香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Large Language Models for Generative Recommendation: A Survey and Visionary Discussions

Lei Li[1], Yongfeng Zhang[2], Dugang Liu[3], Li Chen[1]

[1] Hong Kong Baptist University, [2] Rutgers University, [3] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)

csleili@comp.hkbu.edu.hk

Apr. 22, 2024

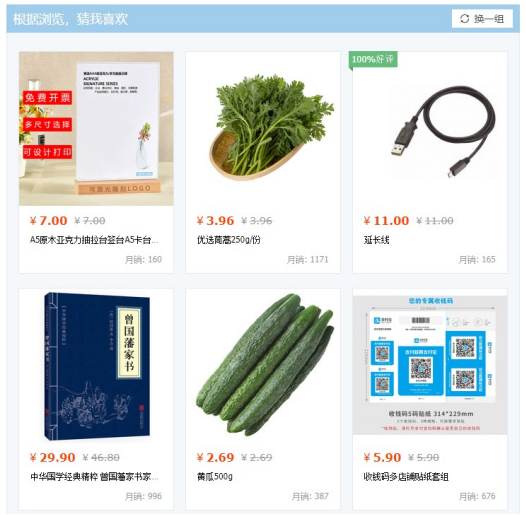
LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Outline

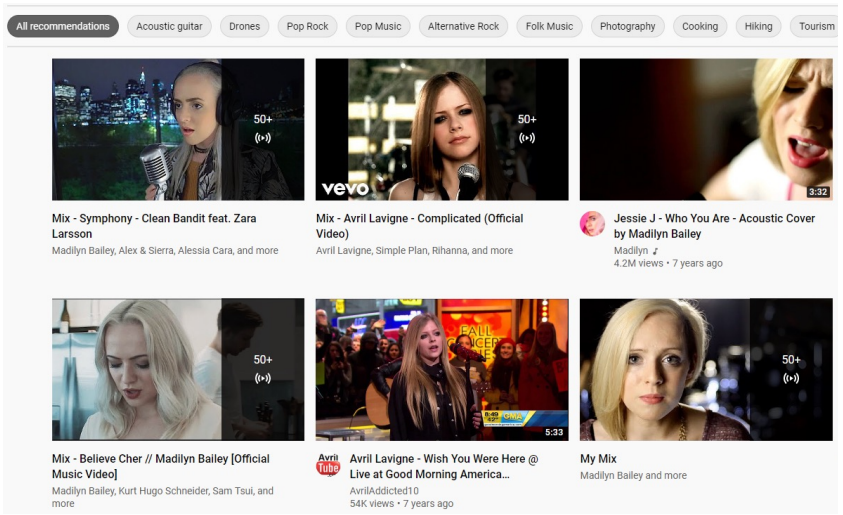
- **Why Generative Recommendation**
- ID Creation Methods
- How to Do Generative Recommendation
- Challenges and Opportunities

Recommendations Everywhere

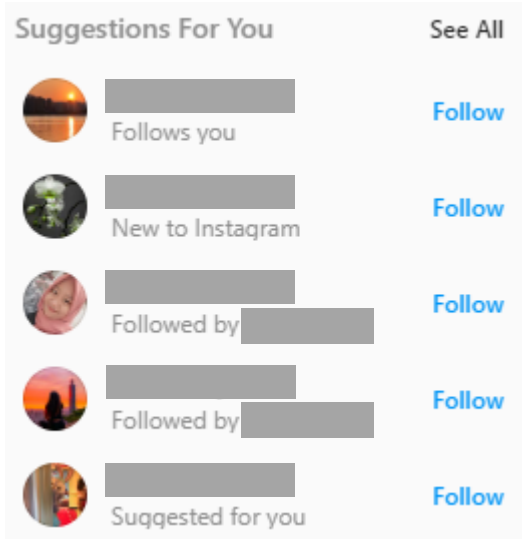
E-commerce
(taobao.com)



Music
(youtube.com)



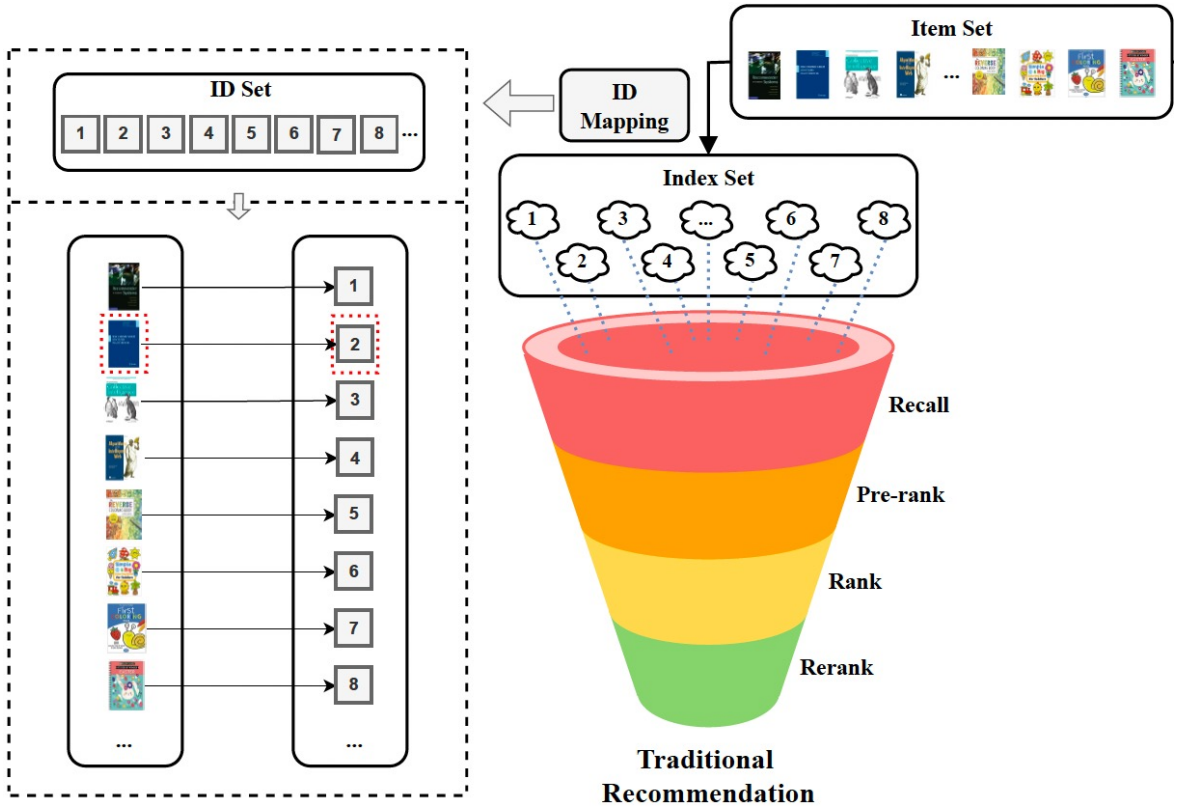
Social Network
(instagram.com)



Movie
(movie.douban.com)

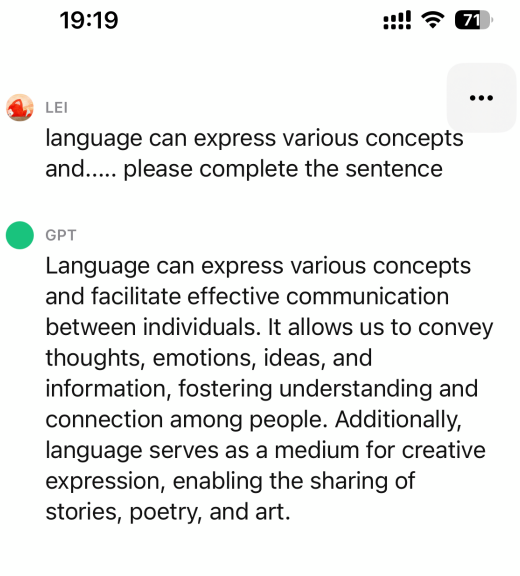
Discriminative Recommendation

- Huge number of items on recommendation platforms
- Computationally expensive score calculation for each item
- Multi-stage filtering to narrow down candidates
  - Simple methods at early stage
  - Complex models at final stage
- Gap between academic research and industrial applications

