05

Summary

Open Challenges and Beyond

Action Tokenization

Human-readable Data 👨 👩

Machine-readable Data 🤖

Tokenize

?

Action Tokenization

Human-readable Data 👨 👩

Machine-readable Data 🤖

Tokenize

i3487, i124, i19240, i772

One token per action? 🤔

Action Tokenization

Human-readable Data 👨👩

Tokenize

Machine-readable Data 🤖

Premium Men's Short Sleeve Athletic Training T-Shirt Made of Lightweight Breathable Fabric, Ideal for Running, Gym Workouts, and Casual Sportswear in All Seasons; High-Performance Breathable Cotton Crew Socks for Men with Arch Support, Cushioned Heel and Toe, and Moisture Control, Perfect for Sports, Walking, and Everyday Comfort; Men's Loose-Fit Basketball Shorts with Elastic Drawstring Waistband, Quick-Dry Mesh Fabric, and Printed Number 11 for Professional and Recreational Play; Official Size 7 Composite Leather Basketball Designed for Indoor and Outdoor Use, Deep Channel Design for Enhanced Grip and Ball Control, Ideal for Training and Competitive Matches;

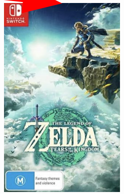
Text description of each action? 🤔

Semantic IDs

(also called: SemID or SID)

A few tokens that jointly index one item.

t3, t321, t643, t1011



Action Tokenization

Can we find a sweet spot between

One token per action	Semantic IDs	Text description of each action
i3487, i124, i19240, i772	t34, t392, t600, t891, t21, t502, t592, t1002, t123, t403, t611, t821, t21, t502, t711, t1022	Premium Men's Short Sleeve Athletic Training T-Shirt Made of Lightweight Breathable Fabric, Ideal for Running, Gym Workouts, and Casual Sportswear in All Seasons; High-Performance Breathable Cotton Crew Socks for Men with Arch Support, Cushioned Heel and Toe, and Moisture Control, Perfect for Sports, Walking, and Everyday Comfort; Men's Loose-Fit Basketball Shorts with Elastic Drawstring Waistband, Quick-Dry Mesh Fabric, and Printed Number 11 for Professional and Recreational Play; Official Size 7 Composite Leather …

Open Challenges

Part 1: What becomes harder?

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

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?

Part 1: What Becomes Harder?

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

Cold-Start Recommendation

Do semantic ID-based models really good at cold-start recommendation?

Cold-Start Recommendation

Do semantic ID-based models really good at cold-start recommendation?

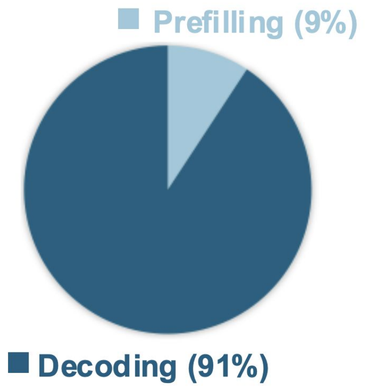
Model	#Params. (M)	Video Games						Cell Phones and Accessories					
		Overall		In-Sample (39.7%)		Unseen (60.3%)		Overall		In-Sample (31.8%)		Unseen (68.2%)	
		R@50	N@50	R@50	N@50	R@50	N@50	R@50	N@50	R@50	N@50	R@50	N@50
UniSRec	2.90	0.0621	0.0200	0.1386	0.0461	0.0118	0.0029	0.0233	0.0077	0.0604	0.0211	0.0060	0.0014
Recformer	233.73	**0.0740**	0.0218	0.1082	0.0333	**0.0514**	**0.0142**	0.0236	0.0070	0.0340	0.0103	**0.0188**	**0.0055**
TIGER	13.26	0.0584	0.0193	**0.1472**	**0.0486**	-	-	0.0232	0.0078	**0.0730**	0.0245	-	-
TIGER_C	13.26	0.0611	0.0198	0.1447	0.0482	0.0061	0.0011	0.0233	0.0078	0.0691	0.0238	0.0019	0.0003
$\textbf{SpecGR}_\textbf{Aux}$	16.16	0.0726	0.0220	0.1399	0.0436	0.0283	0.0078	0.0269	0.0084	0.0722	0.0230	0.0058	0.0015
SpecGR++	13.28	0.0717	**0.0225**	0.1323	0.0439	0.0318	0.0084	**0.0275**	**0.0090**	**0.0730**	**0.0246**	0.0063	0.0017

Ding et al. Inductive Generative Recommendation via Retrieval-based Speculation. AAAI 2026.

