaolei Wang et al. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. EMNLP

Evaluation: Interactive RecSys

- ❑ **Challenge#2: Evaluate interactive recommendation**
  - ❑ Long-term recommendation
    - ❑ Multi-turn user-system interactions
    - ❑ Focus on long-term user engagement, e.g., user retention
  - ❑ How to evaluate long-term engagement is a big challenge.
    - ❑ We have not feedback about the unseen interaction trajectory
    - ❑ Evaluation with agent-based simulator is a potential solution

Evaluation: platform and protocol

- ❑ **Platform: OpenP5**
  - ❑ **Develop, train, and evaluate LLM-based recommenders**
    - ❑ Customized **Dataset**
    - ❑ Customized **item indexing** methods
    - ❑ Personalized **prompt** collection
    - ❑ Extensibility of **multi-task learning**
    - ❑ New **backbone** methods

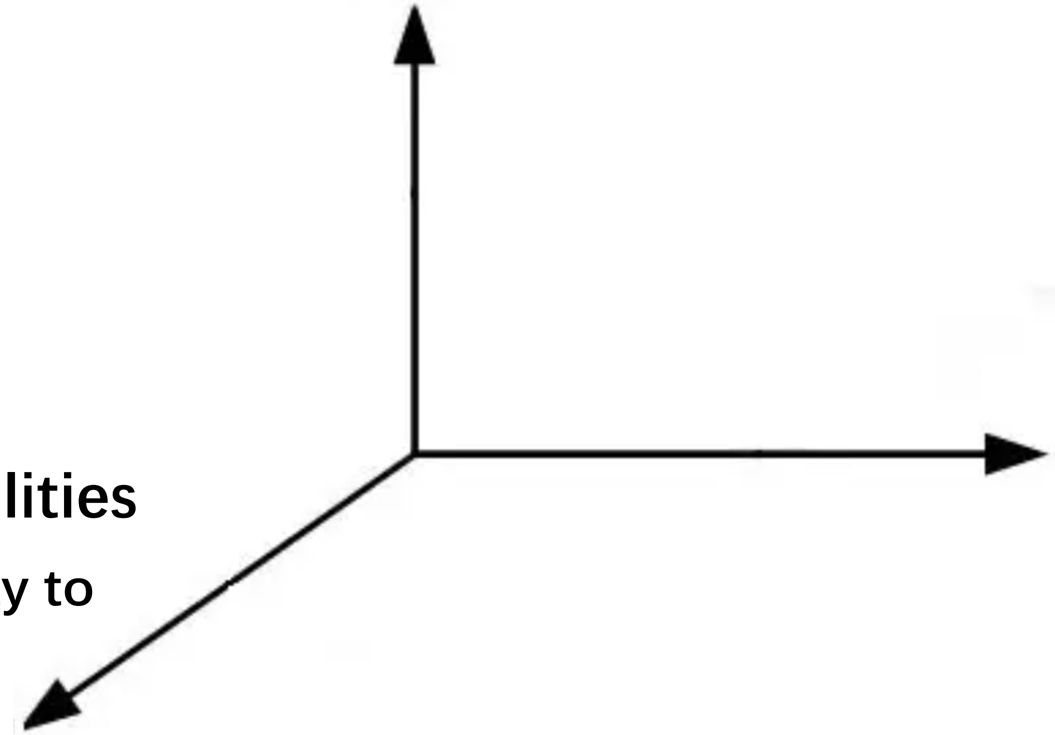
Shuyuan Xu et al. OpenP5: An Open-Source Platform for Developing, Training, and Evaluating LLM-based Recommender Systems. SIGIR'24

Outline

- Introduction
- Background: LM & LM4Rec
- Development of LLMs
- Progress of LLM4Rec
- Open Problems
- **Future Direction & Conclusions**

Progress of LLM4Rec

Three dimensions:



Metrics

From accuracy to trustworthiness

such as privacy, fairness, etc.

How to utilize LLMs

In-context learning

Tuning

Agent

Information modalities

From the text modality to

multiple modalities

Open Problems

Three aspects:

Modeling

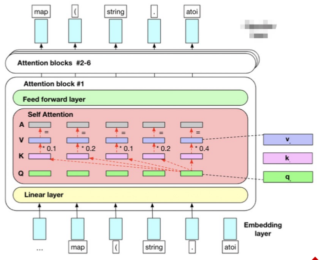
LLM: modeling text/language

Gap

RecSys: modeling behaviors

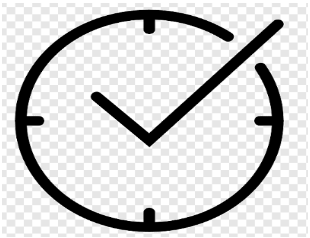
Cost

LLM: high cost/delay



computation/ memory- costly

Gap

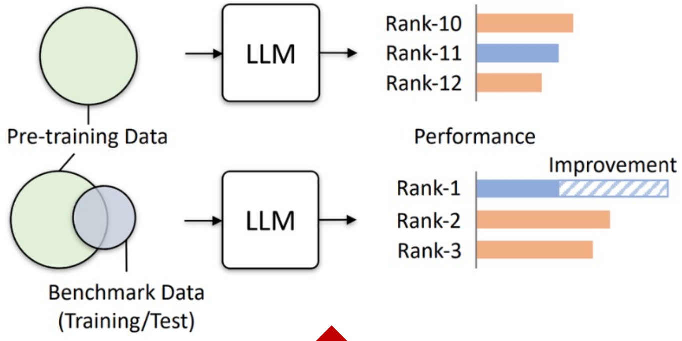


Real-time, focus on cost

RecSys: cost sensitive

Evaluation

LLM: Trained on many data, text-focused, language



Evaluation?

