Recommender Systems Towards Controllable AI

--- Breaking Filter Bubbles via CF Exploration

Jia Wang
Xi'an Jiaotong–Liverpool University
March 18, 2026

CONTENT

Education: Beijing Jiaotong University – KTH Royal Institute of Technology – Hong Kong Polytechnic University.

Publications: Papers in AAAI, IJCAI, ICDM, TBD, TAI, etc. Involved in projects like Sina Weibo Voting and Hong Kong Airport Scheduling.

王佳

西交利物浦大学

推荐系统，数据挖掘，LLM 大模型

江苏 苏州

Website

Email

Github

Google Scholar

ORCID

I am an Assistant Professor at Xi'an Jiaotong-Liverpool University (XJTLU).

Previously, I received my Ph.D. in Computer Science from the Department of Computing at The Hong Kong Polytechnic University, advised by Prof. Jiannong Cao (曹建农). I earned my M.Sc. from KTH Royal Institute of Technology, advised by Prof. Zhibo Pang (庞智博), and my B.Eng. in Communication Engineering from Beijing Jiaotong University. During my doctoral studies, I was a visiting researcher at University of Southern California (USC), where I collaborated with Prof. Yan Liu.

My research interests include recommender systems, embodied intelligence, and multi-agent systems, with a particular focus on building controllable and trustworthy recommendation world models and recommendation agents.

🔥 News

- *2026.02*: 🎉🎉 We release **UniGenRec**: A Unified Generative Recommendation Toolbox Code Code Stars 38
- *2026.02*: 🎉🎉 One paper about Positive Learning Benchmark was accepted by ICLR 2026, congratulations to Dr. Haiyang Zhang and Qiuyi.
- *2026.01*: 🎉🎉 Two paper about Generative Recommendation Systems was accepted by AAAI 2026, congratulations to Peiyu and Leiqi.
- *2025.12*: Tutorial on "SID-based Generative Recommendation Systems" @JD AI-Lab.
- *2025.11*: 🎉🎉 One papers were accepted by NeurIPS 2025, congratulations to Ming Cheng.
- *2025.10*: Talk on "Controllable Recommendation Systems"@(NSFC) Youth Fund Symposium.
- *2025.10*: Tutorial on "Building Robust and Interpretable ECG Models for Early Myocardial Infarction Detection" @宁波大学第一附属医院.
- *2025.10*: Serve as an Session Chair for IJCAI 2025 for Recommendation Session.
- *2025.09*: 🎉🎉 One paper about RAG-LLM Unlearning was accepted by ICDM 2025, congratulations to Haichao.
- *2025.04* 🎉🎉 One paper about Efficient PINNs was accepted by IJCAI 2025, congratulations to Yichen.
- *2025.07* 🏆 Become a member of the Edge Computing Committee of the Association for Automation.
- *2025.03*: 🏆 Embodied AI Through Cloud-Fog Computing: A Framework for Everywhere Intelligence. (IEEE 33rd International Symposium on Industrial Electronics [Best Paper Award]).
- *2025.02* 🎉🎉 One paper about LLM-Based Sequential Recommendation was accepted by APWeb'25, congratulations to Chenke.

1. Research Background: The "Paradox" of Personalized Recommendation

1.1 The Paradox: The Better It "Knows" You, The More It "Traps" You

- **Current Status**: Intelligent recommendation has become the primary portal for information retrieval, pursuing extreme personalization.
- **Problem**: Algorithmic over-catering leads to "narrowing" and "polarization" of information access.
- **Concept**: **"Information Cocoon"** — Users are unconsciously surrounded by homogenized information.

> **Policy Guidance**: CAC (Cyberspace Administration of China) emphasizes "Breaking the Cocoon" to resist information isolation.
>
> **Academic Frontier**: Academician Zhang Bo emphasizes that "Safety and Controllability" are the core of Third-Generation AI.


