

05

Summary

Open Challenges and Beyond

Summary

Scaling Law:

- **larger model + larger dataset → better performance**

Most large models are generative

- (LLMs, Text2Video Models)

Summary

Scaling Law:

- **larger model + larger dataset → better performance**

Most large models are generative

- (LLMs, Text2Video Models)



Large generative rec models?

Summary



How to get a large generative rec model?

- Pre-trained model (e.g., LLMs) → Adaptation;
- From scratch;

Paradigms

Adapt Pre-trained Models

Train from Scratch

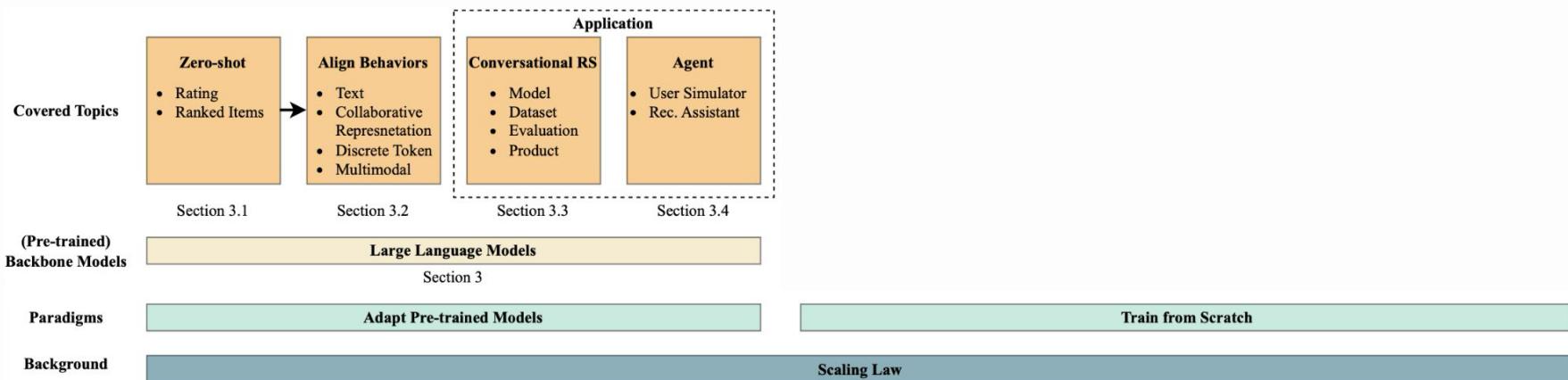
Background

Scaling Law

Summary

Adaptation

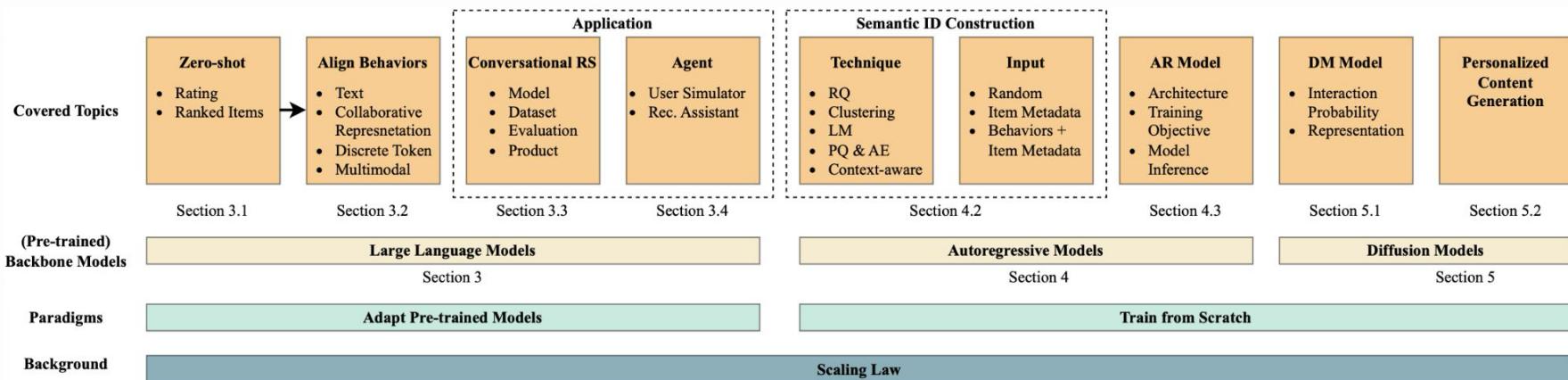
Mainly LLM-based recommendations



Summary

From Scratch

- Autoregressive models (e.g., semantic ID-based);
- Diffusion models;