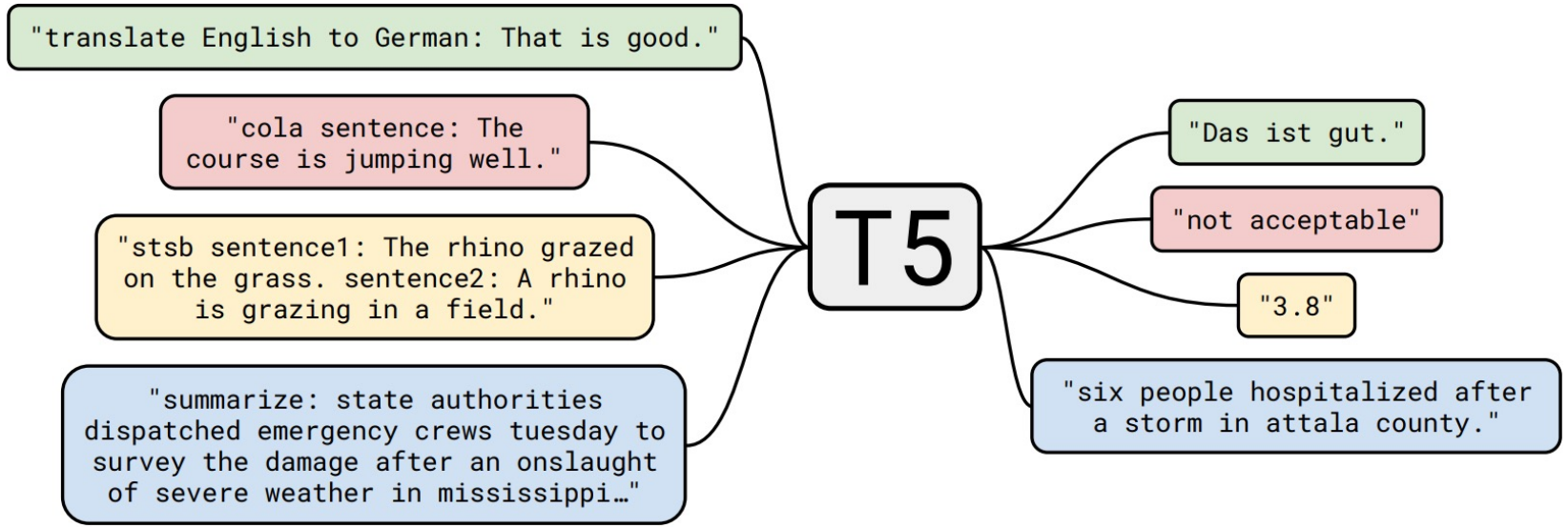


Large Language Models (LLM)

- All tasks formulated as a sequence-to-sequence problem
  - Use previous tokens to predict the next token



ChatGPT [1]



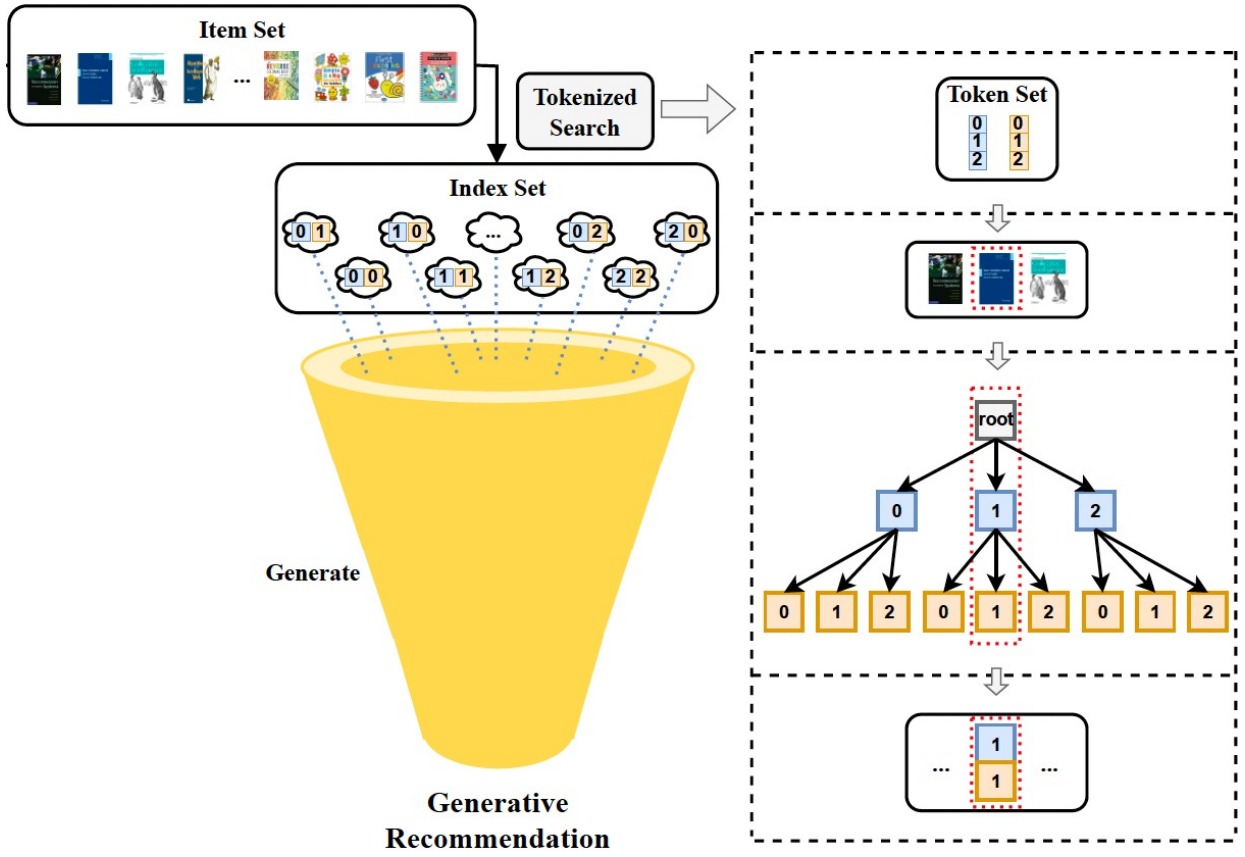
T5 [2]

[1] https://openai.com/chatgpt

[2] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." JMLR'20.

Generative Recommendation

- Simplify recommendation process to one stage
  - Directly generate items for recommendation
  - Implicitly enumerate all items
- Use finite tokens to represent infinite items
  - # tokens = 1000
  - ID length = 10 tokens
  - # items = $1000^{10} = 10^{30}$



Outline

- Why Generative Recommendation
- **ID Creation Methods**
- How to Do Generative Recommendation
- Challenges and Opportunities

Generalized Definition of ID

- *An ID in recommender systems is a sequence of tokens that can uniquely identify an entity, such as a user or an item.*
  - An embedding ID (special case)
  - A sequence of numerical tokens
    - <item><_><73><91> [1]
  - A sequence of word tokens
    - An item title
      - The Lord of the Rings
    - A description of the item
    - A news article
    - A sequence of meaningless words
      - Ring epic journey fellowship adventure [2]

[1] Geng, Shijie, et al. "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)." RecSys'22.
[2] Hua, Wenyue, et al. "How to index item ids for recommendation foundation models." SIGIR-AP'23.

