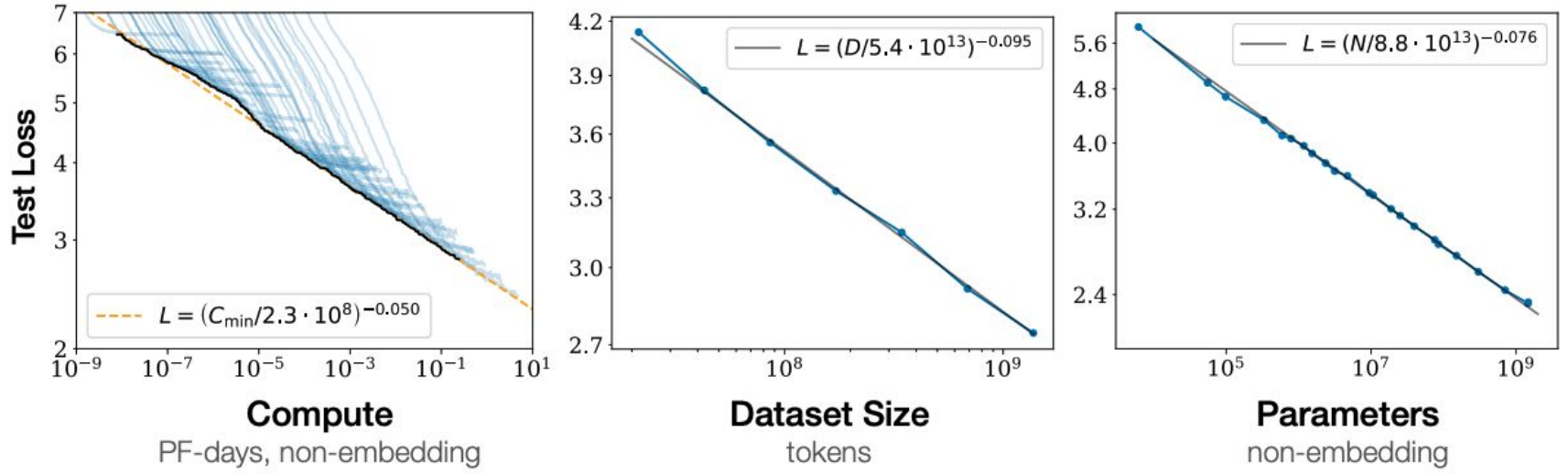
01

Introduction

of Generative Recommendation

Scaling Law as a Pathway towards AGI



Scaling laws provide a framework for understanding how **model size**, **data volume**, and **test-time computing** might lead to advanced AI capabilities.

Understanding Scaling Laws for Recommendation Models. Arxiv 2022.

However ...