Inference Efficiency

Retrieval Models: **K Nearest Neighbor Search**

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

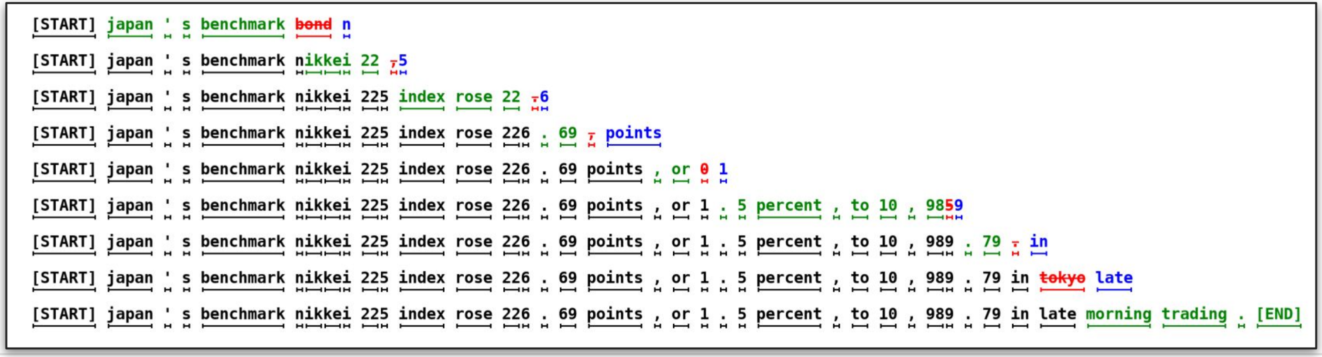


Lin et al. Efficient Inference for Large Language Model-based Generative Recommendation. ICLR 2025.

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)

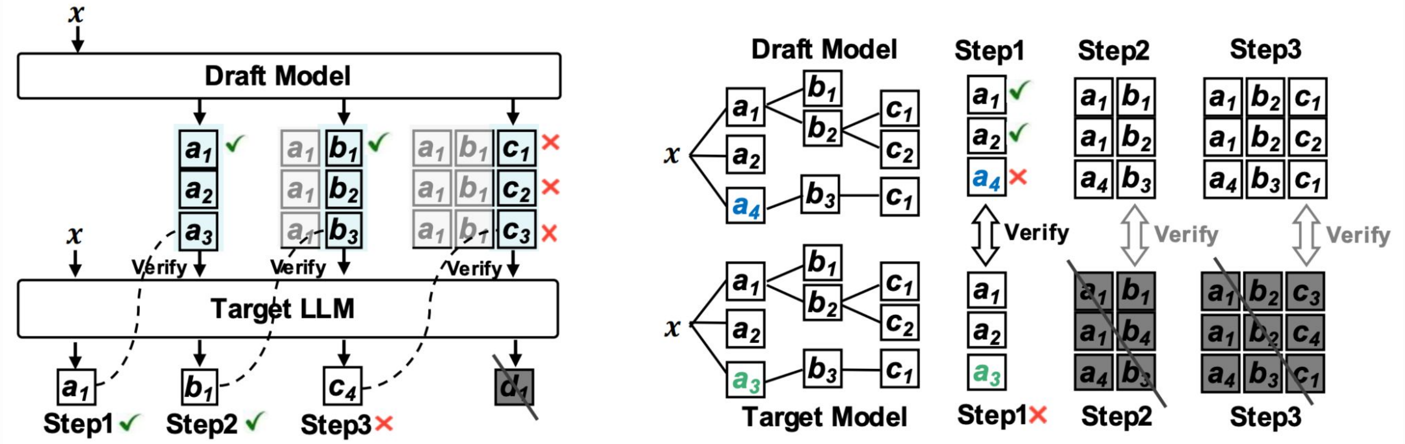


Leviathan et al. Fast Inference from Transformers via Speculative Decoding. ICML 2023.

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)

Lin et al. Efficient Inference for Large Language Model-based Generative Recommendation. ICLR 2025.

Inference Efficiency

In addition to single-model acceleration methods, what about "serving throughout"?