Summary

Open Challenges

Open Challenges	Inference Efficiency	Model Updating	Item Tokenization	Emergent Ability	Test-time Scaling & Reasoning	Unified Retrieval and Ranking	
	Section 6.1	Section 6.2	Section 6.3	Section 6.4	Section 6.5	Section 6.6	
Covered Topics	Zero-shot <ul style="list-style-type: none">RatingRanked Items → Align Behaviors <ul style="list-style-type: none">TextCollaborative RepresentationDiscrete TokenMultimodal	Application Conversational RS <ul style="list-style-type: none">ModelDatasetEvaluationProduct Agent <ul style="list-style-type: none">User SimulatorRec. Assistant	Semantic ID Construction Technique <ul style="list-style-type: none">RQClusteringLMPQ & AEContext-aware Input <ul style="list-style-type: none">RandomItem MetadataBehaviors + Item Metadata		AR Model <ul style="list-style-type: none">ArchitectureTraining ObjectiveModel Inference	DM Model <ul style="list-style-type: none">Interaction ProbabilityRepresentation	Personalized Content Generation
(Pre-trained) Backbone Models	Section 3.1	Section 3.2	Section 3.3	Section 4.2	Section 4.3	Section 5.1	Section 5.2
	Large Language Models			Autoregressive Models		Diffusion Models	
Paradigms	Adapt Pre-trained Models			Train from Scratch			
Background	Scaling Law						

Open Challenges

Part 1: What becomes harder?

Comparing to traditional RecSys, what challenges may large generative models face?

Open Challenges

Inference Efficiency

Section 6.1

Model Updating

Section 6.2

Item Tokenization

Section 6.3

Emergent Ability

Section 6.4

Test-time Scaling & Reasoning

Section 6.5

Unified Retrieval and Ranking

Section 6.6

Open Challenges

Part 1: What becomes harder?

Comparing to traditional RecSys, what challenges may large generative models face?

Part 2: What becomes possible?

What new opportunities may large generative models unlock for recommender systems?

Open Challenges

Inference Efficiency

Section 6.1

Model Updating

Section 6.2

Item Tokenization

Section 6.3

Emergent Ability

Section 6.4

Test-time Scaling & Reasoning

Section 6.5

Unified Retrieval and Ranking

Section 6.6

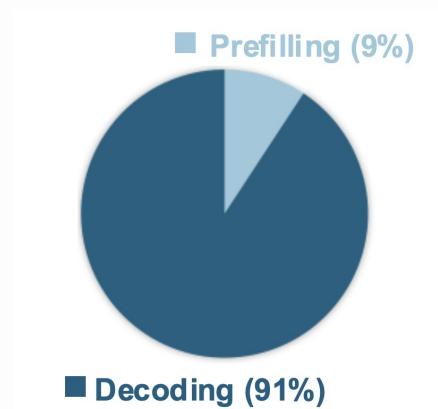
Part 1: What Becomes Harder?

Comparing to traditional RecSys, what challenges may large generative models face?

Inference Efficiency

Retrieval Models: **K Nearest Neighbor Search**

Generative Models (e.g., AR models): **Beam Search**



Inference Efficiency

How to accelerate LLMs? **Speculative Decoding**

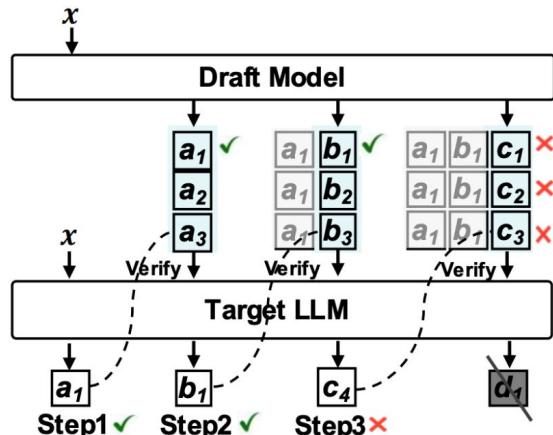
- Use a “cheap” model to generate candidates
- “Expensive” model can **accept** or **reject** (and perform inference if necessary)

```
[START] japan ' s benchmark bond n
[START] japan ' s benchmark nikkei 22 75
[START] japan ' s benchmark nikkei 225 index rose 22 76
[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points
[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points , or 0 1
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 9859
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 , 79 in
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in tokyo late
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]
```

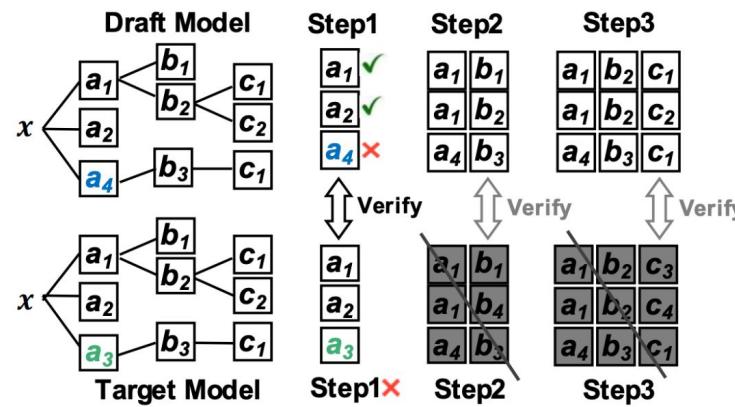
Inference Efficiency

Speculative decoding for generative rec? 