Comparison

1.2 The Essence: Why is "Breaking the Cocoon" So Difficult?

- Core Challenge: The "Three–No" Dilemma of Current Solutions

Dilemma	Manifestation	Key Terms
"Unmeasurable"	Cocoon severity is hard to measure in real–time/online. Both users and platforms remain "unconscious."	Theoretical Modeling Gap
"Opaque"	Recommendation mechanisms are "black boxes"; causal paths are unknown. Users cannot make targeted adjustments.	Lack of Explainability
"Inaccurate"	Simply increasing diversity hurts user experience. Conflicts arise between "new preferences" and "old representations."	Representation Conflict

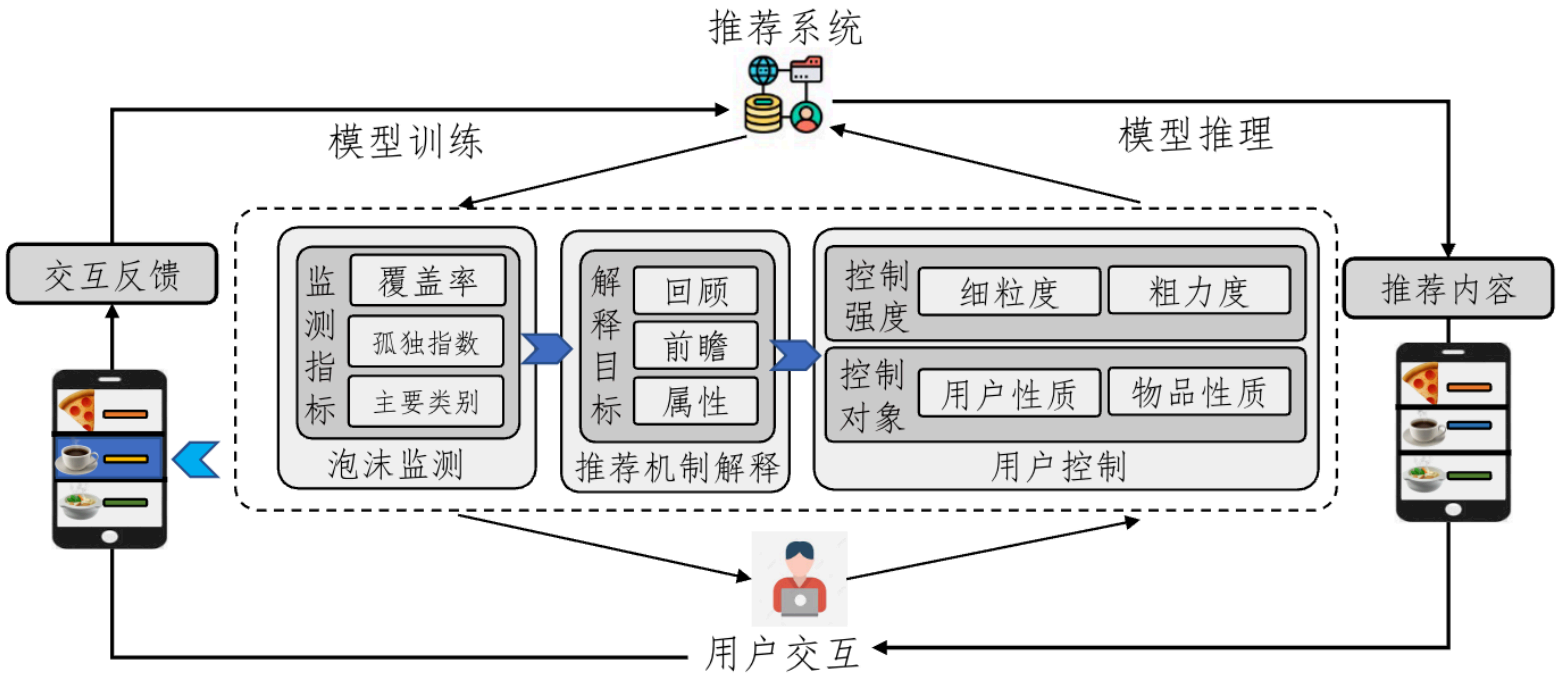
2. Core Idea: Paradigm Shift from "Passive Acceptance" to "Active Control"