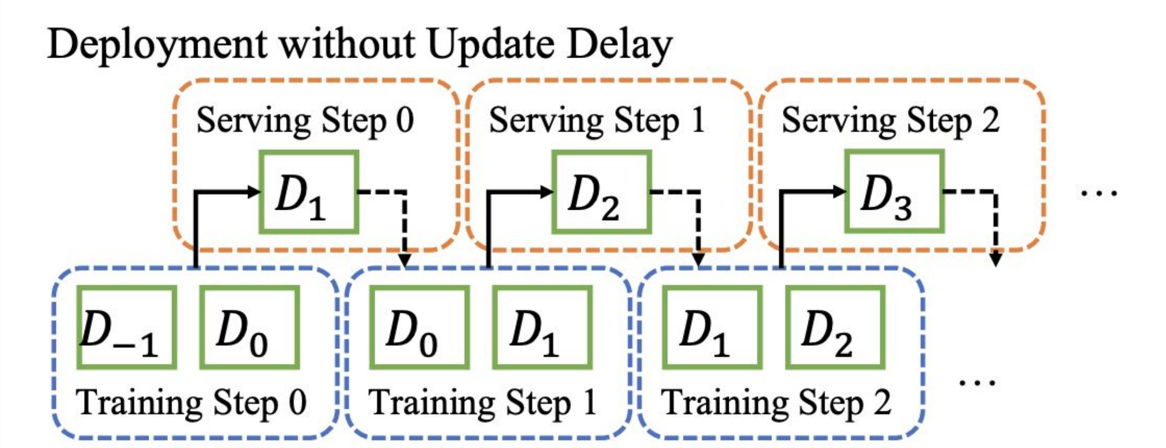
Example: vLLM offers solutions for high-throughput and memory-efficient inference and serving

What's unique for generative rec?

https://github.com/vllm-project/vllm

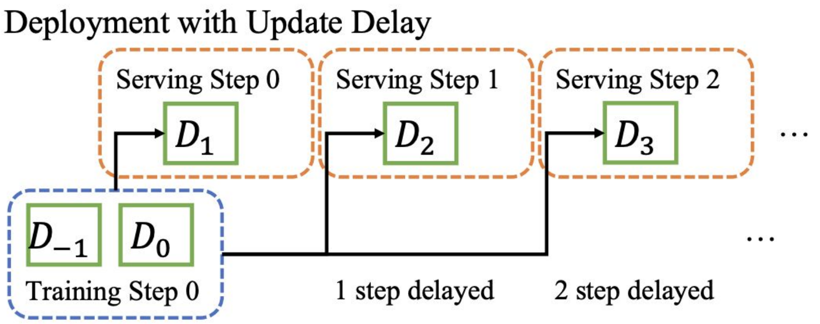
Timely Model Update

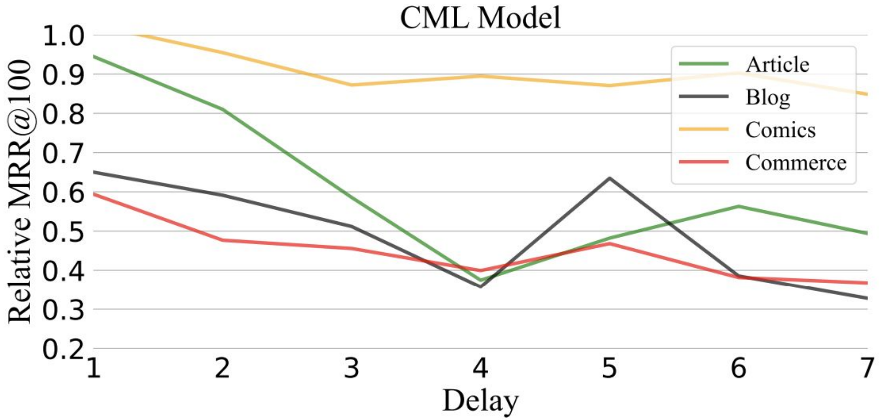
Recommendation models favor timely updates



Timely Model Update

Delayed updates lead to performance degradation





Lee et al. How Important is Periodic Model Update in Recommender Systems? SIGIR 2023.

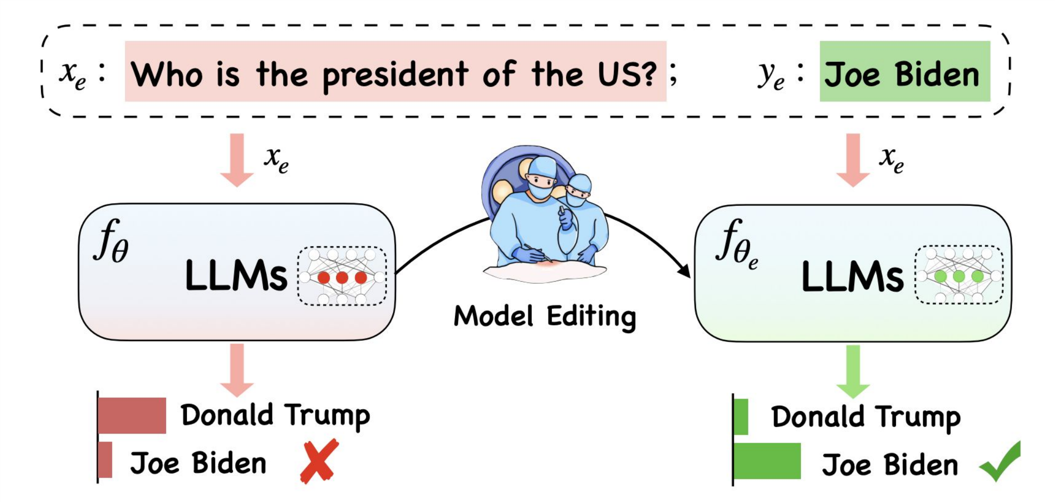
Timely Model Update

How to update large generative rec models timely?