N-to-K verification



(b) N-to-1 Verification of Traditional SD (N=3)



(c) N-to-K Verification of SD with Beam Search (N=K=3)

Inference Efficiency

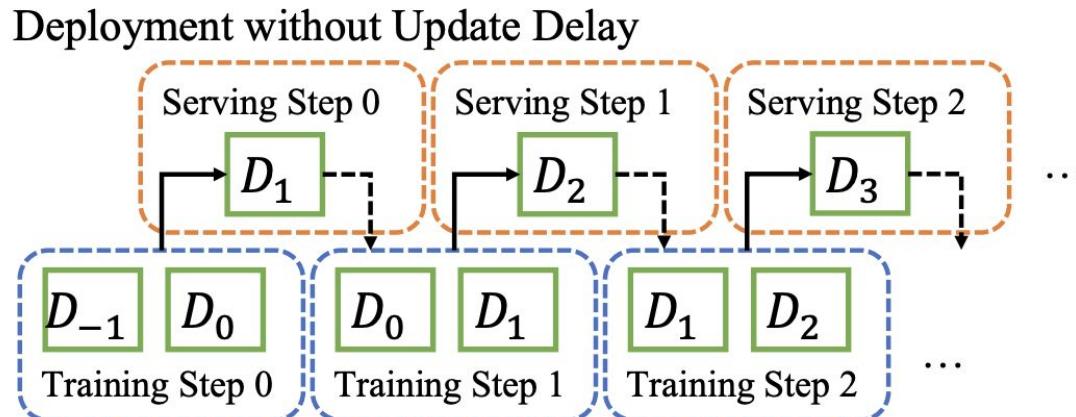
In addition to single-model acceleration methods,
what about “**serving throughout**”?

Example:  LLM offers solutions for high-throughput
and memory-efficient inference and serving

What's unique for generative rec?

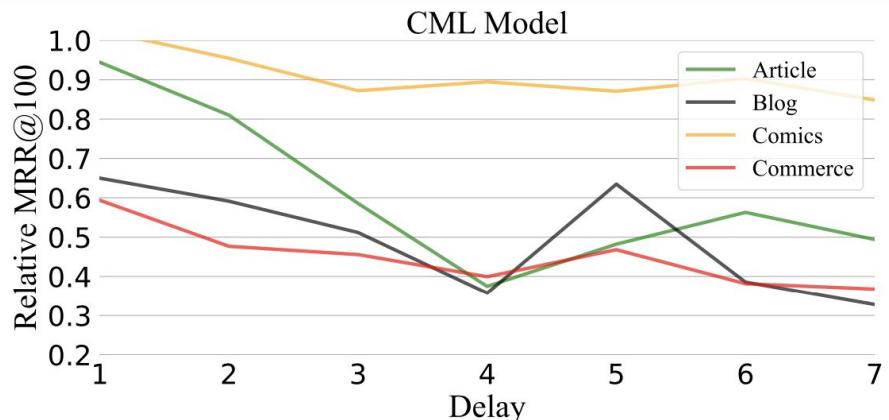
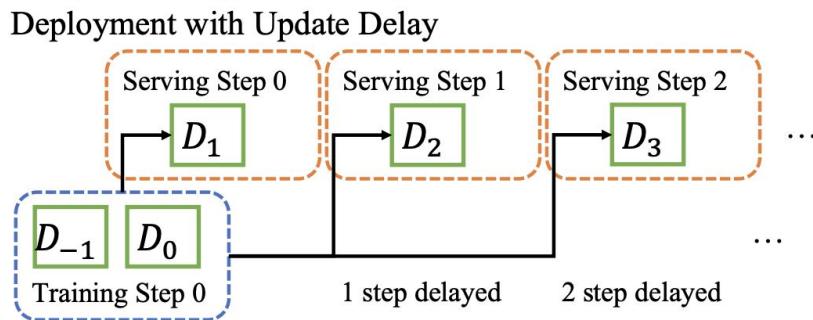
Timely Model Update

Recommendation models favor **timely updates**



Timely Model Update

Delayed updates lead to performance degradation



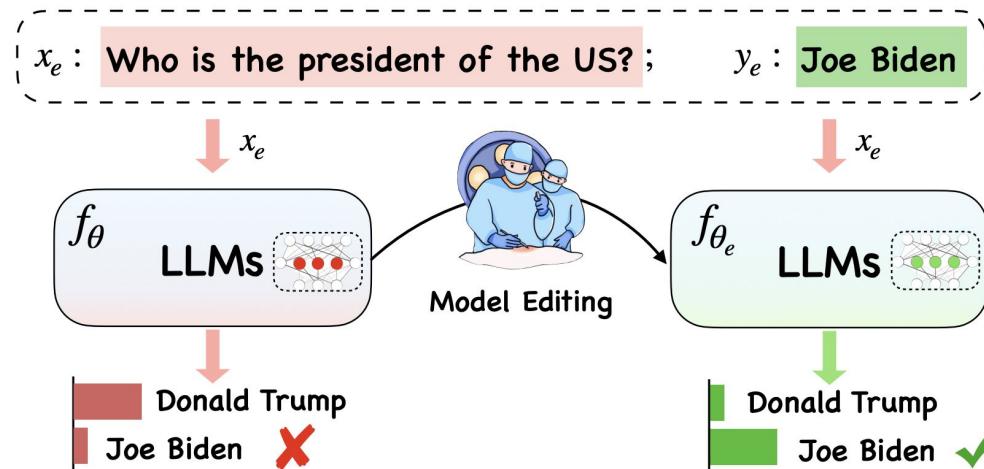
Timely Model Update

How to update large generative rec models timely?