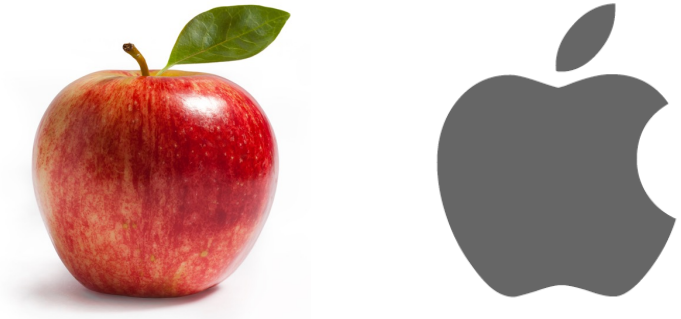
ID Representation in LLM-based Recommendation

- Token sequences as IDs are utilized by a growing number of works
- It is quite common to consider user and item metadata as IDs
- Embedding IDs are rarely used

Item ID	User ID	Related Work
Token Sequence (e.g., "*56 78*")	Token Sequence	P5 (Geng et al., 2022c), UP5 (Hua et al., 2024), VIP5 (Geng et al., 2023), OpenP5 (Xu et al., 2023b), POD (Li et al., 2023b), GPTRec (Petrov and Macdonald, 2023), TransRec (Lin et al., 2023b), LC-Rec (Zheng et al., 2023), (Hua et al., 2023b)
Item Title (e.g., "*Dune*")	Interaction History (e.g., ["*Dune*", "*Her*", ...])	LMRecSys (Zhang et al., 2021), GenRec (Ji et al., 2024), TALLRec (Bao et al., 2023b), NIR (Wang and Lim, 2023), PALR (Yang et al., 2023), BookGPT (Li et al., 2023g), PBNR (Li et al., 2023e), ReLLa (Lin et al., 2024), BIGRec (Bao et al., 2023a), TransRec (Lin et al., 2023b), LLaRa (Liao et al., 2023), Llama4Rec (Luo et al., 2024), Logic-Scaffolding (Rahdari et al., 2024), (Dai et al., 2023; Liu et al., 2023a; Hou et al., 2024; Li et al., 2023f; Zhang et al., 2023c; Wang et al., 2023c; Lin and Zhang, 2023; Di Palma et al., 2023; Li et al., 2023d)
Item Title	Metadata (e.g., age)	InteRecAgent (Huang et al., 2023), (Zhang et al., 2023b; He et al., 2023)
Metadata	Metadata	M6-Rec (Cui et al., 2022), LLMRec (Liu et al., 2023b), RecMind (Wang et al., 2023d), TransRec (Lin et al., 2023b), (Wu et al., 2024)
Embedding ID	Embedding ID	PEPLER (Li et al., 2023a)

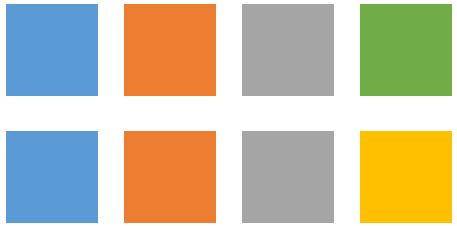
Problems with Existing IDs

- Metadata-based IDs
  - Long IDs
    - Computationally expensive to conduct generation
    - Difficult to find an exact match in database
  - Short IDs
    - Difficult to distinguish two items
      - Apple fruit vs. Apple company
- Embedding IDs
  - Not compatible with LLM as OOV tokens
  - Cost a lot of memory to store



LLM-compatible IDs

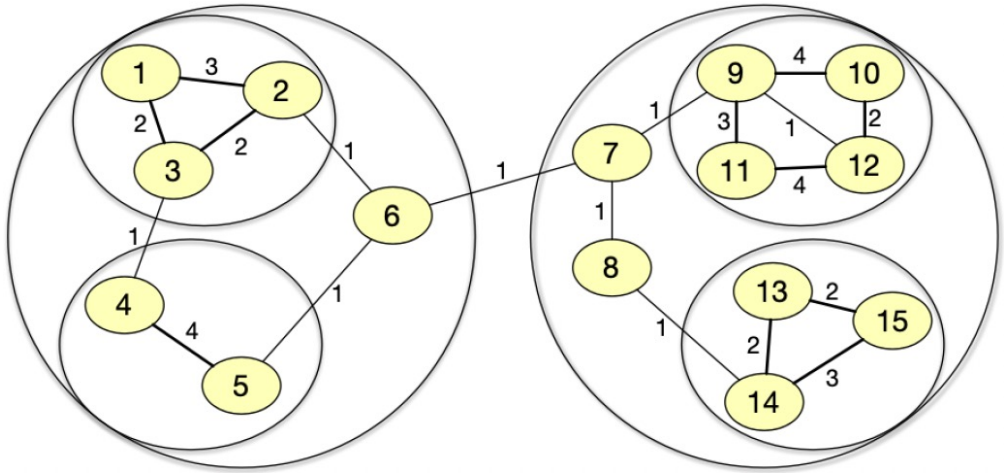
- Retain collaborative information of IDs in LLM environment
  - User-user
  - Item-item
  - User-item
- Short and exact representations of IDs
  - Similar users/items share more tokens while the remaining tokens are used to distinguish them



An illustration of two ID sequences

Spectral Clustering

- Compose an item ID with nodes on a hierarchical tree
  - Construct an item graph with co-occurring frequency of two items
  - Recursively group similar items into the same cluster
  - Construct a hierarchical tree



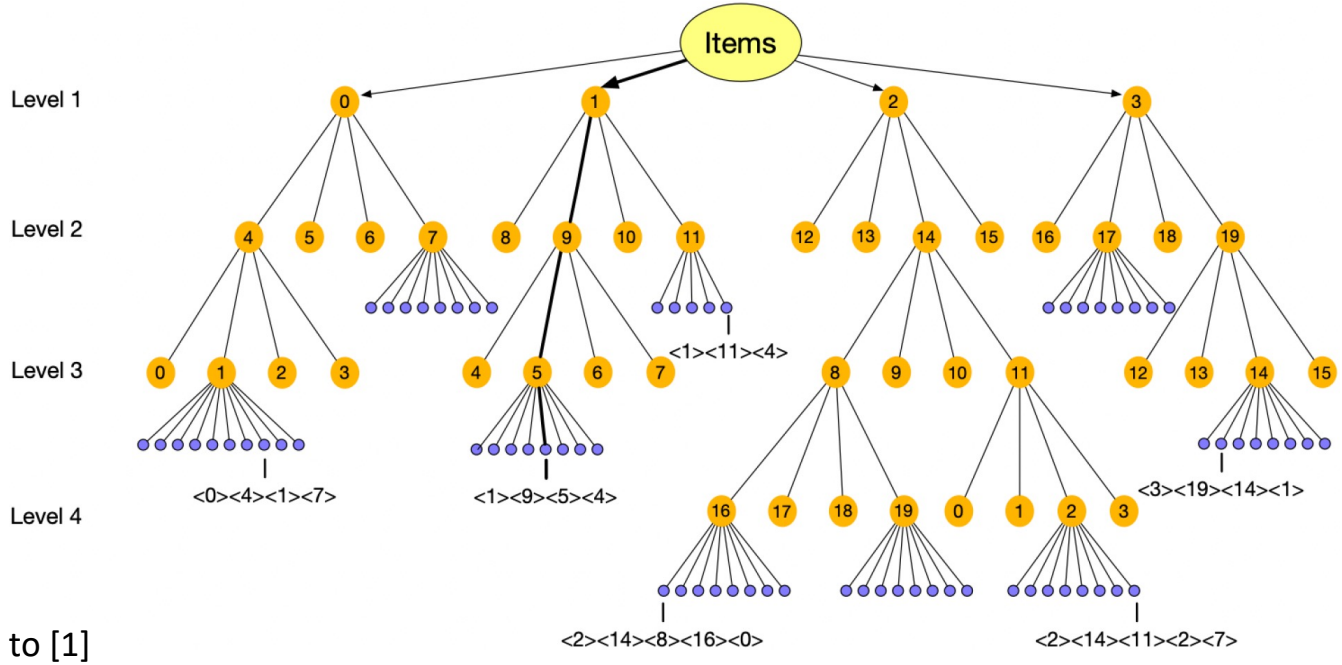


Image credit to [1]

[1] Hua, Wenyue, et al. “How to index item ids for recommendation foundation models.” SIGIR-AP’23.