(Frequently retraining large generative models may be resource consuming)

Timely Model Update

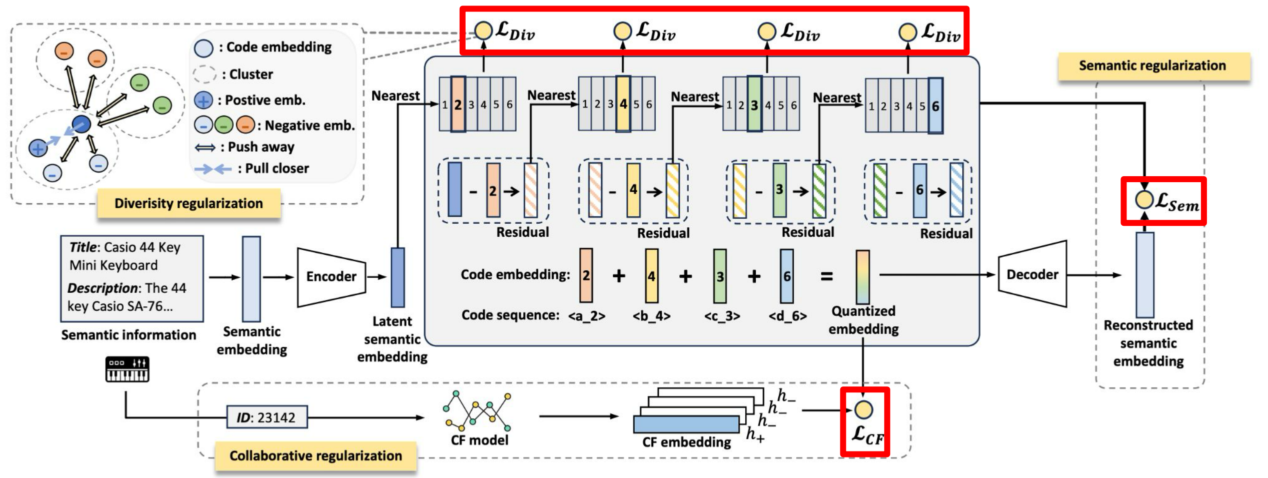
How to update large generative rec models timely?



Knowledge editing?

Item Tokenization

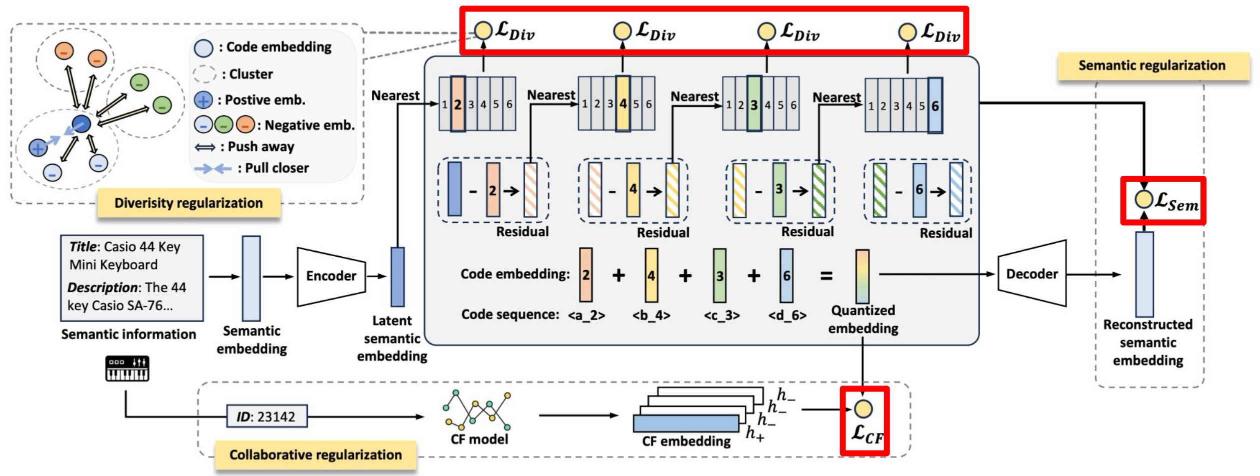
Multiple objectives for optimizing item tokenization …