(Frequently retraining large generative models may be resource consuming)

Timely Model Update

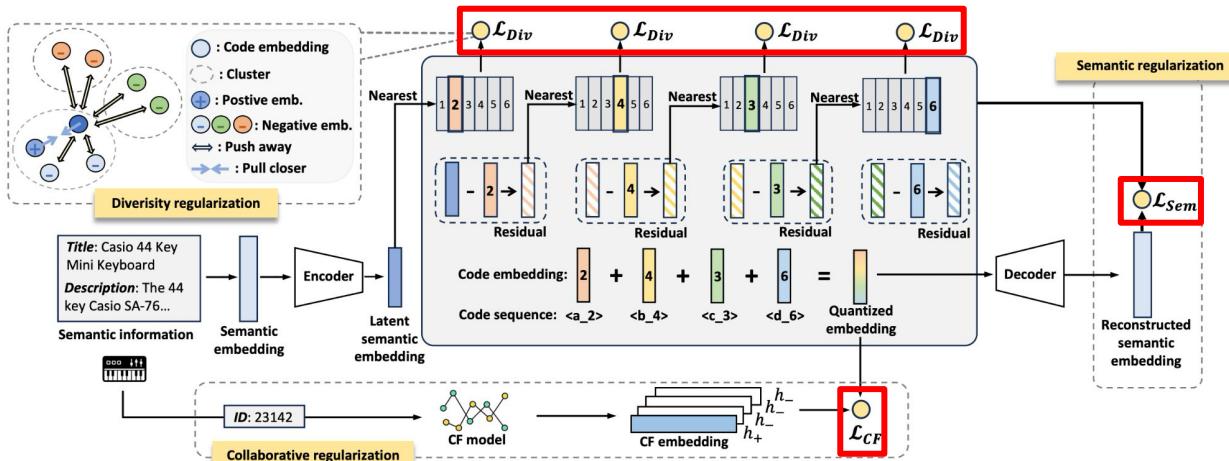
How to update large generative rec models timely?



Knowledge editing?

Item Tokenization

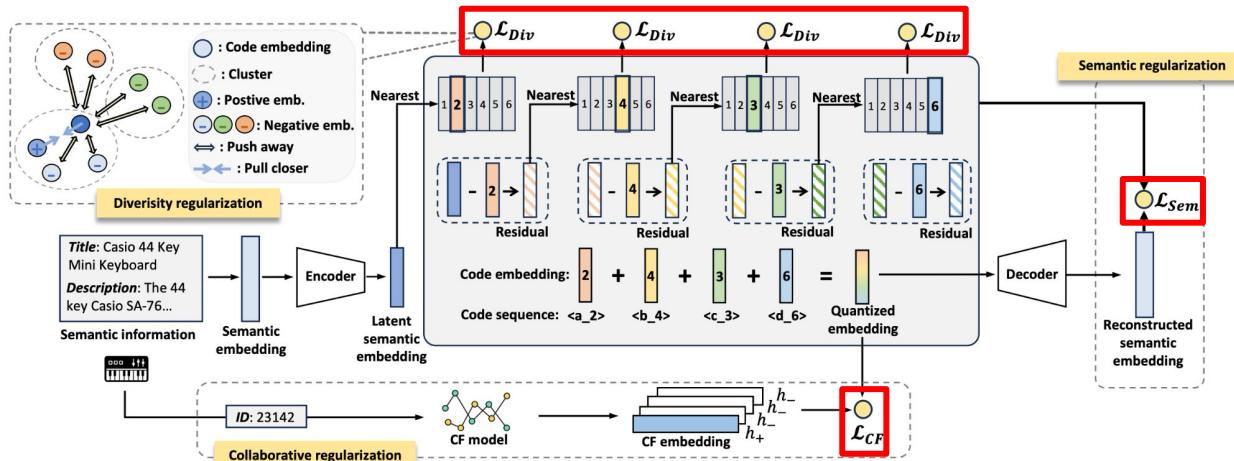
Multiple objectives for optimizing item tokenization ...



Item Tokenization

Multiple objectives for optimizing item tokenization ...

But **none** of them is **directly related to rec performance**



Item Tokenization

reconstruction loss ≠ downstream performance

How to connect tokenization objective with
recommendation performance?

Zipf's distribution? Entropy? Linguistic metrics?