Singular Value Decomposition

- Acquire an item's ID tokens from its latent factors

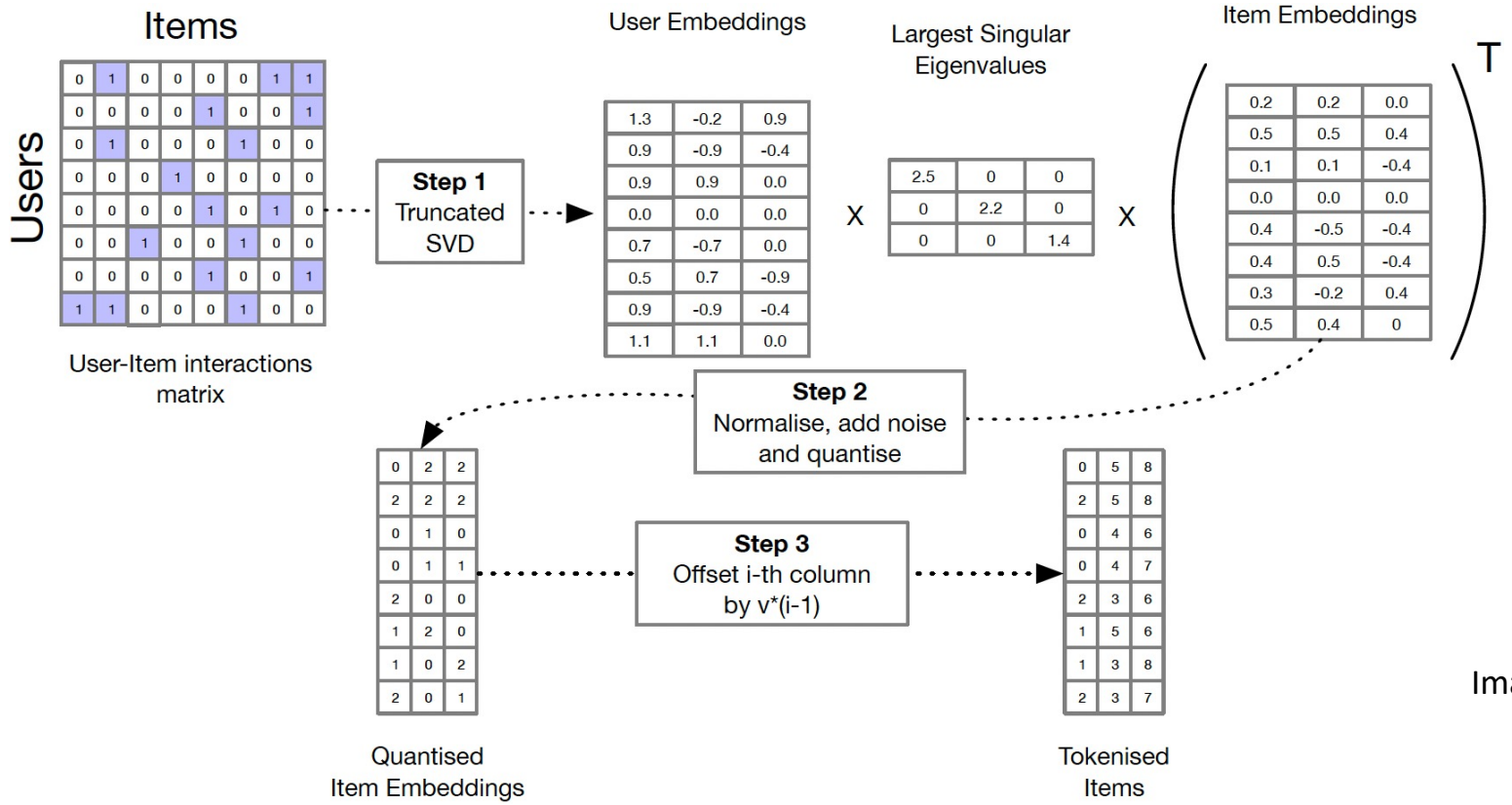


Image credit to [1]

[1] Petrov, Aleksandr V., and Craig Macdonald. "Generative Sequential Recommendation with GPTRec." SIGIR'23 Workshop.

Residual-Quantized Variational AutoEncoder

- Obtain item embedding with its description and pass it through RQVAE's encoder
- Find the nearest embedding to residual vector and keep its index at each step

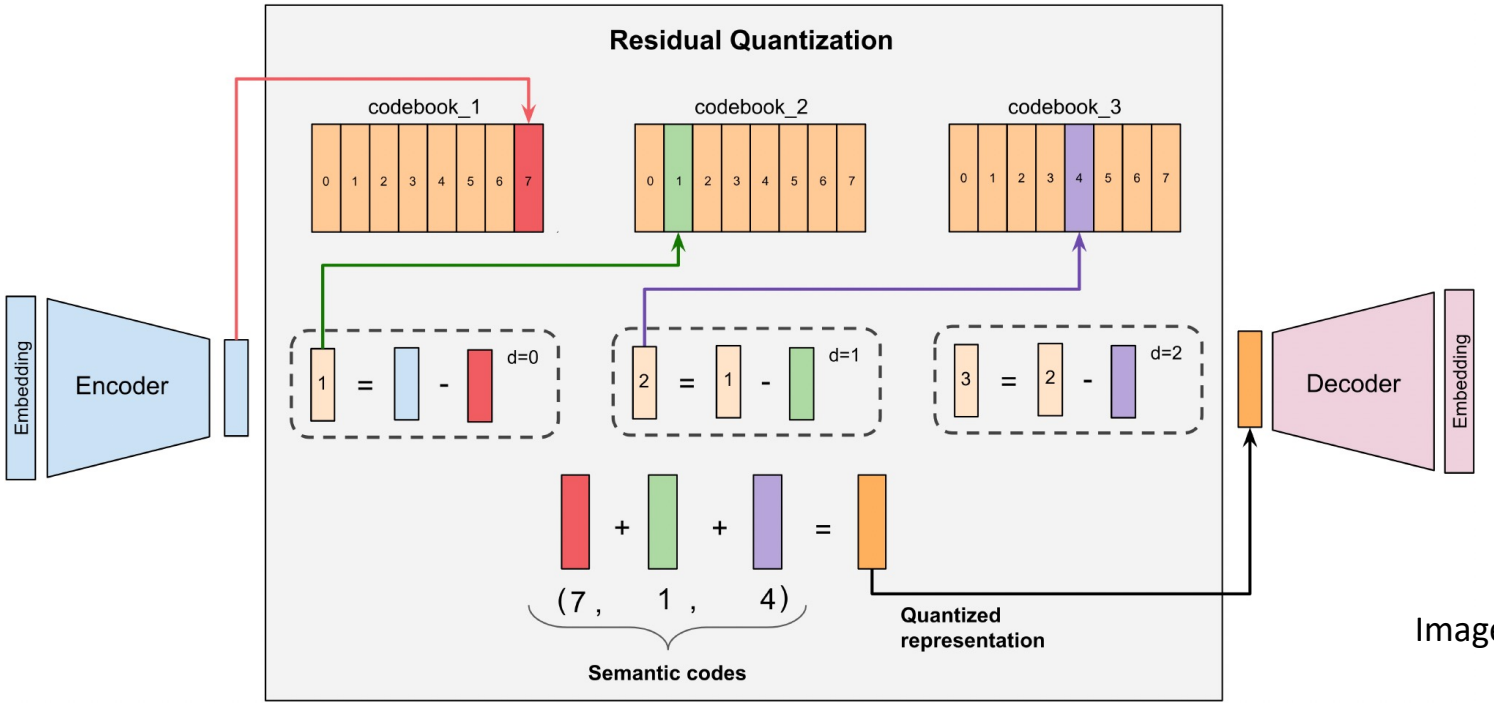


Image credit to [1]

[1] Rajput, Shashank, et al. "Recommender systems with generative retrieval." NeurIPS'23.

Outline

- Why Generative Recommendation
- ID Creation Methods
- **How to Do Generative Recommendation**
- Challenges and Opportunities

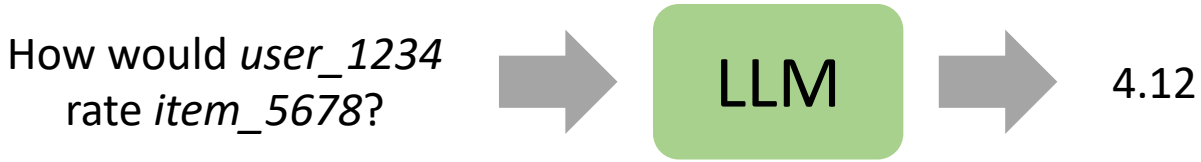
How to Do Generative Recommendation

- Top-N recommendation and sequential recommendation are popular
- Key steps to conduct generation for each task
  - Prompt construction
  - ID filling
  - Auto-regressive generation