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 / Pctx / **?**

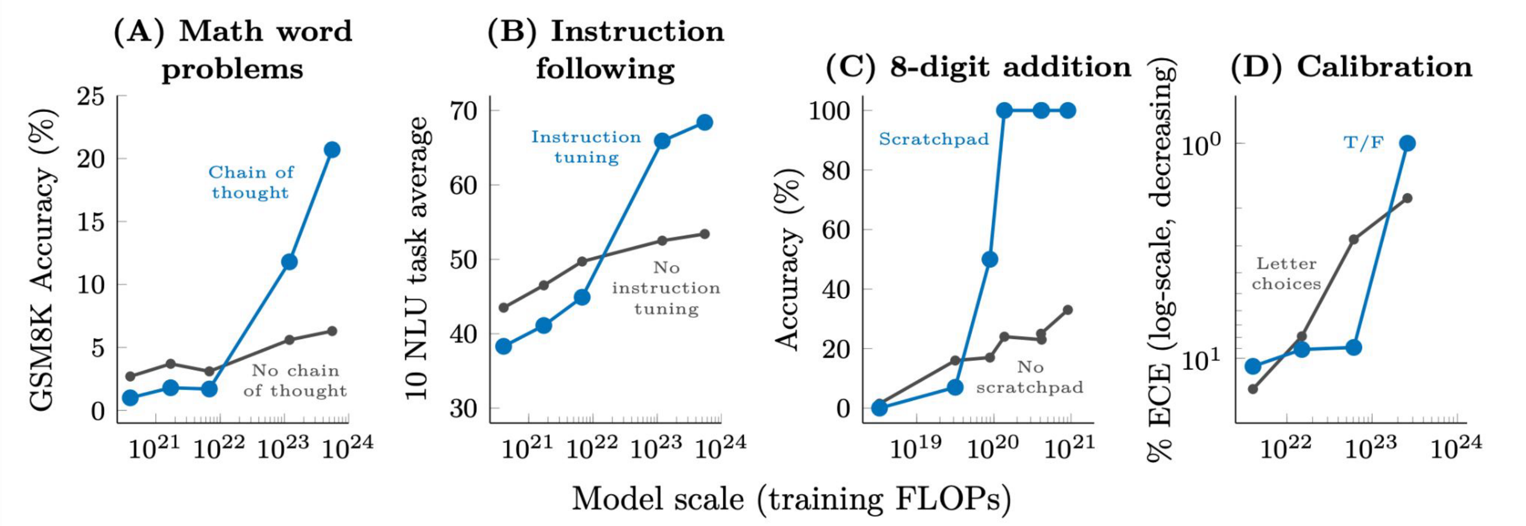
Hou et al. ActionPiece: Contextually Tokenizing Action Sequences for Generative Recommendation. arXiv:2502.13581.

Part 2: What Becomes Possible?

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

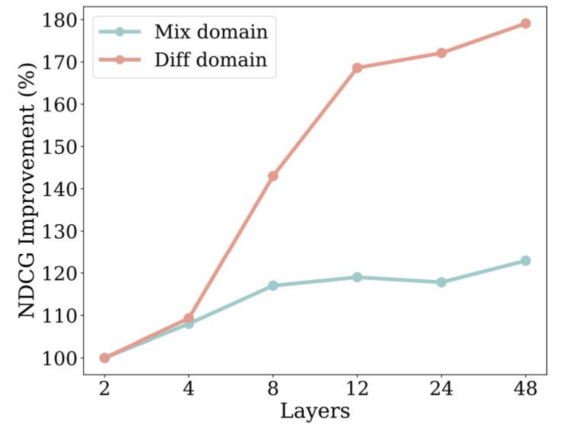
Emergent Ability

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

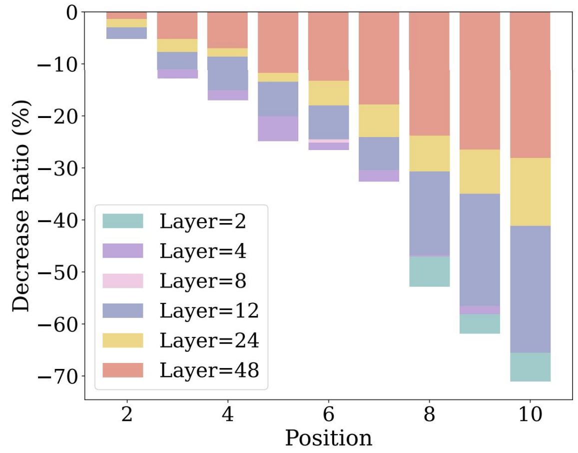


Emergent Ability

Do we have emergent abilities in large generative recommendation models?