Item Tokenization

Language Tokenization

2014~2015:

Word / Char

Context-independent ⇒ Context-aware

Item Tokenization

Language Tokenization

2014~2015:

Word / Char

2016~present:

BPE / WordPiece

Context-independent ⇒ Context-aware

Item Tokenization

Language Tokenization

2014~2015:

Word / Char

2016~present:

BPE / WordPiece

Context-independent ⇒ Context-aware

SemID Construction

2023~2024:

RQ / PQ / Clustering /
LM-based Generator

Item Tokenization

Language Tokenization

2014~2015:

Word / Char

2016~present:

BPE / WordPiece

Context-independent ⇒ Context-aware

SemID Construction

2023~2024:

RQ / PQ / Clustering /
LM-based Generator

2025:

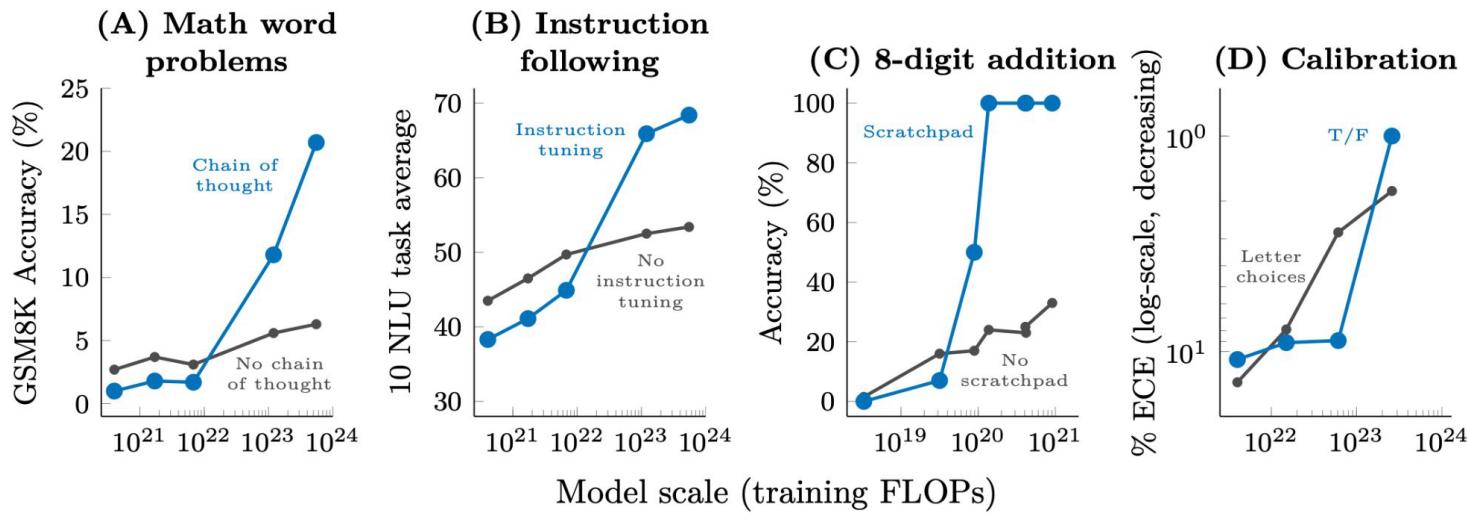
ActionPiece / ?

Part 2: What Becomes Possible?

What new opportunities may large generative models unlock for recommender systems?

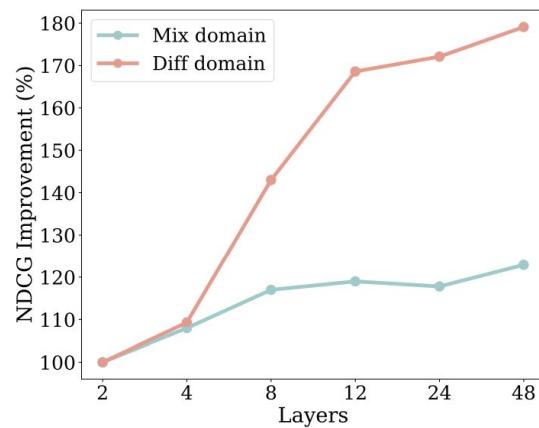
Emergent Ability

Abilities not present in smaller models but is present in larger models

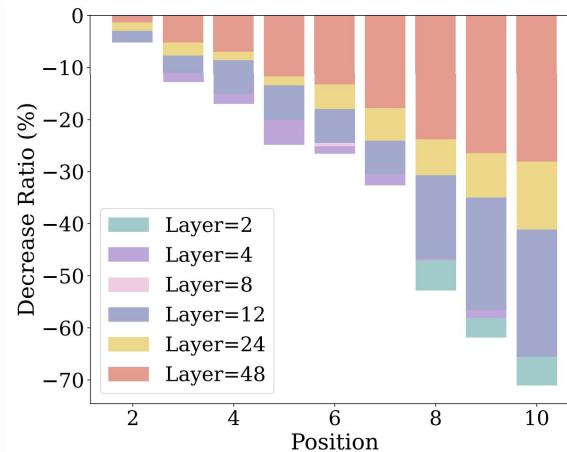


Emergent Ability

Do we have **emergent abilities** in large generative recommendation models?