RecSys research: interactions, offline, anonymous data

Generative Recommendation Paradigm

❑ Generative AI for recommendation

- Personalized **content generation**, including item repurposing and creation.
  - **Application**: News, fashion products, micro-videos, virtual products in games, etc.

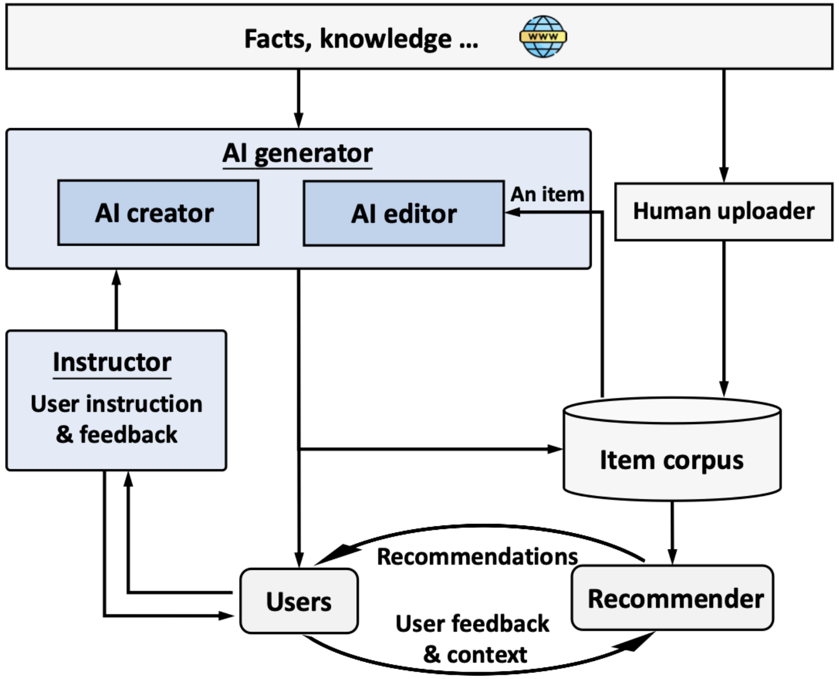


Figure 4: A demonstration of GeneRec. The instructor collects user instructions and feedback to guide content generation. The AI editor aims to repurpose existing items in the item corpus while the AI creator directly creates new items.

Instructor:

- Pre-process user instructions and feedback to guide the content generation of the AI generator.

AI Editor:

- Refine or repurpose existing items according to personalized user instructions and feedback.
- External facts and knowledge might be used for content generation.

AI Creator:

- Generate new items based on personalized user instructions and feedback.

AI Checker:

- Generation quality checks.
- Trustworthiness checks.

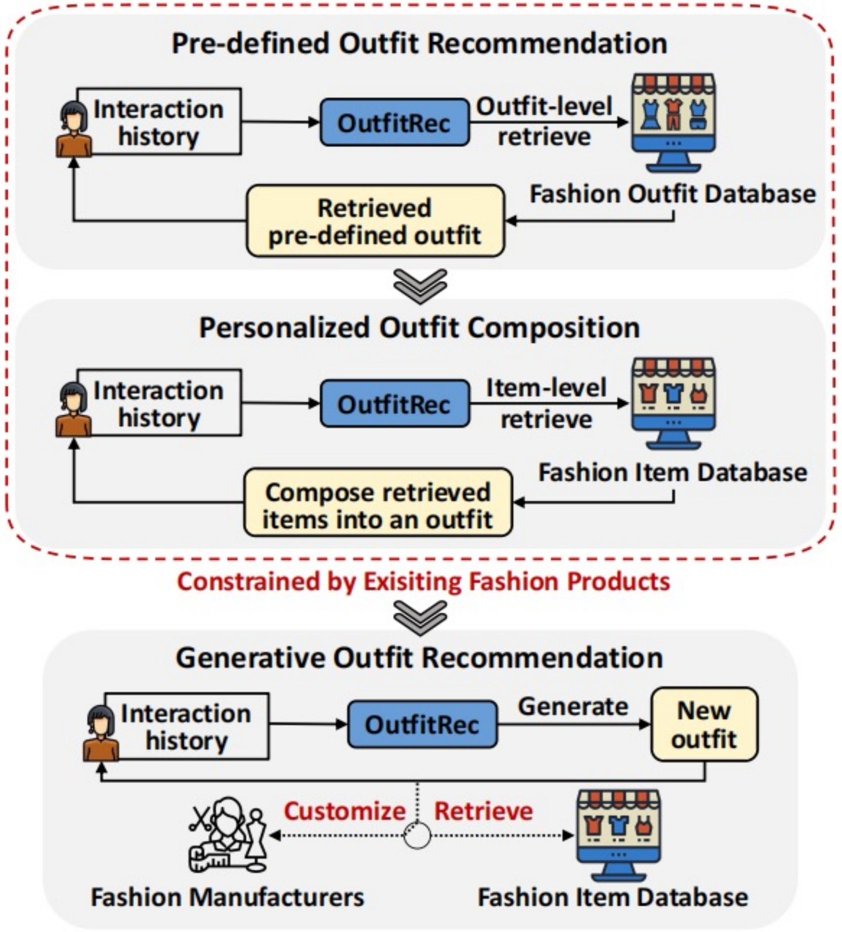
Applicable to many domains, including images, micro-videos, movies, news, books, and even products (for manufacture).