Rating Prediction	Top-N Recommendation	Sequential Recommendation	Explainable Recommendation	Review Generation	Review Summarization	Conversational Recommendation
P5 (Geng et al., 2022c), BookGPT (Li et al., 2023g), LLMRec (Liu et al., 2023b), RecMind (Wang et al., 2023d), Llama4Rec (Luo et al., 2024), (Liu et al., 2023a; Dai et al., 2023; Li et al., 2023d)	P5 (Geng et al., 2022c), UP5 (Hua et al., 2024), VIP5 (Geng et al., 2023), OpenP5 (Xu et al., 2023b), POD (Li et al., 2023b), GPTRec (Petrov and Macdonald, 2023), LLMRec (Liu et al., 2023b), RecMind (Wang et al., 2023d), NIR (Wang and Lim, 2023), Llama4Rec (Luo et al., 2024), (Zhang et al., 2023b,c; Liu et al., 2023a; Li et al., 2023f; Dai et al., 2023; Di Palma et al., 2023; Carraro and Bridge, 2024)	P5 (Geng et al., 2022c), UP5 (Hua et al., 2024), VIP5 (Geng et al., 2023), OpenP5 (Xu et al., 2023b), POD (Li et al., 2023b), GenRec (Ji et al., 2024), GPTRec (Petrov and Macdonald, 2023), LMRecSys (Zhang et al., 2021), PALR (Yang et al., 2023), LLMRec (Liu et al., 2023b), RecMind (Wang et al., 2023d), BIGRec (Bao et al., 2023a), TransRec (Lin et al., 2023b), LC-Rec (Zheng et al., 2023), LLaRa (Liao et al., 2023), (Hua et al., 2023b; Liu et al., 2023a; Hou et al., 2024; Zhang et al., 2023c)	P5 (Geng et al., 2022c), VIP5 (Geng et al., 2023), POD (Li et al., 2023b), PEPLER (Li et al., 2023a), M6-Rec (Cui et al., 2022), LLMRec (Liu et al., 2023b), RecMind (Wang et al., 2023d), Logic-Scaffolding (Rahdari et al., 2024), (Liu et al., 2023a)	-	P5 (Geng et al., 2022c), LLMRec (Liu et al., 2023b), RecMind (Wang et al., 2023d), (Liu et al., 2023a)	M6-Rec (Cui et al., 2022), RecLLM (Friedman et al., 2023), InteRecAgent (Huang et al., 2023), PECRS (Ravaut et al., 2024), (Wang et al., 2023c; Lin and Zhang, 2023; He et al., 2023)

Rating Prediction

- Given a user and an item, estimate a score that the user would give the item
- Many studies are based on ChatGPT
- An overall assessment of the target item can be useful
- Predicting a user's rating is less practical
  - Difficulty in collecting explicit feedback



Top-*N* Recommendation

- Recommend *N* items that a user never interacted with
  - Easier to collect implicit feedback
    - Clicking and purchasing

$$\text{Top}(u, i) := \arg\max^{N}_{i \in \mathcal{I}/\mathcal{I}_u} \hat{r}_{u,i}$$

- Due to LLM's context length limit, many works provide a candidate list
  - A testing item and sampled negative items
- LLM with beam search can produce *N* different item IDs

Select one item as recommendation
for *user_1234* from the following candidates:
item_6783, ..., *item_9312*, *item_2834*

Sequential Recommendation

- Given a user's interaction history, predict the next item that the user will interact with
  - Step further than top-*N* recommendation by considering interaction order
- Generating "Yes" or "no" for recommendation is discriminative
  - LLM need to do this for every item

Given *user_1234*'s interaction history *item_3456*, ..., *item_4567*, *item_5678*, predict the next item that the user will click

6789

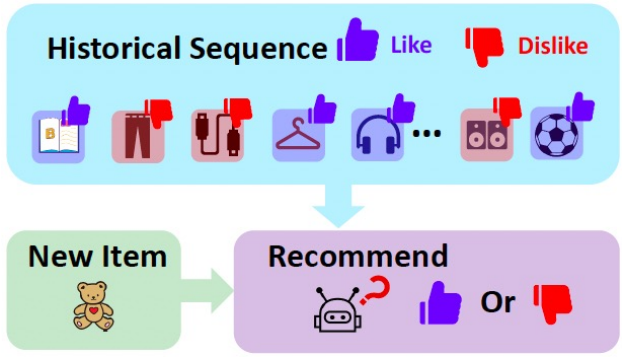


Image credit to [1]

[1] Bao, Keqin, et al. "Tallrec: An effective and efficient tuning framework to align large language model with recommendation." RecSys'23.

Explainable Recommendation

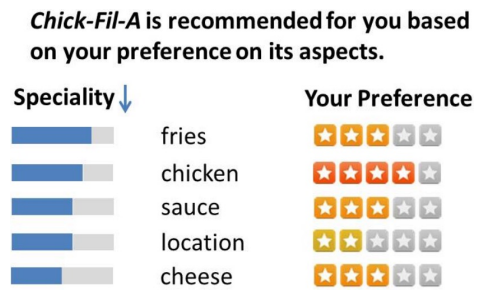


Image credit to [1]

Image credit to [2]

- Various ways to explain a recommendation
  - Explicit item features [1]
  - Visual highlights [2]
- In the context of LLM, generate a sentence to explain why an item is recommended to a user
- Provide item features in the prompt to guide the generation
  - Acting or plot

Explain to *user_1234* why *item_5678* is recommended → LLM → The movie is top-notch

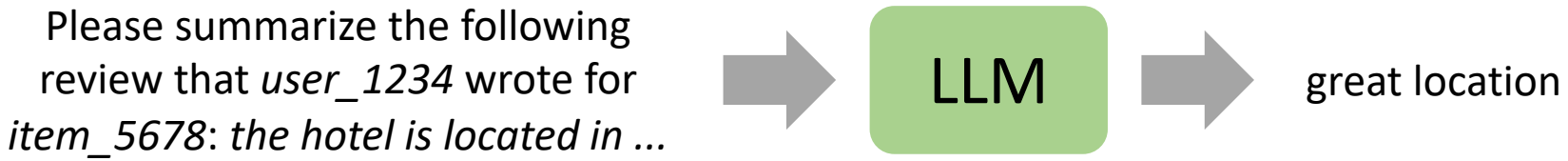
[1] He, Xiangnan, et al. “Trirank: Review-aware explainable recommendation by modeling aspects.” CIKM’15.
[2] Chen, Xu, et al. "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation." SIGIR’19.

Review Generation

- An automatic tool that can draft reviews would make it easier and more efficient for users to leave a comment
- The data would facilitate the development of recommendation research
  - Explainable recommendation
  - Conversational recommendation
- Unexplored with LLM on this problem
  - Too similar to explanation generation

Review Summarization

- A concise review summary could help users quickly know an item's pros and cons
- Many methods target how to summarize a user's own review
  - Unnecessary because the user knows about the target item
- More meaningful to conduct multi-review summarization that summarizes different users' opinions on the same item



Conversational Recommendation

- Engage users in a dialogue to refine preferences and suggest items
  - Advantage: users can freely state their preferences in natural language
- The community has not reached a consensus on how to formulate the task
- Labels are usually adopted to mark the speaker of an utterance

USER: I like action movie.
SYSTEM: How about Mission Impossible?
USER: I have watched it before.
SYSTEM:

What about Heart of Stone?

Evaluation Protocols

- Recommendation tasks
  - Offline metrics
    - RMSE and MAE for rating prediction
    - NDCG, precision and recall for top-*N* recommendation and sequential recommendation
  - Online A/B tests
- Generation tasks
  - Automatic evaluation
    - BLEU and ROUGE for text similarity
    - Problematic to over-emphasize the matching with annotated data [1]
    - More advanced metrics are needed
  - Human evaluation
    - Limited number of participants

[1] Wang, Xiaolei, et al. “Rethinking the evaluation for conversational recommendation in the era of large language models.” EMNLP’23.