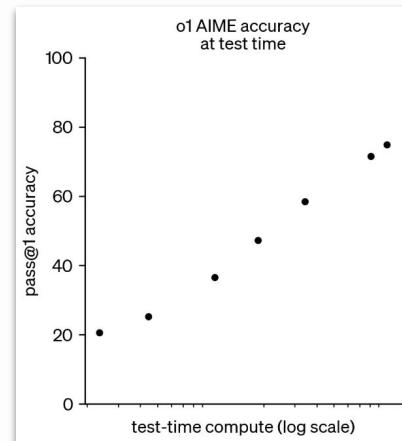
Cross-domain



Trajectory Prediction

Test-time Scaling

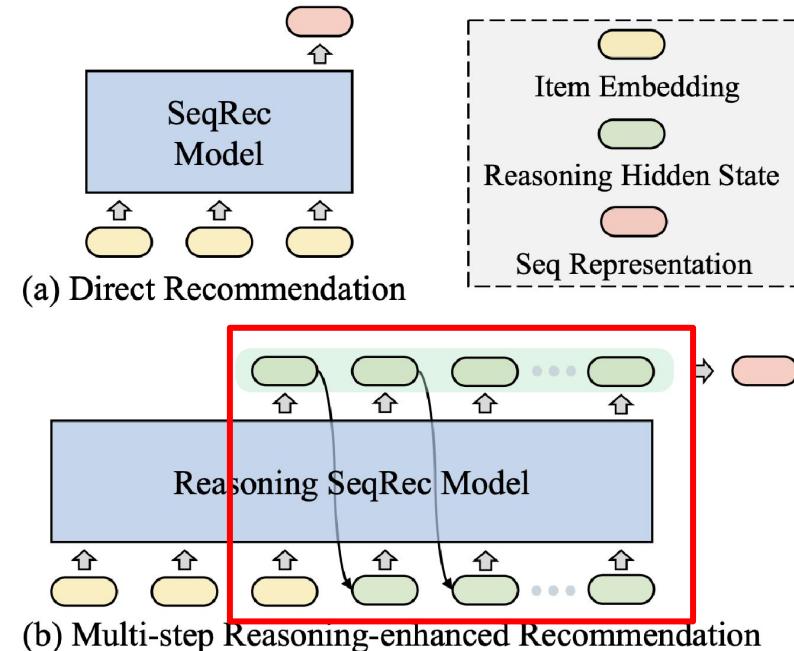
There have been explorations on **model / data scaling** of recommendation models



Test-time scaling is still under exploration

Test-time Scaling

Reasoning over **latent hidden states** to scale up test-time computation.



Large Diffusion Model

End-to-end personalized content generation with diffusion model.

- Existing work largely uses **pretrained diffusion model** to generate personalized content.
- An **end-to-end training** could bring mere personalization.

Agentic

Next: personalization & agentic system

Customize ChatGPT

Introduce yourself to get better, more personalized responses ⓘ

What should ChatGPT call you?

What do you do?

What traits should ChatGPT have? ⓘ

I love Iron Man. Recommend some scientific movies for me.