Wenjie Wang et al. Generative Recommendation: Towards Next-generation Recommender Paradigm. arxiv 2023

Generative Recommendation Paradigm

❑ Generative Recommendation in Fashion Domain

The Evolution of Fashion Outfit Recommendation

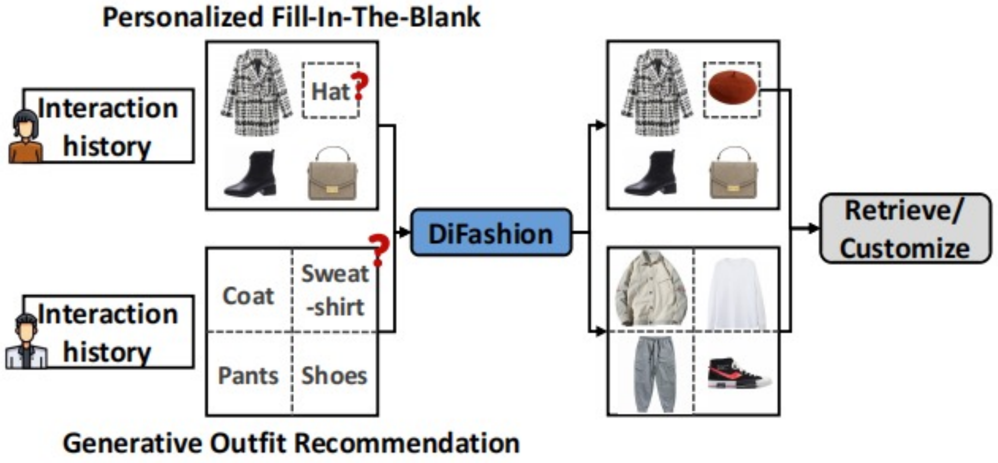


New Task

Generative Outfit Recommendation

Objective: generating a set of new personalized fashion products to compose a visually compatible outfit catering to users' fashion tastes.

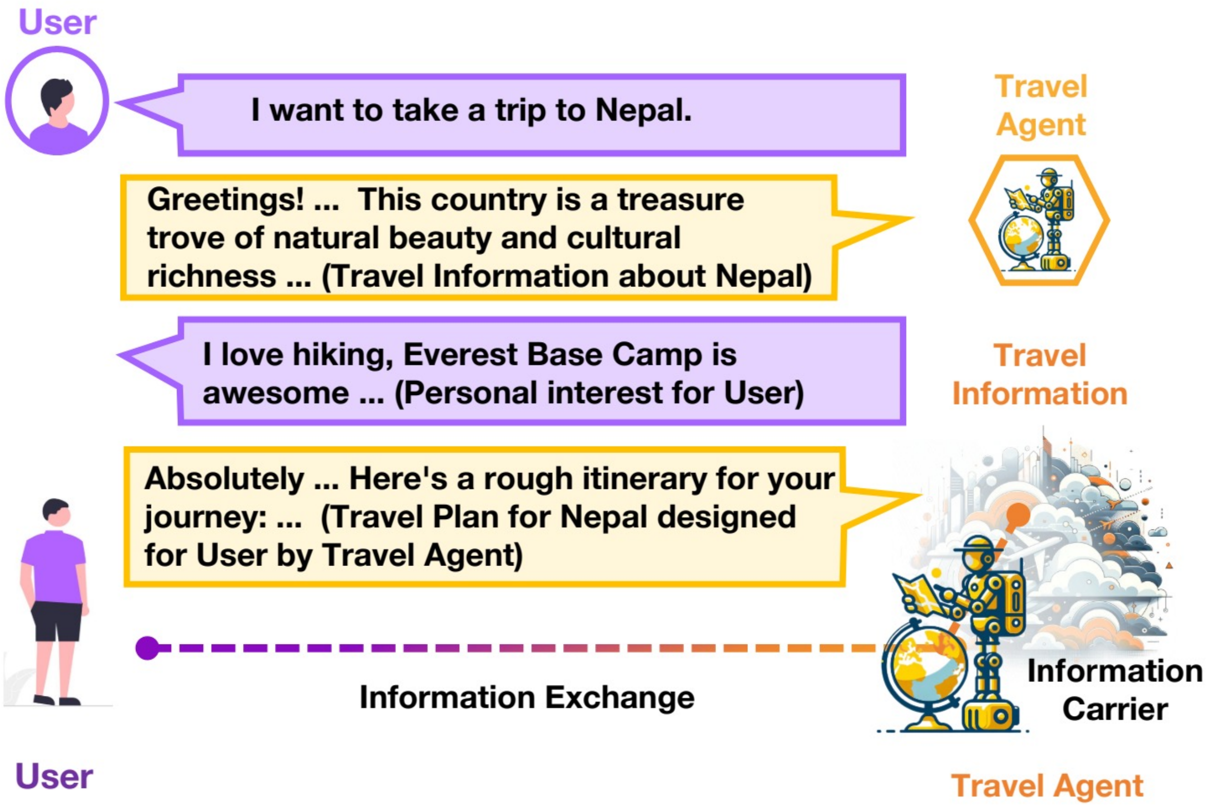
Practical Implementation: retrieve or customize



Yiyan Xu et al. Diffusion Models for Generative Outfit Recommendation. In SIGIR '24.