Cross-domain

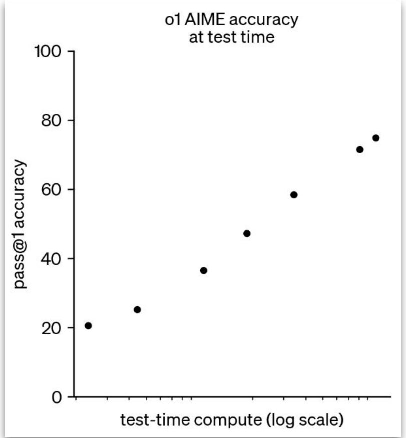


Trajectory Prediction

Zhang et al. Scaling Law of Large Sequential Recommendation Models. RecSys 2024.

Test-time Scaling & Reasoning

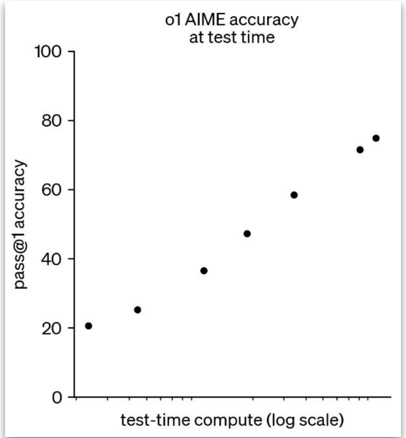
There have been explorations on model / data scaling of recommendation models



Test-time scaling is still under exploration

Test-time Scaling & Reasoning

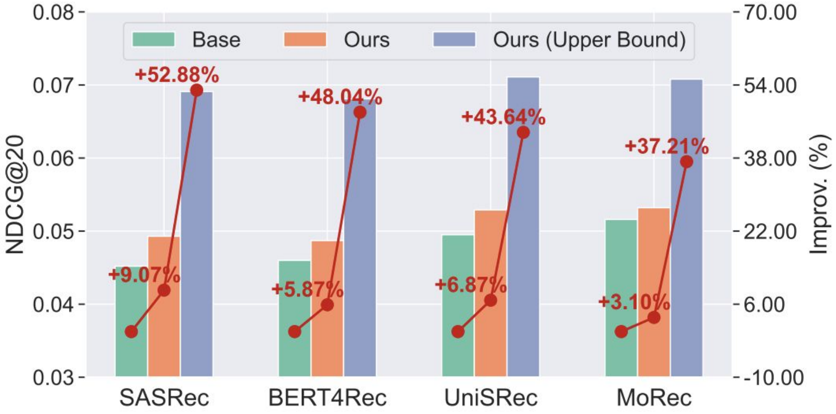
There have been explorations on model / data scaling of recommendation models

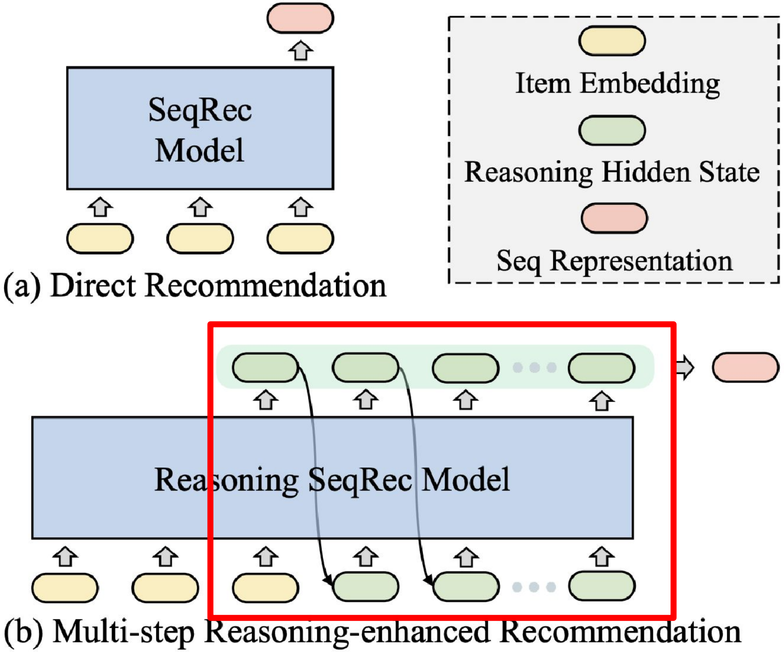


Test-time scaling is ~~still under~~ being actively exploration

Test-time Scaling & Reasoning

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





Tang et al. Think Before Recommend: Unleashing the Latent Reasoning Power for Sequential Recommendation. arXiv:2503.22675.