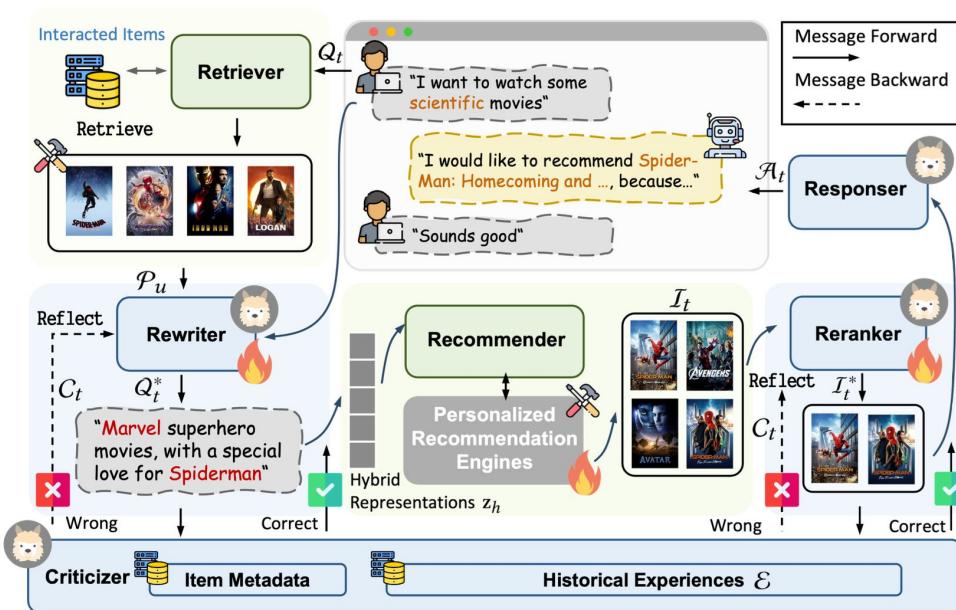
Q. Completed
44s · 22 sources

- Thinking
- Exploring scientific movies
- Evaluating movie themes

- All these movies have strong scientific elements, either based on real events or accurate concepts. The Imitation Game shows early computing, while Interstellar uses real physics theories.
- I'm checking if they fit the user's taste for adventure and innovation, like Iron Man. The Martian's survival story and 2001's space exploration seem to match well.
- Hidden Figures tells a true story of NASA mathematicians, which is inspiring and scientific. I think these choices will engage the user with their blend of science and drama.
- Let me ensure they're not just fiction but grounded in science. The Martian is known for realism, and 2001 incorporates real science from Arthur C. Clarke.
- The Martian (2015) follows an astronaut using science to survive on Mars. It's problem-solving and engineering, much like Tony Stark's character, and focuses on botany and space exploration.
- Hidden Figures (2016) celebrates African-American women at NASA, focusing on math and aerospace engineering. Its inspiring story aligns with Iron Man's themes of intelligence and overcoming challenges.

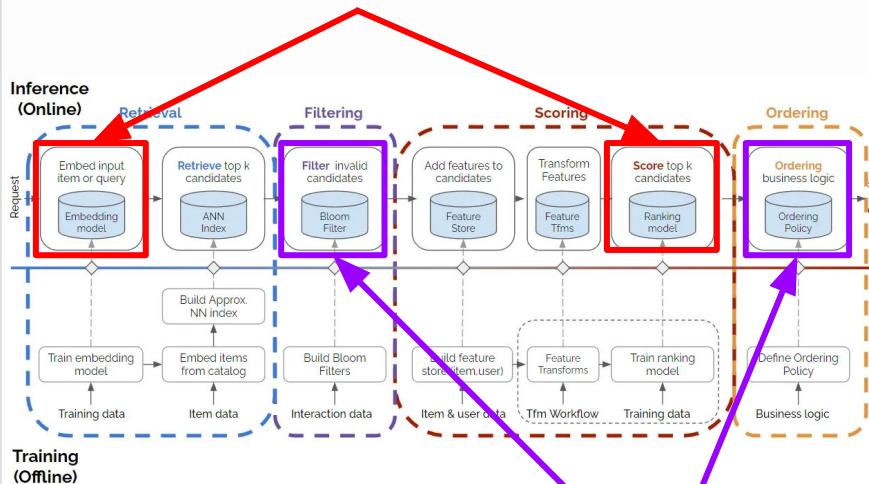
22 web pages

Agentic



Large agentic system
for precise user
modeling and better
recommendation

Unified Retrieval & Ranking models



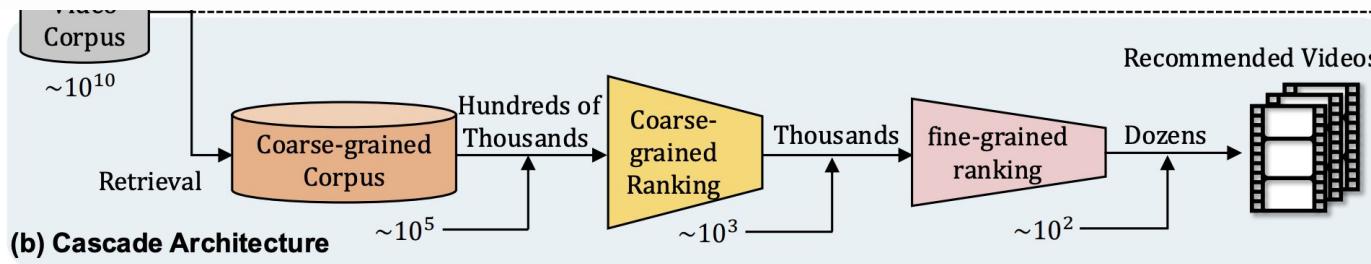
Complicated Architecture

- Difficult to be optimized in an **end-to-end** way
- Latency** between / within different modules