Recommender for Agent Platform

- □ Existing agent platforms such as GPTs (OpenAI), Poe (Quora), and DouBao (ByteDance) possess a vast number of LLM-based agents.
- □ How to recommend LLM-based Agent to the user?



I want to take a trip to Nepal.

Greetings! ... This country is a treasure trove of natural beauty and cultural richness ... (Travel Information about Nepal)

I love hiking, Everest Base Camp is awesome ... (Personal interest for User)

Absolutely ... Here's a rough itinerary for your journey: ... (Travel Plan for Nepal designed for User by Travel Agent)

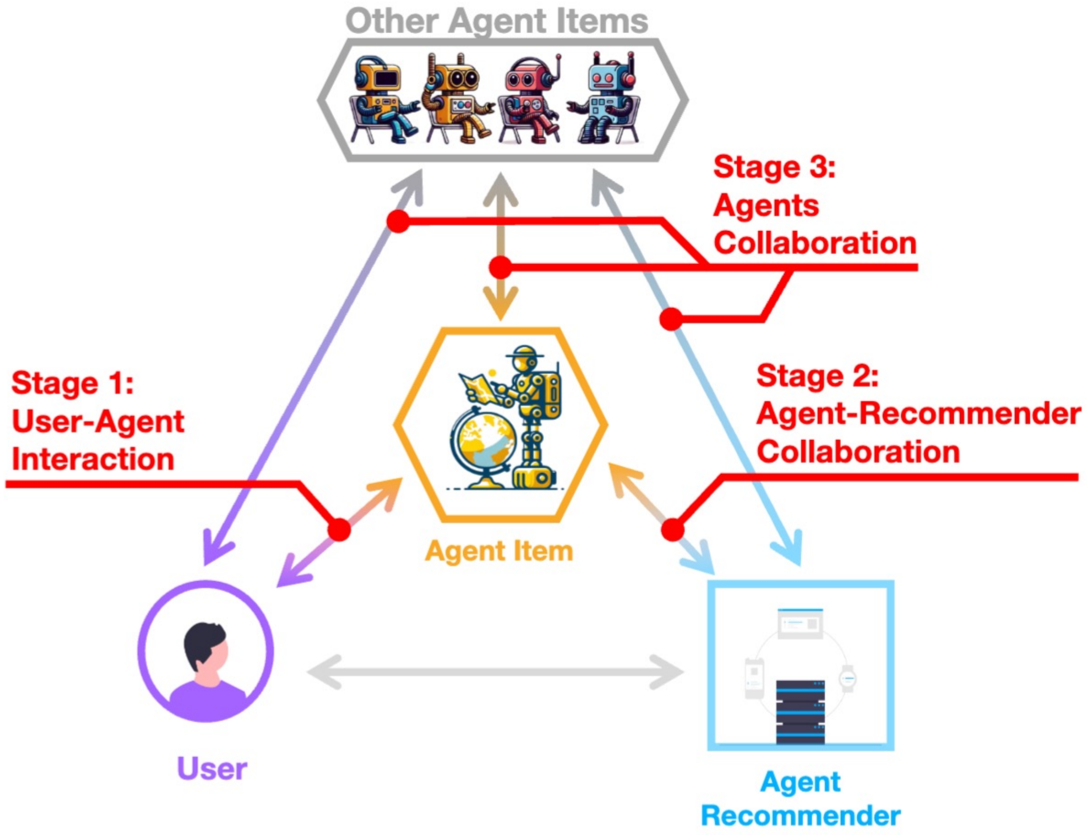
Different from Items in Traditional Recommender System, LLM-based Agent holds the potential to extend the format of information carriers and the way of information exchange.

-> Formulate new Information System

-> New Rec paradigm Rec4Agentverse

Jizhi Zhang et al. Prospect Personalized Recommendation on Large Language Model-based Agent Platform Arxiv 2024

Rec4Agentverse



Three stages of Rec4Agentverse . The bidirectional arrows depicted in the Figure symbolize the flow of information.

- User-Agent interaction stage:

Information flows between the user and Agent Item.