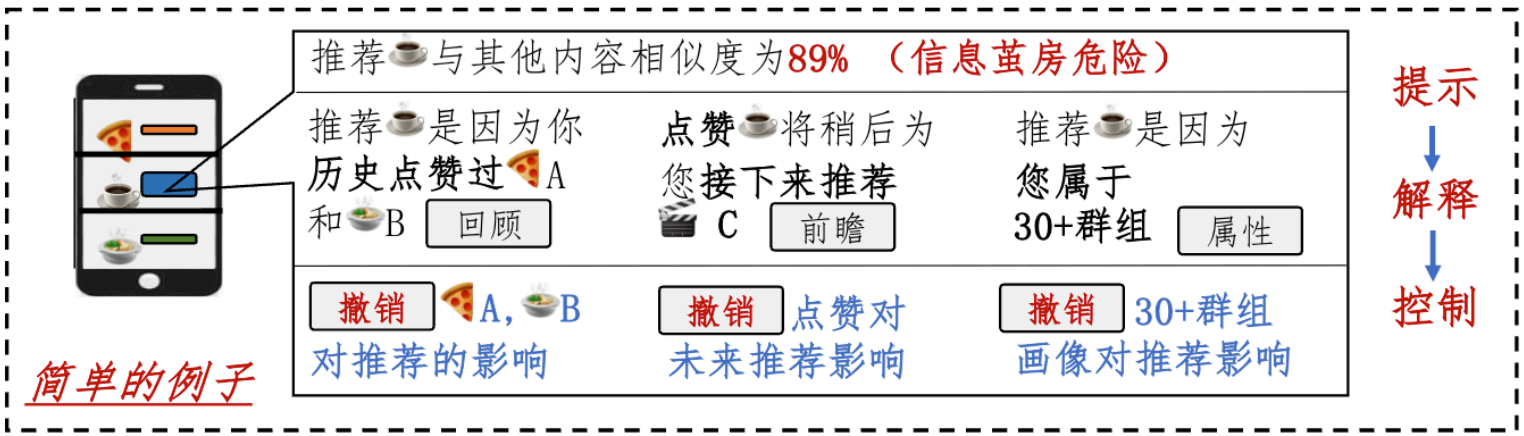
2.1 Our Solution: Empowering Users with "Counterfactual" Control

Three Core Pillars

- **Quantifiable (Measure):**
  - Real-time perception of cocoon severity.
- **Explainable (Understand):**
  - Insights into the causal factors of cocoon formation.
- **Controllable (Act):**
  - Precise intervention in recommendation results.

1. **Perceive**: Model multi-dimensional indicators to quantify and warn about information cocoons.
2. **Understand**: Use causal counterfactual reasoning to generate understandable "cocoon-breaking" path explanations for users.
3. **Act**: Implement precise user control over recommended content based on causal intervention.


推荐系统
模型训练
模型推理
交互反馈
监测指标
覆盖率
孤独指数
主要类别
泡沫监测
解释目标
回顾
前瞻
属性
推荐机制解释
控制强度
细粒度
粗力度
控制对象
用户性质
物品性质
用户控制
推荐内容
用户交互


推荐☕与其他内容相似度为89% （信息茧房危险）
推荐☕是因为你历史点赞过🍕A和☕B
回顾
点赞☕将稍后为您接下来推荐🎬C
前瞻
推荐☕是因为您属于30+群组
属性
撤销 🍕A, ☕B 对推荐的影响
撤销 点赞对未来推荐影响
撤销 30+群组画像对推荐影响
提示
解释
控制
简单的例子

3. Research Plan: Technical Framework

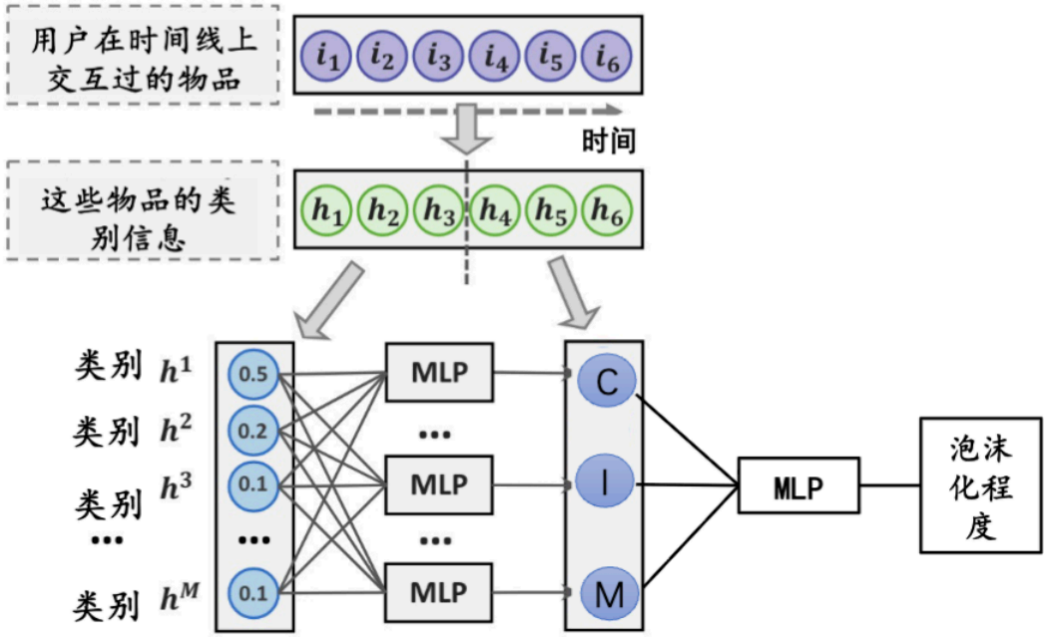
3.1 Tech I: Building a "Dashboard" for Information Cocoons