rules, strategies, heuristics

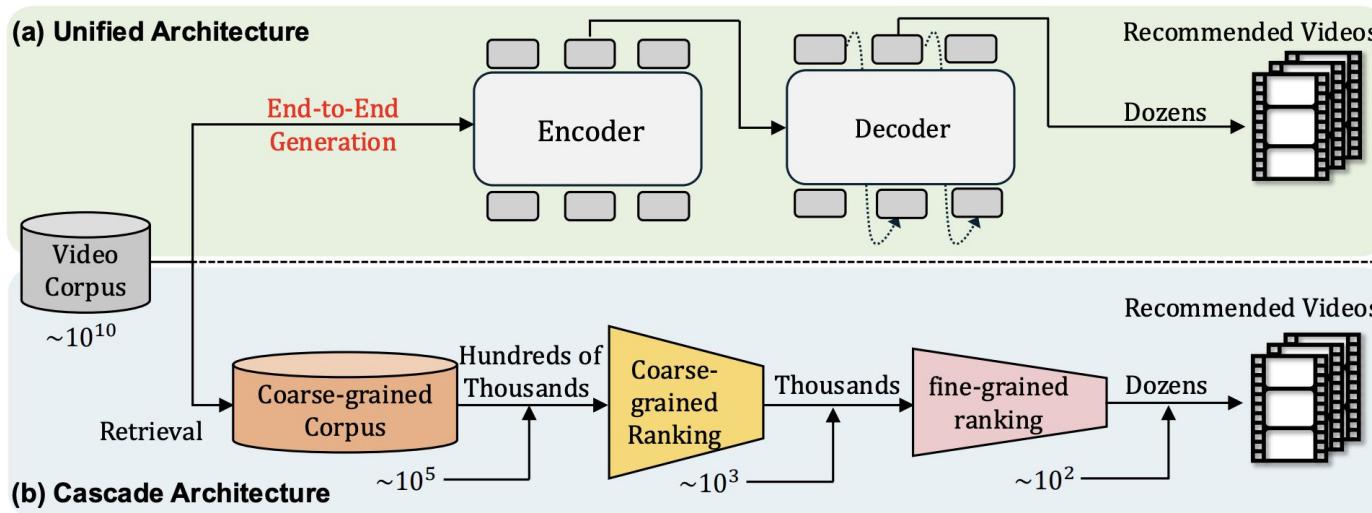
Unified Retrieval & Ranking

Is it possible to replace traditional cascade architecture



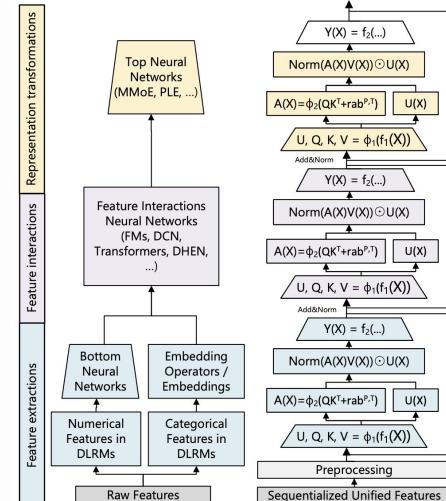
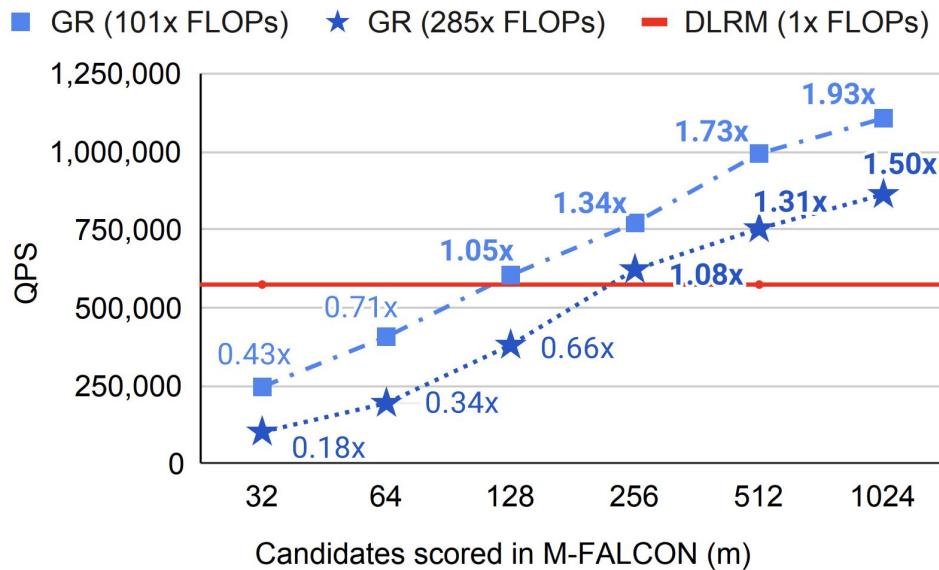
Unified Retrieval & Ranking

Is it possible to replace traditional cascade architecture with a **unified generative model**?



Unified Retrieval & Ranking

Better throughout when ranking more candidates



Q & A

Thank you for coming!



Please refer to

large-genrec.github.io

for *slides, paper list,*

- [Papers](#)
 - [Surveys](#)
 - [LLM-based Generative Recommendation](#)
 - [LLM as Sequential Recommender](#)
 - [Early Efforts: Zero-shot Recommendation with LLMs](#)
 - [Aligning LLMs for Recommendation](#)
 - [Training Objectives & Inference](#)
 - [LLM as Conversational Recommender & Recommendation Assistant](#)
 - [LLM as User Simulator](#)
 - [Semantic ID-based Generative Recommendation](#)
 - [Semantic ID Construction](#)
 - [Quantization](#)
 - [Hierarchical Clustering](#)
 - [Contextual Action Tokenization](#)
 - [Behavior-aware Tokenization](#)
 - [Language Model-based Generator](#)
 - [Architecture](#)
 - [Dense & Generative Retrieval](#)
 - [Unified Retrieval and Ranking](#)
 - [Aligning with LLMs](#)
 - [Diffusion Model-based Generative Recommendation](#)
 - [Diffusion-enhanced Recommendation](#)
 - [Diffusion as Recommender](#)
 - [Personalized Content Generation with Diffusion](#)
 - [Resources](#)