- Agent-Recommender Collaboration stage:

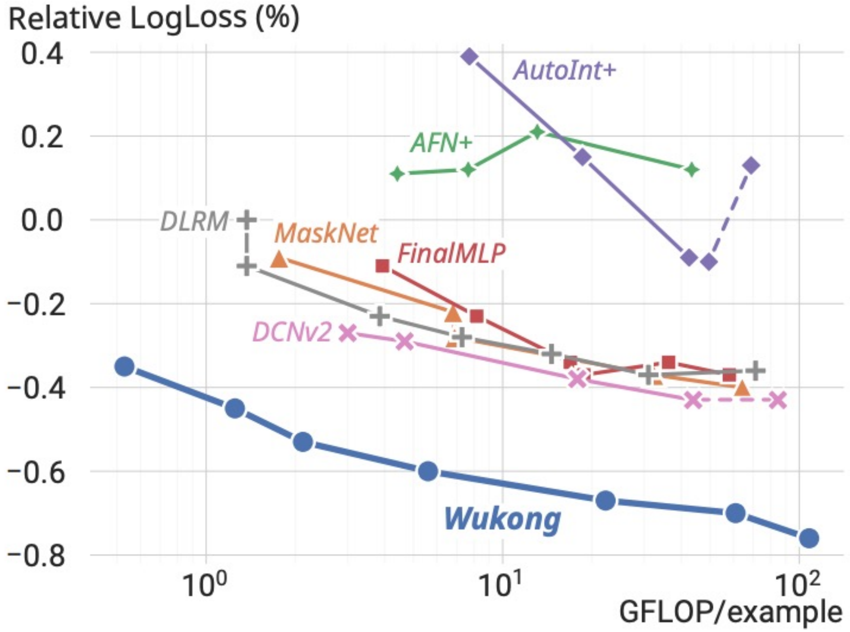
Information flows between Agent Item and Agent Recommender.

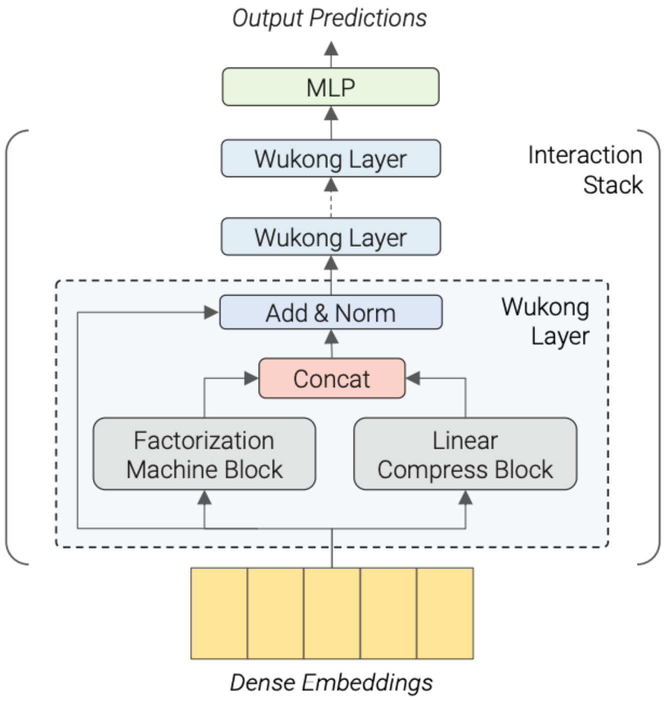
- Agents Collaboration stage:

Information flows between Agent Items.

Jizhi Zhang et al. Prospect Personalized Recommendation on Large Language Model-based Agent Platform Arxiv 2024