- **Multi-dimensional Metrics:**

  i. **Coverage**: Measures the **breadth** of recommended content.

  ii. **Isolation Index**: Measures the **segregation** between user groups (Sociological perspective).

  iii. **Measures**: Measures the **concentration** of content categories .



3.2 Tech II: Making the "Black Box" Talk

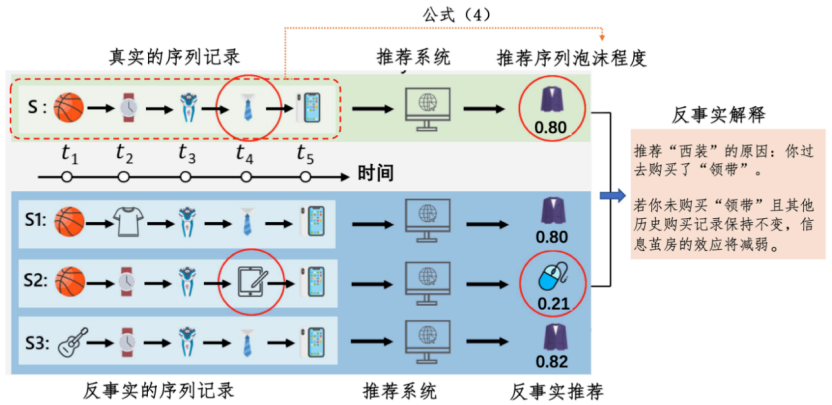
- **Core Idea: Answering "What-if" Questions**

> "If you hadn't clicked those videos in the past, would the system still recommend these to you now?"

- **Method: Causal Counterfactual Explanations**

i. **Retrospective**: "System content is homogenized because you liked [Item A]. To stop this, please revoke that action."

ii. **Prospective**: "Due to your current click, homogenization is rising. To prevent this, please undo the action."



3.3 Key Tech III: Achieving "Precise Control"

- **Core Challenges**:
  - Feature modifications may lead to "imprecise control" due to **confounders**.
  - Solving the **"New vs. Old Representation Conflict"** introduced by user control.
- **Methods**:

  i. **Causal Intervention**: Adjust predictions by subtracting the influence of "old preferences."

  ii. **Parameter-Efficient Fine-Tuning (PEFT)**: Control via Prompt Tuning while freezing the backbone model.

  iii. **LLM-based Control**: Utilizing "Natural Language" via Prompt Engineering or Knowledge Editing.

3.3 Key Tech III: "Precise Control" (Implementation Paths)

Path 1: Causal Intervention	Path 2: LLM Prompt + RAG	Path 3: Small Model Prompt Tuning
"Structured" Control	**"Zero-shot" Control**	**"Domain-Adaptive" Control**
Idea: Decouple confounding for precise intervention. Treat control as a `do`-operation on the causal graph to cut unwanted associations.	**Idea: Understand open intent without training.** Use powerful NLU of fixed LLMs to map natural language to recommendation strategies.	**Idea: Efficient tuning for precise adaptation.** Freeze the LLM and train "soft prompts" to adapt the model to control tasks.
Implementation: Calculate recommendation scores after removing specific feature influences via counterfactual inference. `Final Score = Adjusted Pred – α * (Original – Counterfactual)`	**Implementation:** Design sophisticated text prompts to guide LLMs to generate control signals or re-rank item lists.	**Implementation:** Use user instructions and history as input; output optimal control vectors via fine-tuned Soft Prompts.