Language Modeling

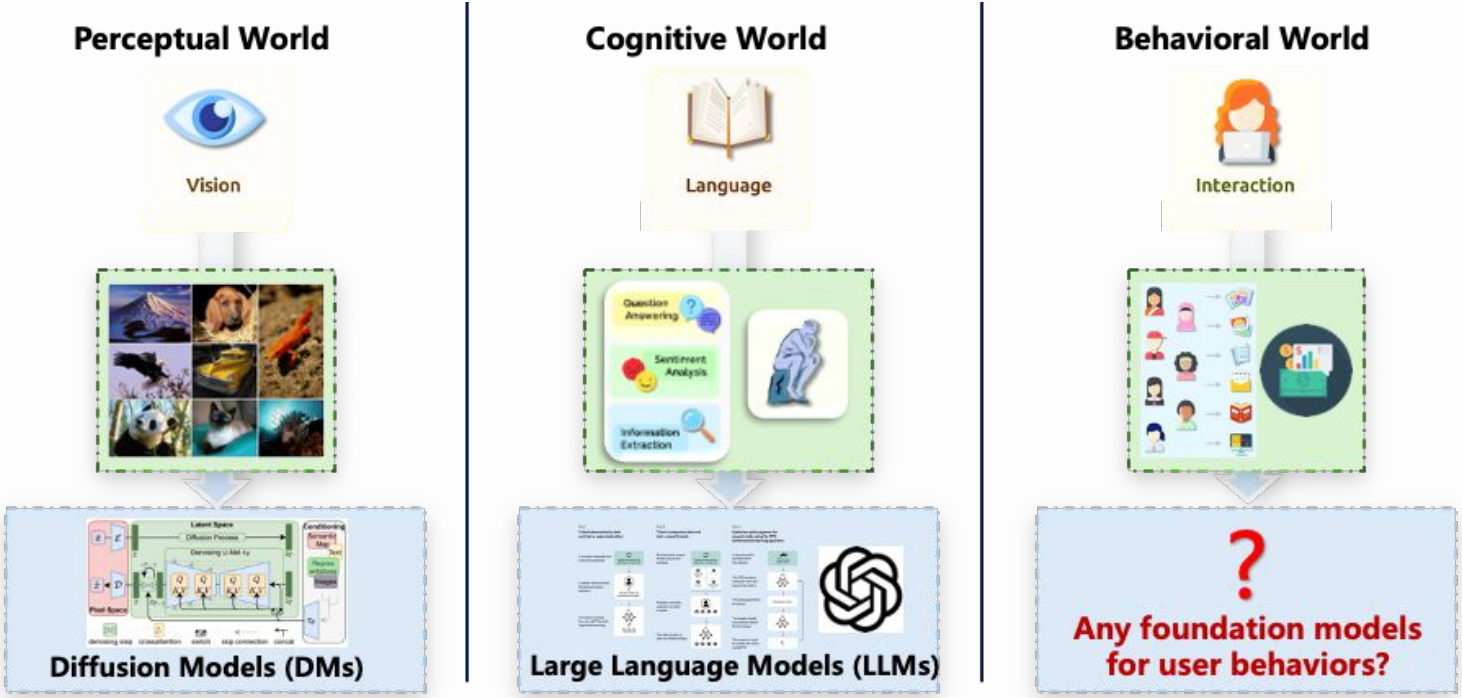
- Dense world knowledge
- Text tokens (Ten thousands level)

User Behavior Modeling

- Sparse user-item interactions
- Items (Billion to trillion level)

Scaling laws rarely apply to traditional recommendation models.

As the Reflection of Real World,



What are Generative Models & Why?

A generative model learns the underlying distribution of data and can generate new samples from it.

Learning

$p_\theta(x)$

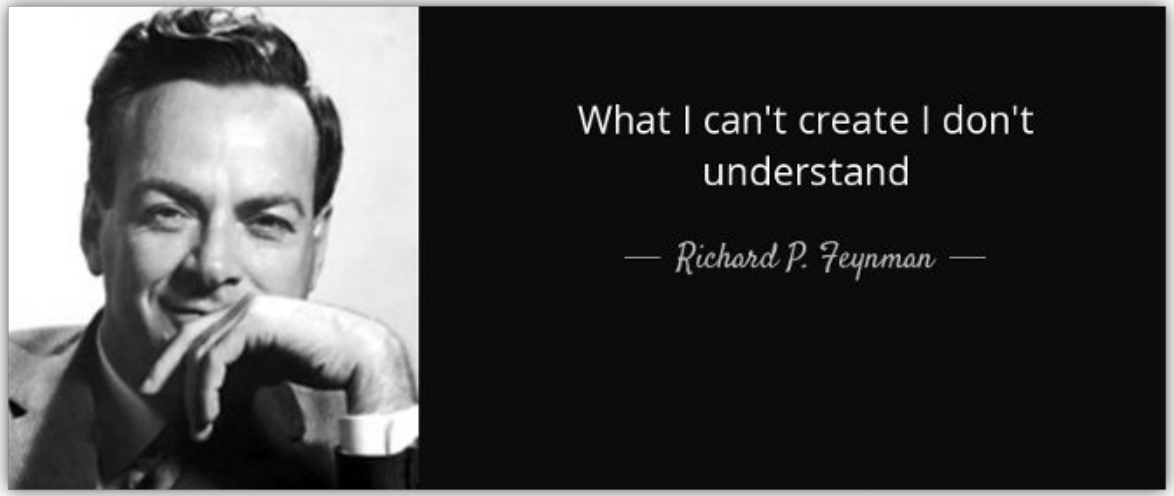
Generating

Training data~$p_{data}(x)$

Data generation distribution

New samples

A Potential Solution: “Generative” Recommendation



“What User Behaviors LLMs can not Generate, LLMs do not Understand.”