Wukong

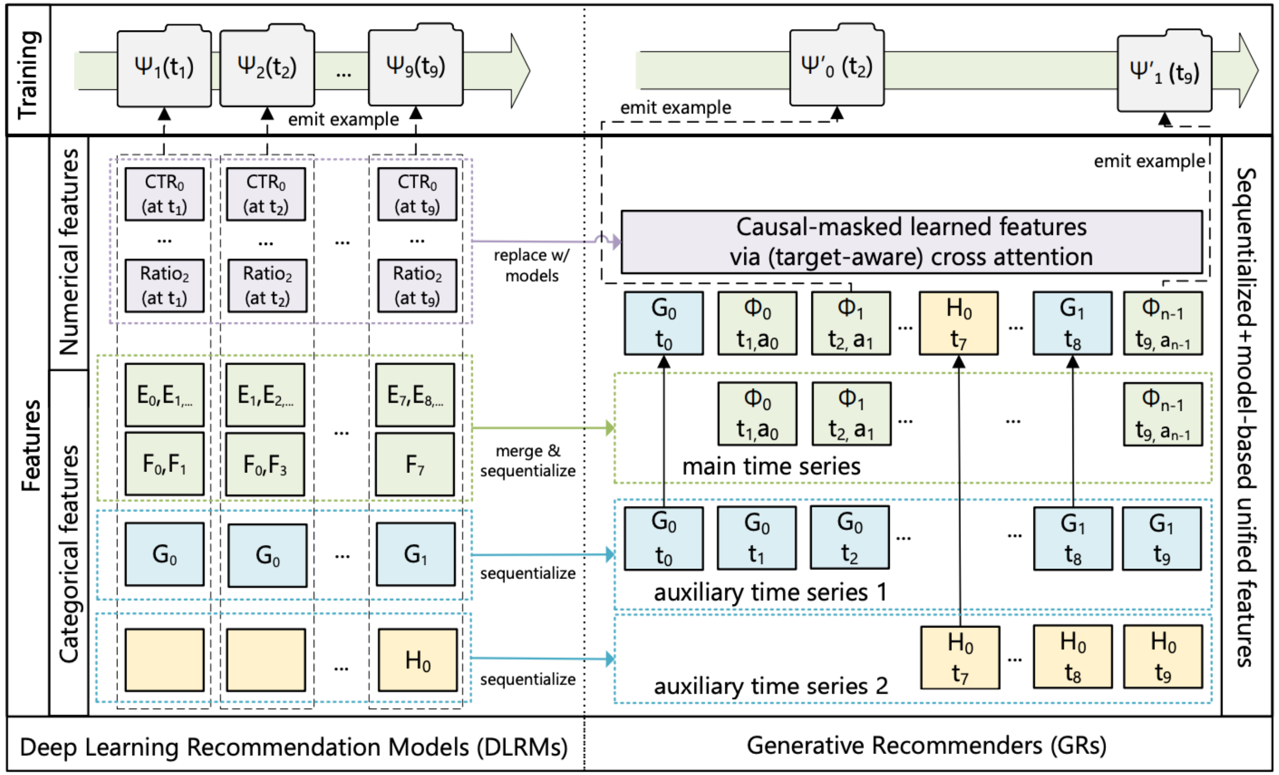


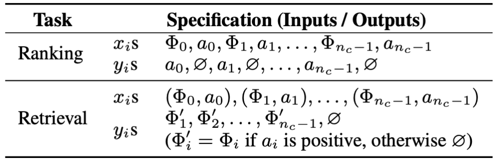


- ❑ The scaling properties of the CTR model have been verified, showing excellent performance on both internal and open-source data.
- ❑ Demonstrates the possibility of increasing the size of CTR models through clever structural design and appropriate scaling settings
- ❑ Exhibits better scaling performance than previous models.

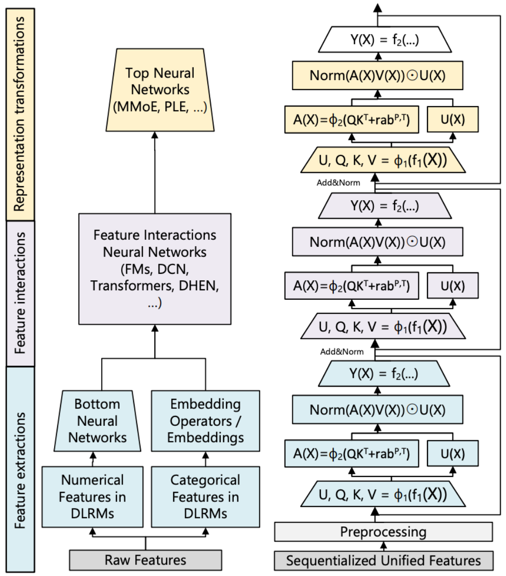
Buyun Zhang et al., 2024 Wukong: Towards a Scaling Law for Large-Scale Recommendation arxiv 2024

Action Speaker Louder than Words





Task		Specification (Inputs / Outputs)
Ranking	x_is	$\Phi_0, a_0, \Phi_1, a_1, \ldots, \Phi_{n_c-1}, a_{n_c-1}$
	y_is	$a_0, \varnothing, a_1, \varnothing, \ldots, a_{n_c-1}, \varnothing$
Retrieval	x_is	$(\Phi_0, a_0), (\Phi_1, a_1), \ldots, (\Phi_{n_c-1}, a_{n_c-1})$
	y_is	$\Phi'_1, \Phi'_2, \ldots, \Phi'_{n_c-1}, \varnothing$ ($\Phi'_i = \Phi_i$ if a_i is positive, otherwise $\varnothing$)


Representation transformations
Feature interactions
Feature extractions
Top Neural Networks (MMoE, PLE, ...)
Feature Interactions Neural Networks (FMs, DCN, Transformers, DHEN, ...)
Bottom Neural Networks
Embedding Operators / Embeddings
Numerical Features in DLRMs
Categorical Features in DLRMs
Raw Features
Y(X) = f2(...)
Norm(A(X)V(X))⊙U(X)
A(X)=ϕ2(QKT+rabP,T)
U(X)
U, Q, K, V = ϕ1(f1(X))
Add&Norm
Preprocessing
Sequentialized Unified Features

- ❑ Generative Recommender
- ❑ New model architecture and feature processing methods.
  - ❑ Unify user behavior sequence
  - ❑ Process both ranking and retrieval tasks

Jiaqi Zhai et al., 2024 Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations ICML 2024