Outline

- Why Generative Recommendation
- ID Creation Methods
- How to Do Generative Recommendation
- **Challenges and Opportunities**

LLM-based Agents for Simulation

- User behavior simulators for recommendation algorithms [1]
  - Address data-scarcity problem
- Paradox
  - Useless if the simulator cannot precisely simulate a user's preference
  - Recommendation algorithms not needed if the simulator perfectly simulates a user's preference

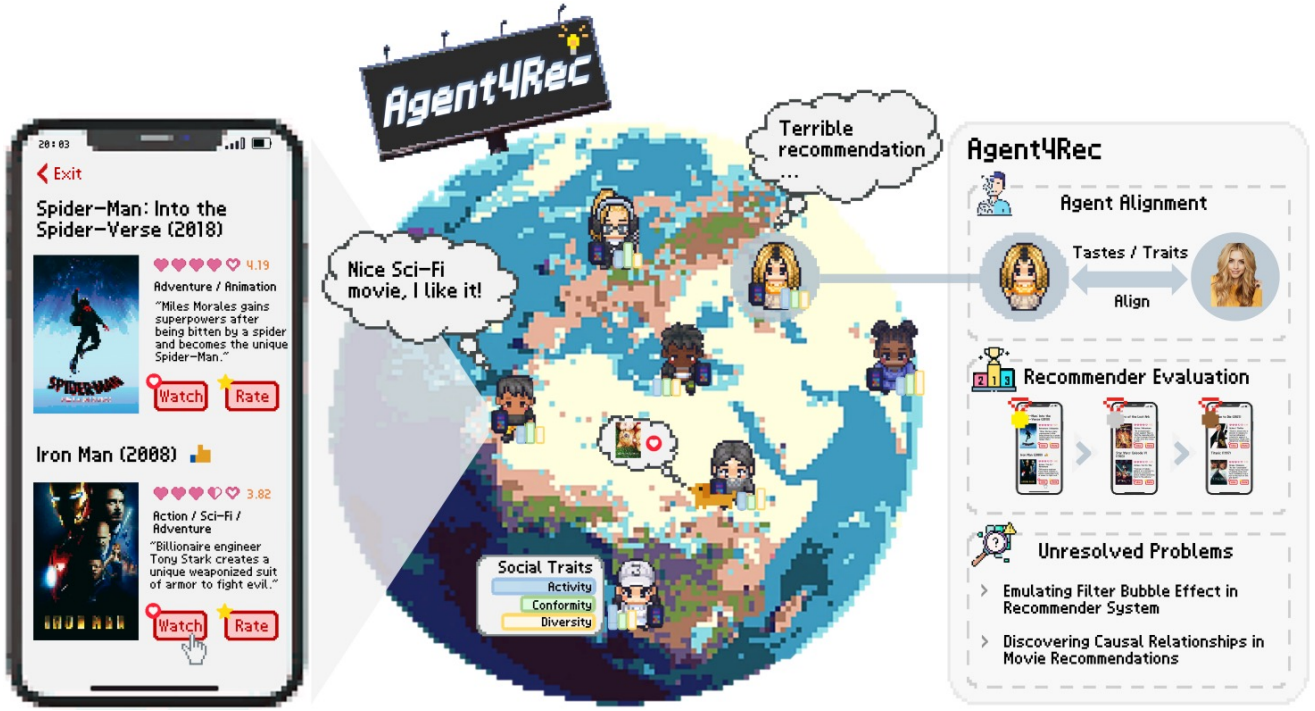


Image credit to [1]

[1] Zhang, An, et al. "On generative agents in recommendation." arXiv'23.

LLM-based Agents for Trip Recommendation

- LLM can call tools, APIs, and expert models to solve complex tasks
- Trip recommender system can cater to a user's personalized needs
  - Duration
  - Budget
  - Attractions
- Then draft an itinerary by looking up real-time information
  - Opening hours
  - Transportation time

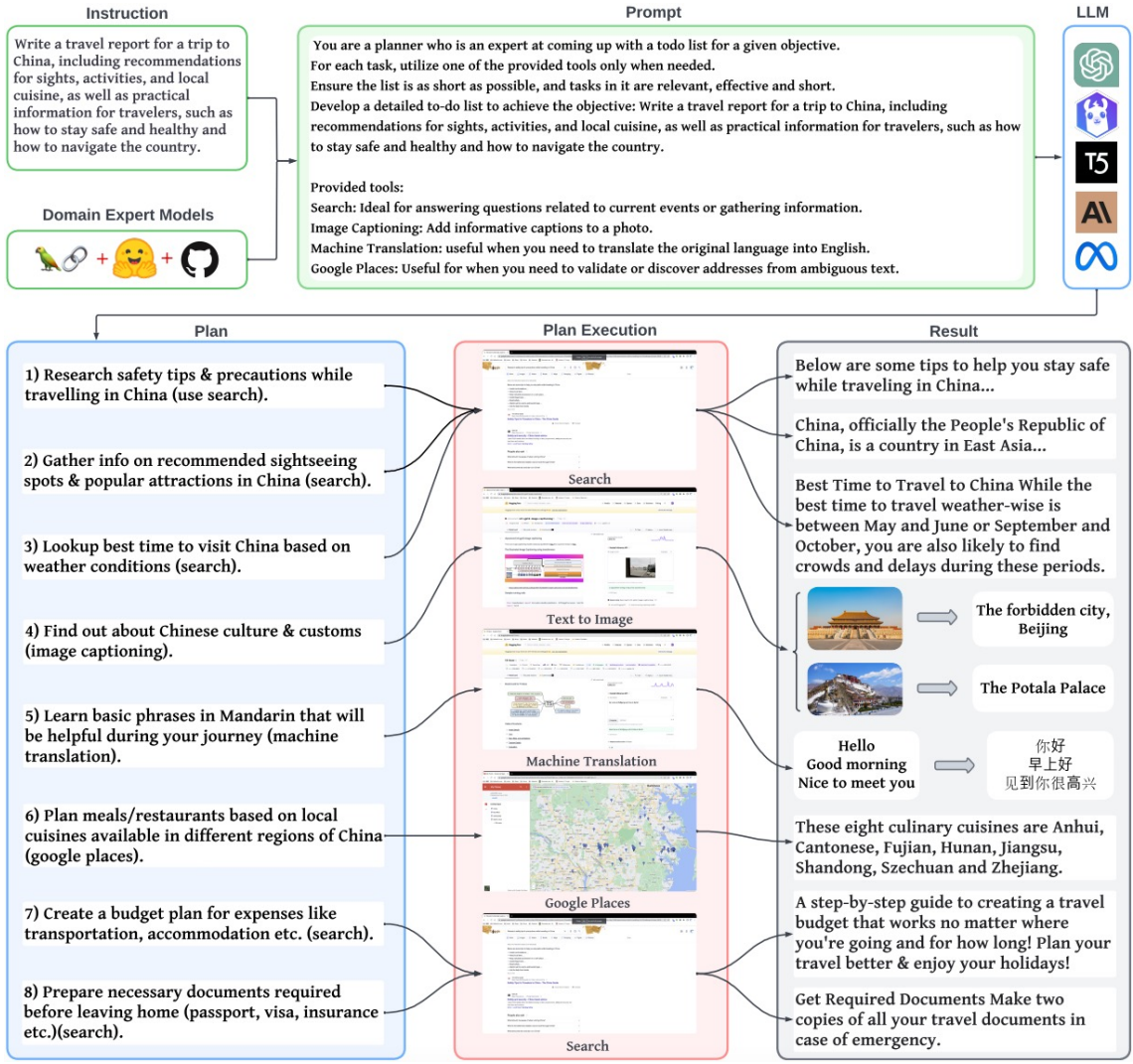


Image credit to [1]

[1] Ge, Yingqiang, et al. "Openagi: When llm meets domain experts." NeurIPS'24.

LLM-based Agents for In-vehicle Recommendation

- Vehicles equipped with intelligent recommender systems might change how people live just as how smart phones did

- Before a trip
  - Routing
  - In-vehicle settings
    - Seat adjustment
- During a trip
  - Out-of-vehicle services
    - Gas/charging stations
    - Restaurants
  - Infotainment
    - News
  - Re-planning
- After a trip
  - Control of connected devices
    - Air conditioner at home

[1] Luettin, Juergen, et al. “Future of in-vehicle recommendation systems @ Bosch." RecSys’19.