Where are We Now?

In language and vision:

- Large language/diffusion models have been established.
- Scaling law has been witnessed.

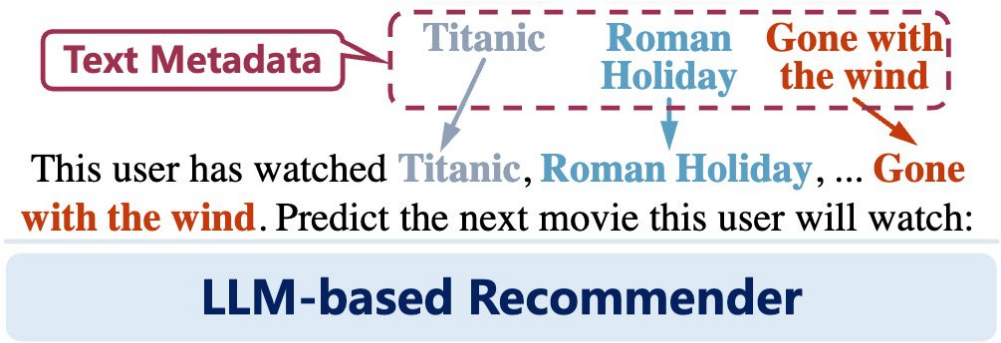
In recommendation:

- Incorporat generative components in traditional recommender.
- Initial attempts on generative recommendation.

Pathways towards Scalable Generative Recommendation

Adapt Pre-trained Models

- Large Language Models



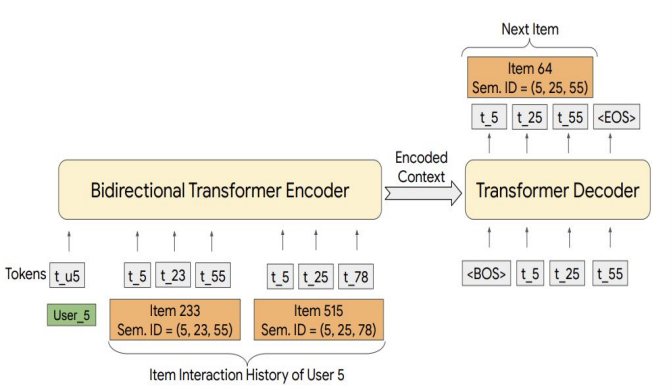
Adapting LLMs for recommendation task

Liao et al. LLaRA: Large Language-Recommendation Assistant. SIGIR 2024.

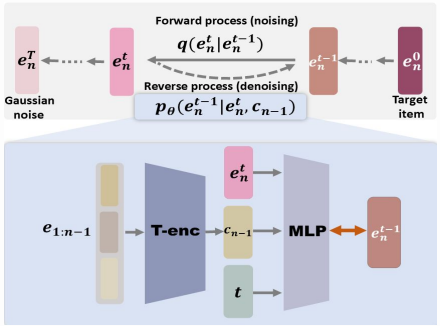
Pathways towards Scalable Generative Recommendation

Train from Scratch

- Autoregressive Models
  - Semantic ID



- Diffusion Models



Schedule Overview

Time (AEST)	Session	Presenter
9:00 - 9:10	Part 1: Background and Introduction	Tat-Seng Chua
9:10 - 10:10	Part 2: LLM-based Generative Recommendation	Leheng Sheng
10:10 - 10:30	Part 3.1: Introduction of Semantic IDs	Yupeng Hou
10:30 - 11:00	Coffee Break & QA Session	
11:00 - 11:40	Part 3.2: SemID-based Generative Recommendation	Yupeng Hou
11:40 - 12:10	Part 4: Diffusion-based Generative Recommendation	Jiancan Wu (proxy speaker of Zhengyi)
12:10 - 12:30	Part 5: Open Challenges and Beyond	Yupeng Hou