Action Speaker Louder than Words

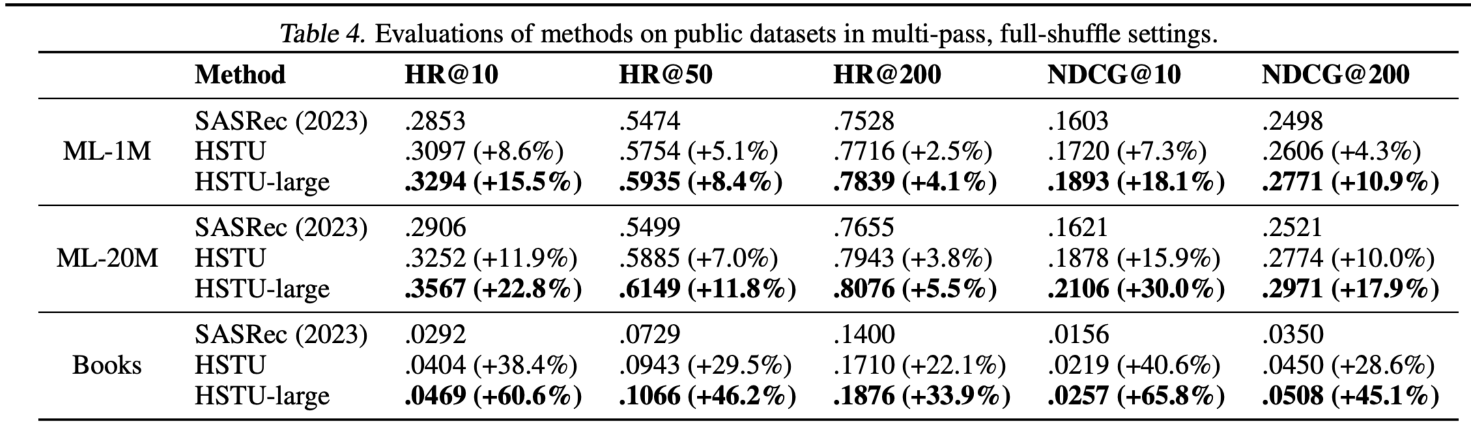


Table 4. Evaluations of methods on public datasets in multi-pass, full-shuffle settings.

	Method	HR@10	HR@50	HR@200	NDCG@10	NDCG@200
ML-1M	SASRec (2023)	.2853	.5474	.7528	.1603	.2498
	HSTU	.3097 (+8.6%)	.5754 (+5.1%)	.7716 (+2.5%)	.1720 (+7.3%)	.2606 (+4.3%)
	HSTU-large	**.3294 (+15.5%)**	**.5935 (+8.4%)**	**.7839 (+4.1%)**	**.1893 (+18.1%)**	**.2771 (+10.9%)**
ML-20M	SASRec (2023)	.2906	.5499	.7655	.1621	.2521
	HSTU	.3252 (+11.9%)	.5885 (+7.0%)	.7943 (+3.8%)	.1878 (+15.9%)	.2774 (+10.0%)
	HSTU-large	**.3567 (+22.8%)**	**.6149 (+11.8%)**	**.8076 (+5.5%)**	**.2106 (+30.0%)**	**.2971 (+17.9%)**
Books	SASRec (2023)	.0292	.0729	.1400	.0156	.0350
	HSTU	.0404 (+38.4%)	.0943 (+29.5%)	.1710 (+22.1%)	.0219 (+40.6%)	.0450 (+28.6%)
	HSTU-large	**.0469 (+60.6%)**	**.1066 (+46.2%)**	**.1876 (+33.9%)**	**.0257 (+65.8%)**	**.0508 (+45.1%)**

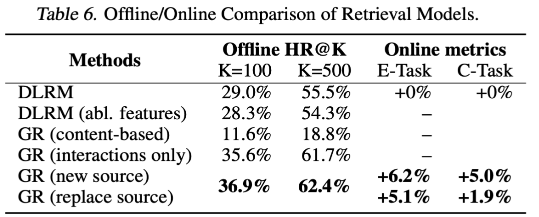


Table 6. Offline/Online Comparison of Retrieval Models.

Methods	**Offline HR@K**		**Online metrics**	
	K=100	K=500	E-Task	C-Task
DLRM	29.0%	55.5%	+0%	+0%
DLRM (abl. features)	28.3%	54.3%	–	
GR (content-based)	11.6%	18.8%	–	
GR (interactions only)	35.6%	61.7%	–	
GR (new source)	**36.9%**	**62.4%**	**+6.2%**	**+5.0%**
GR (replace source)			**+5.1%**	**+1.9%**

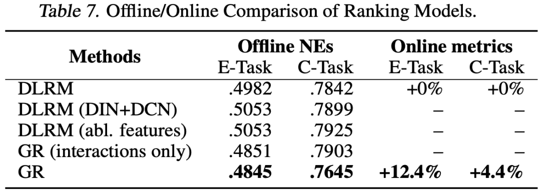
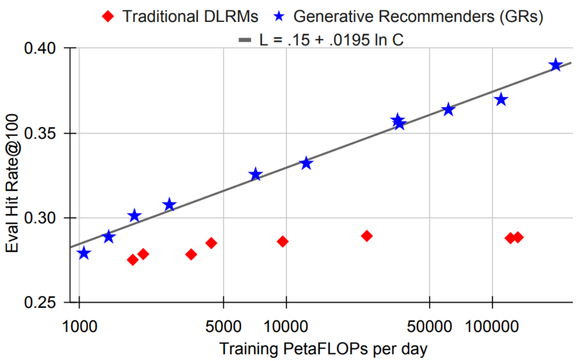
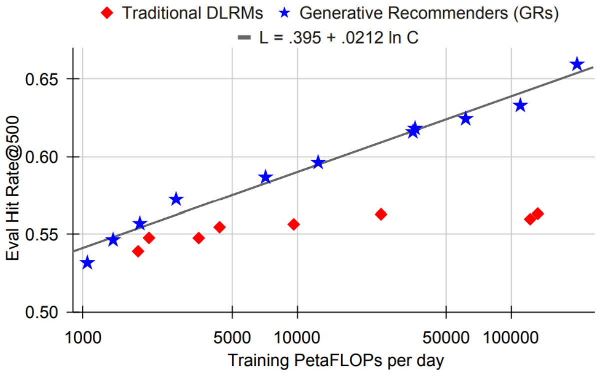


Table 7. Offline/Online Comparison of Ranking Models.