Hallucination

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.
GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Image credit to [1]

- Hallucinated recommendations may cause severe losses for users
  - Drug recommendation
  - Medical treatment recommendation
  - Financial investment recommendation
- Possible solutions
  - Meticulously designed IDs [2]
  - Augmented retrieval [3]

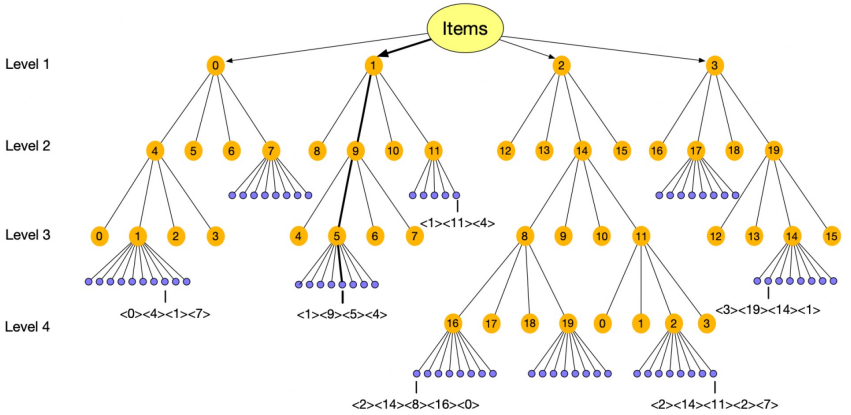


Image credit to [2]

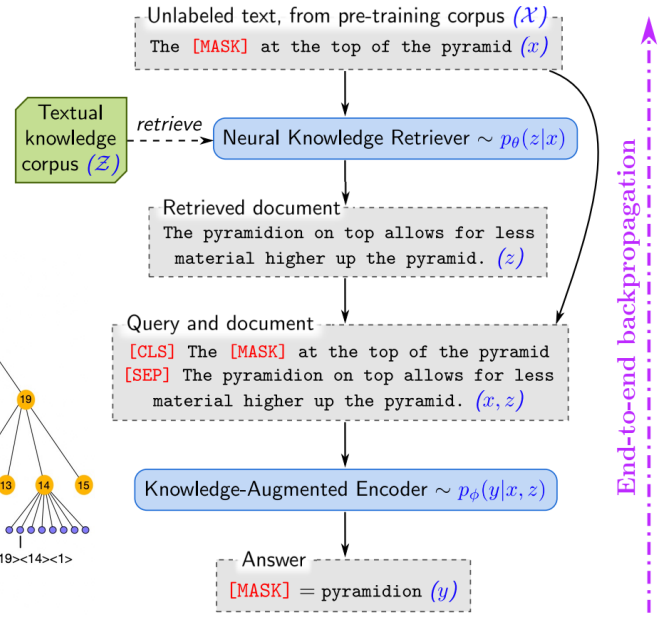


Image credit to [3]

[1] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog (2019).
[2] Hua, Wenyue, et al. "How to index item ids for recommendation foundation models." SIGIR-AP'23.
[3] Guu, Kelvin, et al. "Retrieval augmented language model pre-training." ICML'20.

Content Bias

- Machine-generated explanations for male users are longer than those for female users in game domain
- Training data are adapted from user reviews of games

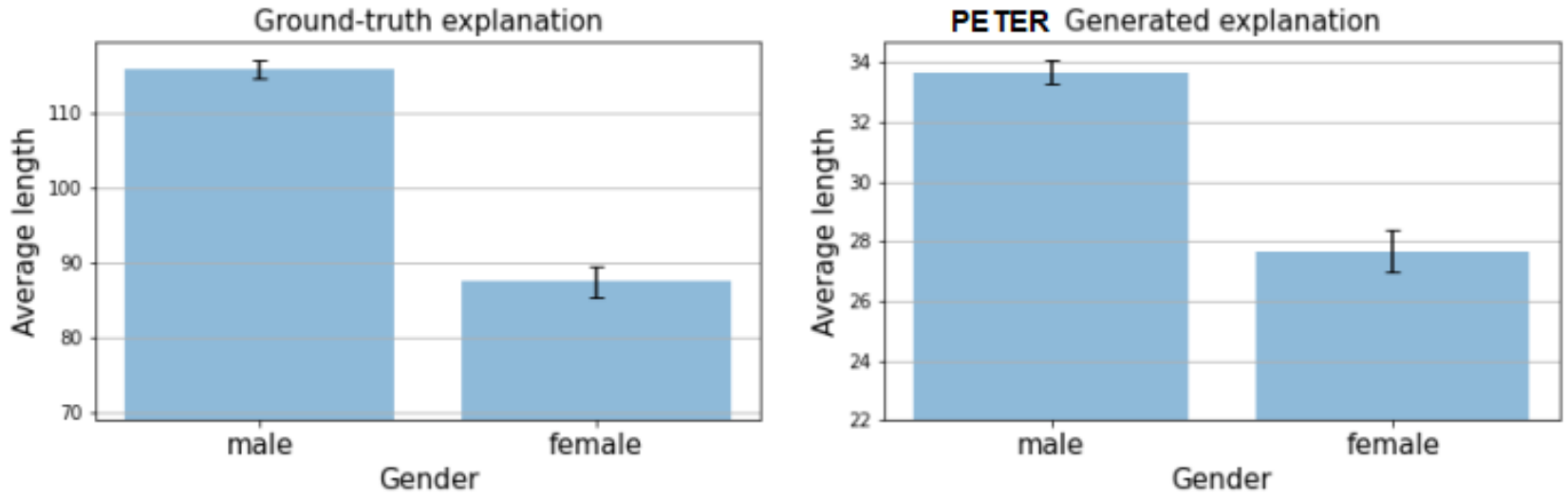


Image credit to [1]

[1] Wang, Nan, et al. "COFFEE: Counterfactual Fairness for Personalized Text Generation in Explainable Recommendation." EMNLP'23.

Recommendation Bias

- Music recommendations made by ChatGPT for people with different demographic attributes are dissimilar [1]
  - Could also be a type of personalization
- More work can be done from the perspective of fairness definition and bias mitigation
  - *What is the boundary between bias and personalization?*

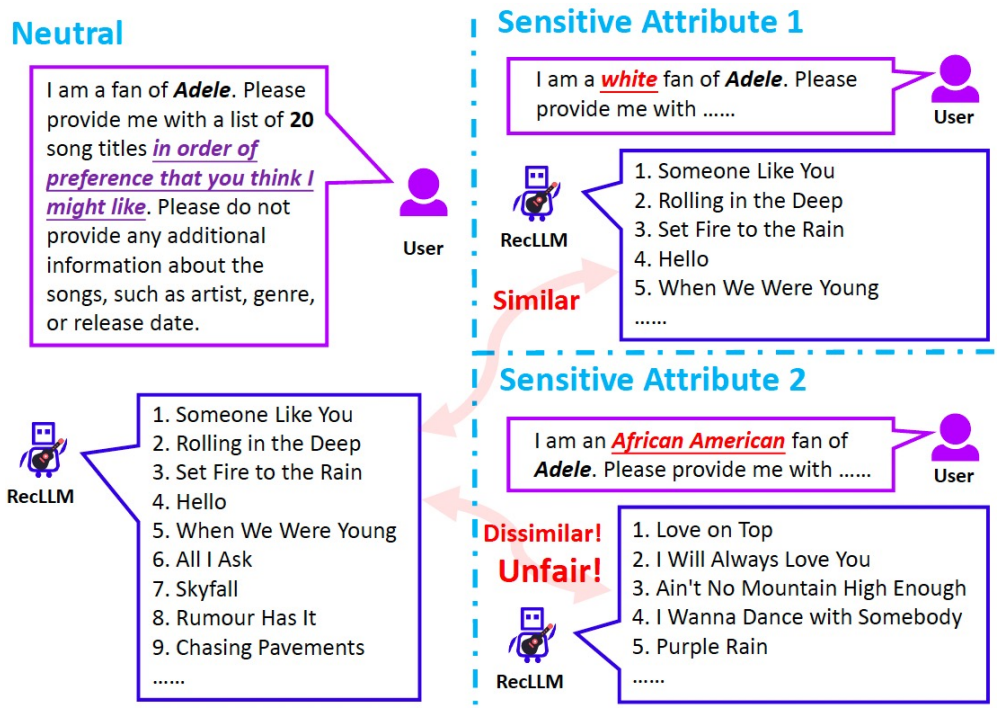


Image credit to [1]

[1] Zhang, Jizhi, et al. “Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation." RecSys’23.