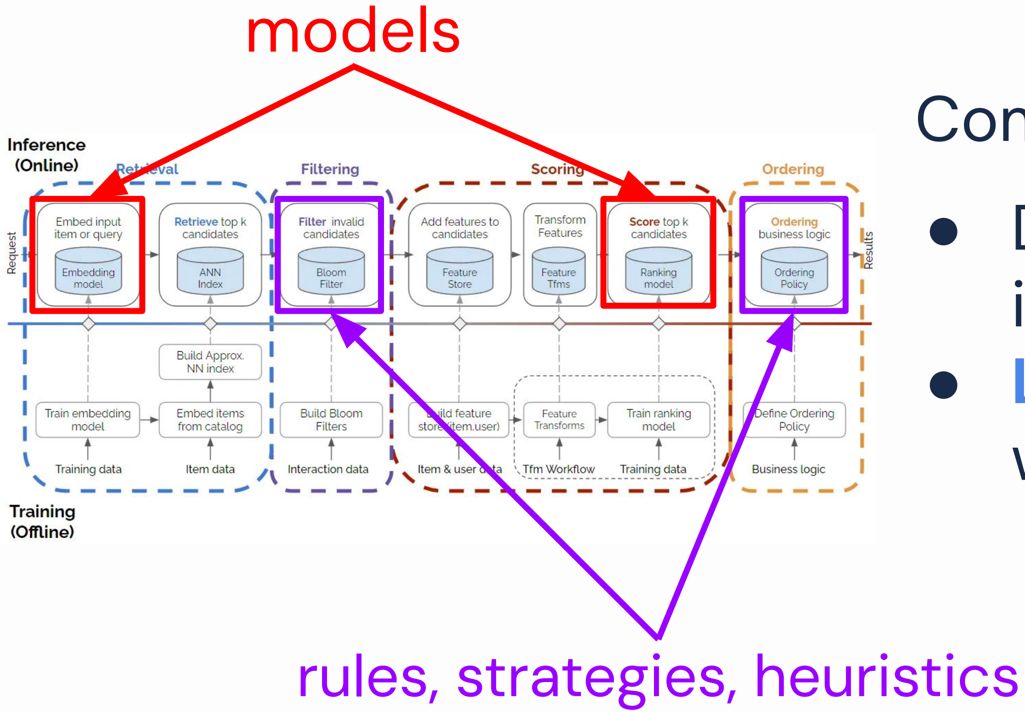
Test-time Scaling & Reasoning

Reasoning over

text tokens

Liu et al. OneRec-Think: In-Text Reasoning for Generative Recommendation. arXiv:2510.11639.