Methods	**Offline NEs**		**Online metrics**	
	E-Task	C-Task	E-Task	C-Task
DLRM	.4982	.7842	+0%	+0%
DLRM (DIN+DCN)	.5053	.7899	–	–
DLRM (abl. features)	.5053	.7925	–	–
GR (interactions only)	.4851	.7903	–	–
GR	**.4845**	**.7645**	**+12.4%**	**+4.4%**

- ❑ Better performance than traditional models on in-house data and open-source data





- ❑ Far more scaling ability than traditional DLRMs

Jiaqi Zhai et al., 2024 Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations ICML 2024

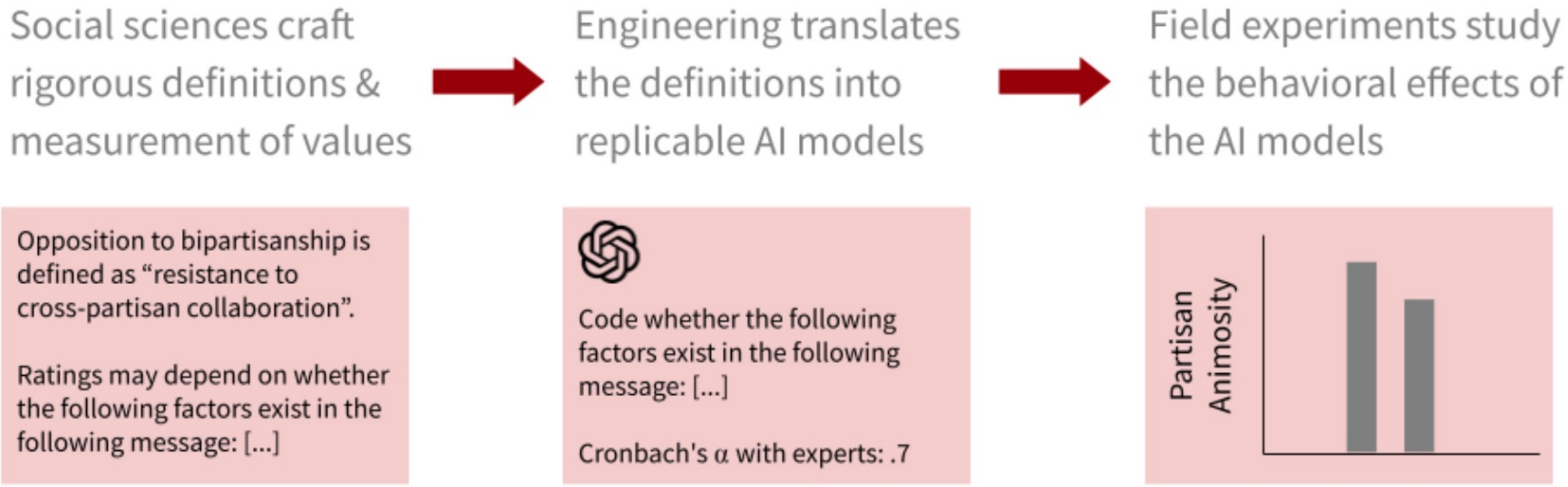
Large Behavior Model

- ❑ **What we have know?**
    - ❑ Scaling Law remains effective on industrial-scale data when combined with an appropriate architecture in the context of recommendation scenario.
    - ❑ When the model is large enough and captures high-order information, it exhibits a certain generalization ability across scenes and domains.

- ❑ **What we do not know?**
    - ❑ How to integrate world knowledge and LLMs for generative recommendation.
    - ❑ How do we model complex user behaviors, including short-term and long-term behaviors? And how is the scalability of these behaviors manifested?
    - ❑ How to model the collaborative information across user-item interactions?

Social Value Alignment of LLMRec

- **Social media AI already embed values** --- maximize each user's individual experience---as predicted through likes in RecSys
- **It can harm societal values** --- wellbeing, social capital, mitigating harm to minoritized groups, democracy, and maintaining pro-social norms.
- **Could we directly encode societal values into RecSys?**



Social sciences craft rigorous definitions & measurement of values

Engineering translates the definitions into replicable AI models

Field experiments study the behavioral effects of the AI models

Opposition to bipartisanship is defined as "resistance to cross-partisan collaboration".

Ratings may depend on whether the following factors exist in the following message: [...]

Code whether the following factors exist in the following message: [...]

Cronbach's α with experts: .7

Michael S. Bernstein, et al. Tuning Our Algorithmic Amplifiers: Encoding Societal Values into Social Media AIs.

Thanks for Your Listening !

Tutorial on Large Language Models for Recommendation

Find our slides at

https://generative-rec.github.io/tutorial/

Survey: A Survey of Generative Search and Recommendation in the Era of Large Language Models

https://arxiv.org/pdf/2404.16924

Follow our WeChat account “智荐阁”!

智荐阁

Tutorial

Survey