Transparency and Explainability

- Generate natural language explanations for recommended items
  - Largely explored
- Explain the internal working mechanism of LLM
  - Possible solution: align LLM with an explicit knowledge base

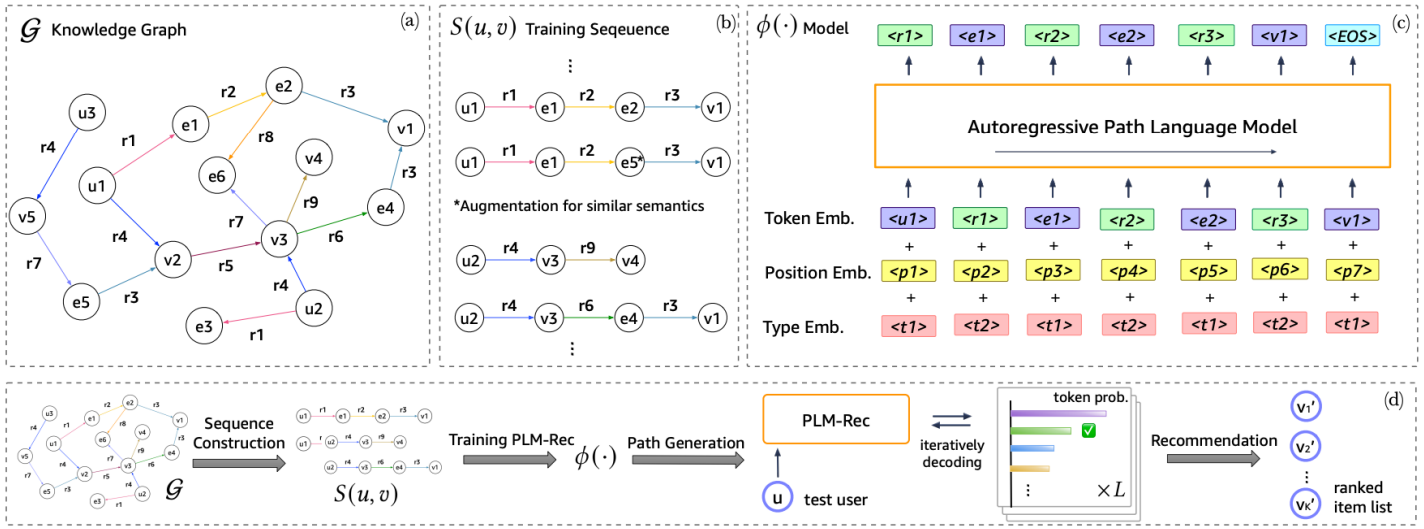


Image credit to [1]

[1] Geng, Shijie, et al. “Path language modeling over knowledge graphs for explainable recommendation." WWW’22.

Controllability

- The lack of controllability may make LLM generate inappropriate content
  - Harassing content
  - Misinformation
- Control the feature of an explanation [1]
- Control the feature of a recommended item
  - Price
  - Color
  - Brand

Rating	Feature	Explanation
4		***The rooms are spacious and the bathroom has a large tub.***
3.90	bathroom	The **bathroom** was large and had a separate shower.
	tub	The bathroom had a separate shower and **tub**.
	rooms	The **rooms** are large and comfortable.

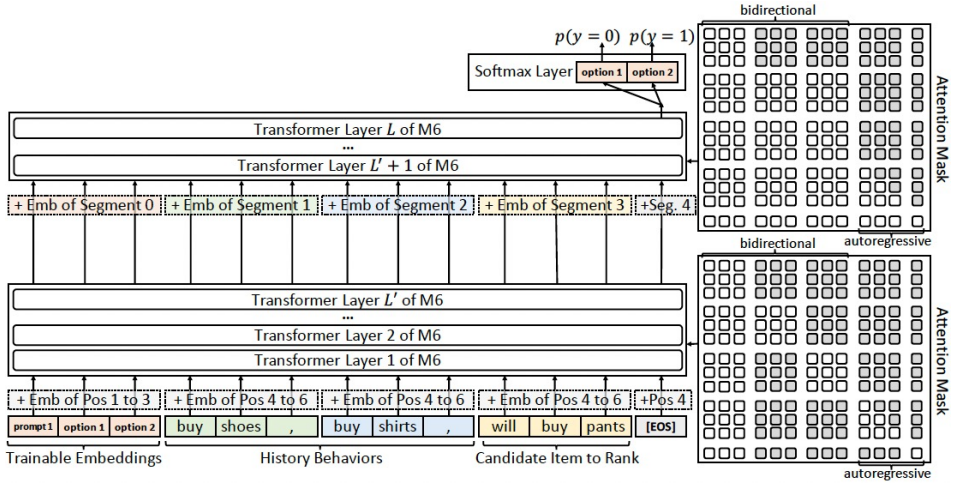
Image credit to [1]

[1] Li, Lei, Yongfeng Zhang, and Li Chen. "Generate neural template explanations for recommendation." CIKM'20.

Inference Efficiency

- Recommender systems are latency-sensitive but LLM contain a huge amount of parameters
  - Pre-compute the first few layers of LLM and cache the results [1]
  - Remove prompt template [2]
- Much room to improve LLM's inference efficiency

Image credit to [1]



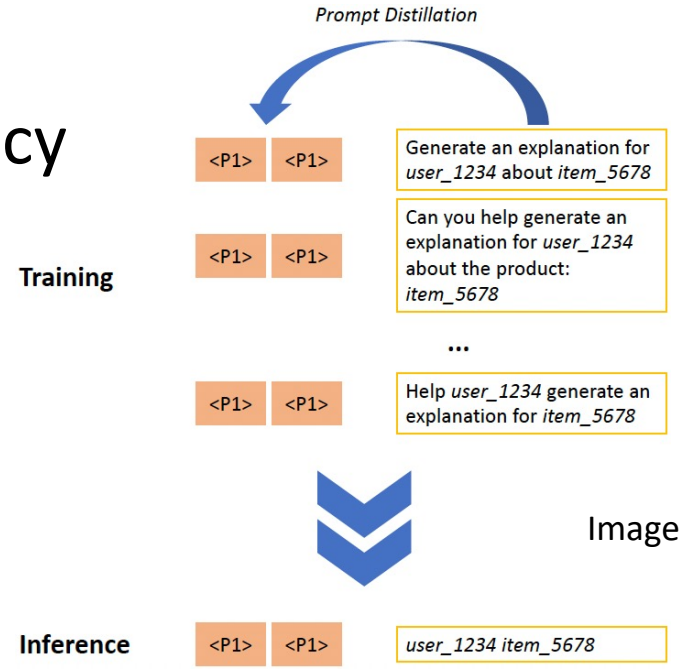


Image credit to [2]

[1] Cui, Zeyu, et al. "M6-rec: Generative pretrained language models are open-ended recommender systems." arXiv'22.
[2] Li, Lei, et al. "Prompt Distillation for Efficient LLM-based Recommendation." CIKM'23.

Multimodal Recommendation

- Data of other modalities can also be represented as a sequence of tokens
  - Suno [1] for audio generation
  - SORA [2] for video generation
  - DALL·E [3] for image generation
    - Visual explanation [4]
    - Product design [5]
- Create new items when existing items do not match users' interests

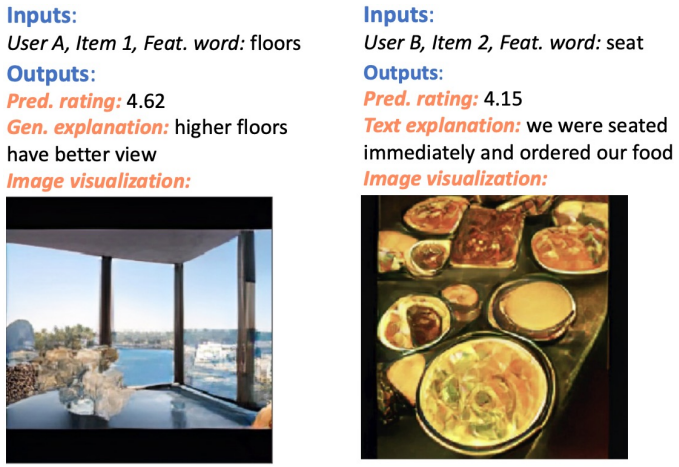


Image credit to [4]

Image credit to [5]

[1] https://suno.com/
[2] https://openai.com/sora
[3] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." ICML'21.
[4] Geng, Shijie, et al. "Improving personalized explanation generation through visualization." ACL'22.
[5] Cui, Zeyu, et al. "M6-rec: Generative pretrained language models are open-ended recommender systems." arXiv'22.

Cold-start Recommendation

- Recommender systems may fail to make recommendations for new users or items
  - Limited or no interactions
- Users' preferences and items' attributes can be represented in natural language
- LLM learned world knowledge during pre-training
  - Perform recommendation even not fine-tuned on recommendation-specific data

Image credit to [1]

[1] https://zhuanlan.zhihu.com/p/376798647

Conclusion

- Elaborated on the advantages of LLM-based generative recommendation
  - Generalized ID's definition
  - Summarized ID creation methods
- Provided general formulation for each generative recommendation task and reviewed the progress
- Acknowledged challenges that are worth exploration
- Research in line with the trend of AI
  - Discriminative AI -> generative AI

Q&A

Thank you!

csleili@comp.hkbu.edu.hk

lileipisces.github.io