Unified Retrieval & Ranking

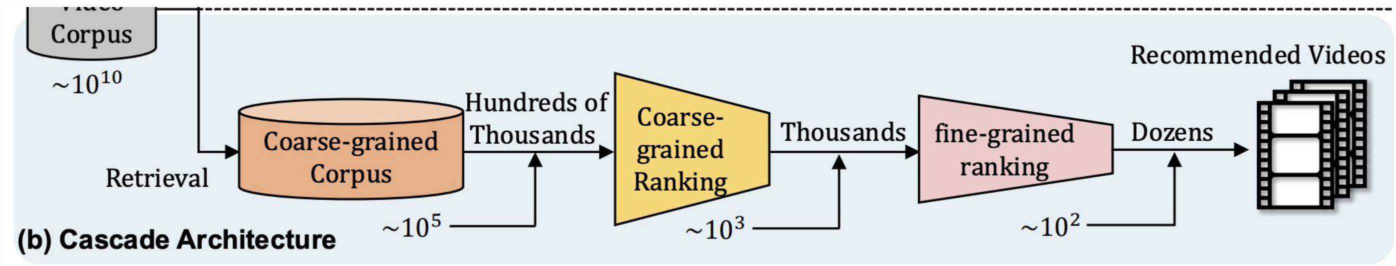


Complicated Architecture

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

Unified Retrieval & Ranking

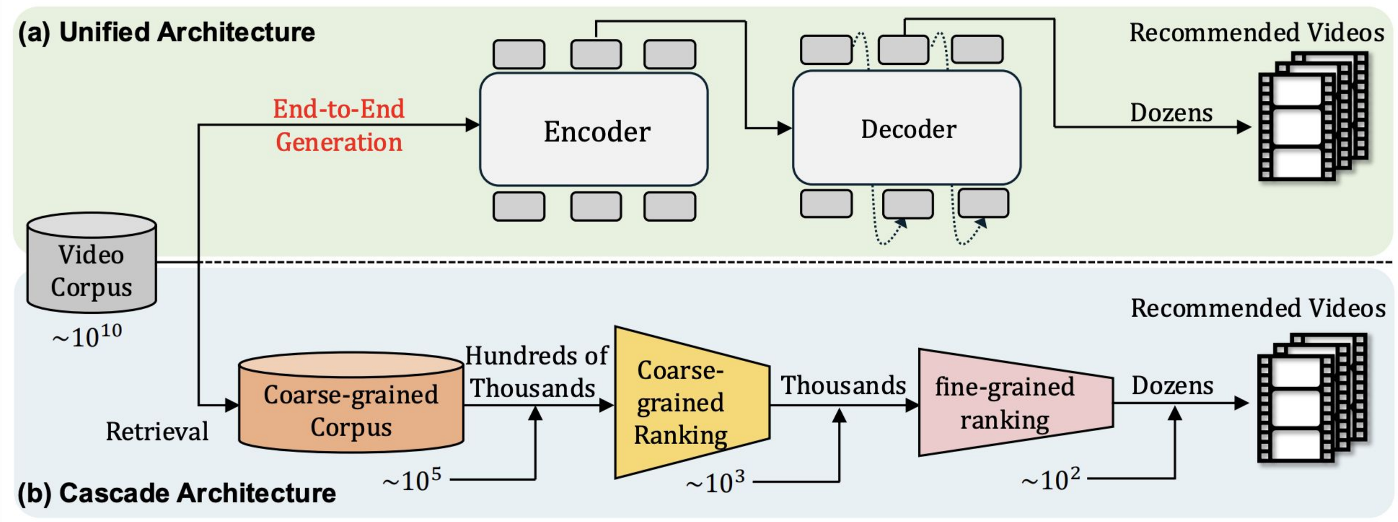
Is it possible to replace traditional cascade architecture



Deng et al. OneRec: Unifying Retrieve and Rank with Generative Recommender and Iterative Preference Alignment. arXiv:2502.18965.

Unified Retrieval & Ranking

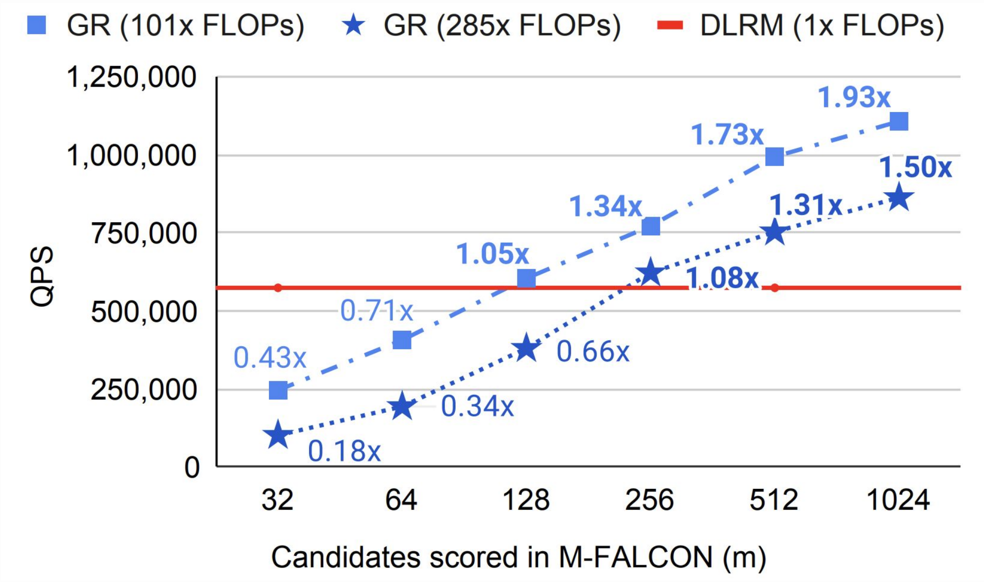
Is it possible to replace traditional cascade architecture with a **unified generative model**?

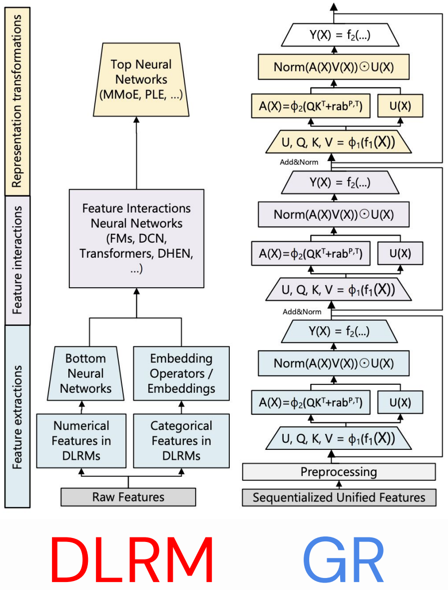


Deng et al. OneRec: Unifying Retrieve and Rank with Generative Recommender and Iterative Preference Alignment. arXiv:2502.18965.

Unified Retrieval & Ranking

Better throughout when ranking more candidates





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

Q & A

Thank you for coming!

Please kindly refer to

large-genrec.github.io

for *slides*, *paper list*, ……