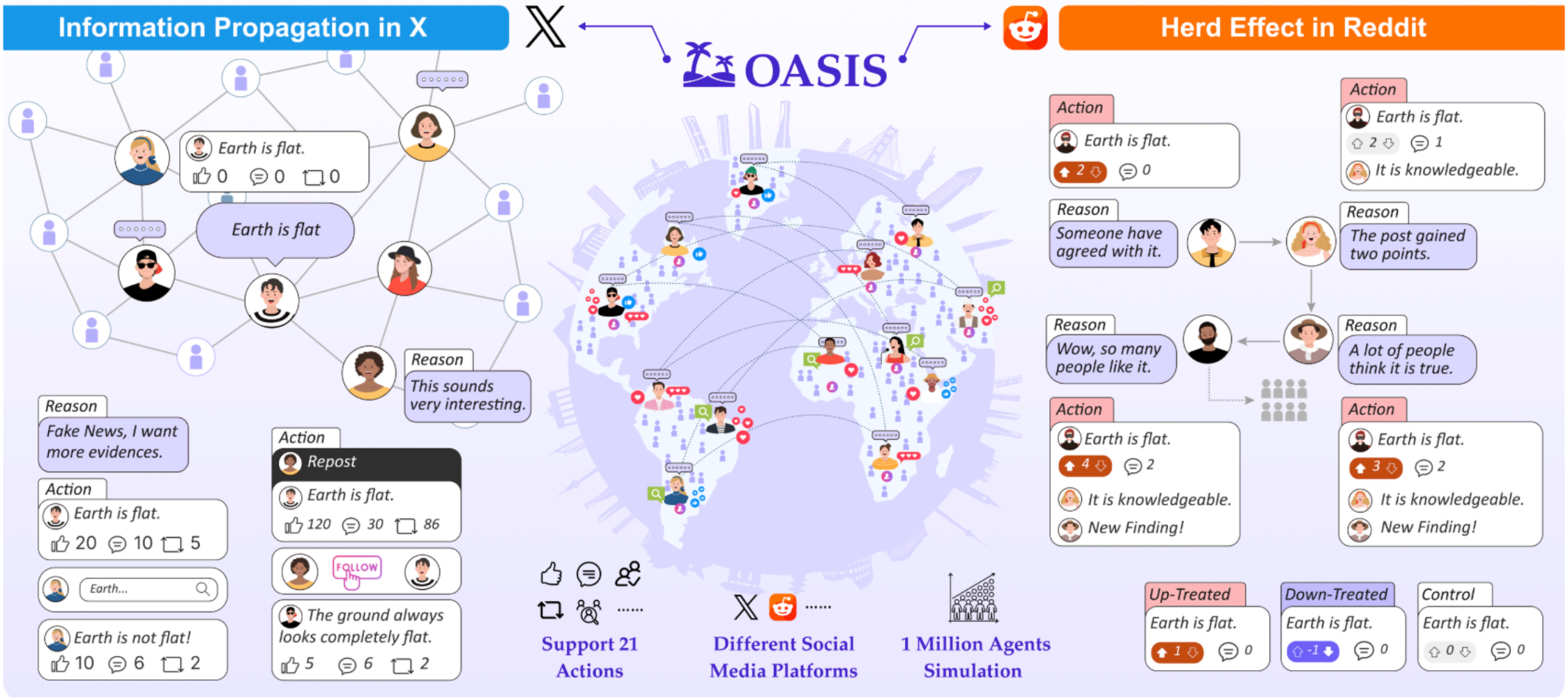
4. Recent Results: Project Progress

ReLAX: Training an AI Agent to Explain "Black-Box" Models

Core Idea: Train an agent to "clear the game" (change model predictions) using the minimum number of "moves" (modifying features), generating efficient, concise, and model-agnostic counterfactual explanations.

> ReLAX models the search for counterfactual explanations as a **Sequential Decision Task**:

- **Agent**: A Deep Reinforcement Learning agent.
- **Goal**: Change the prediction of an input sample (e.g., "Not Recommended → Recommended").
- **Action**:
  1. Feature selection (Discrete action)
  2. Modification magnitude (Continuous action)
- **Reward**:
  * Prediction Flip → High Reward ✓
  * Feature Modification → Penalty Mechanism ✗



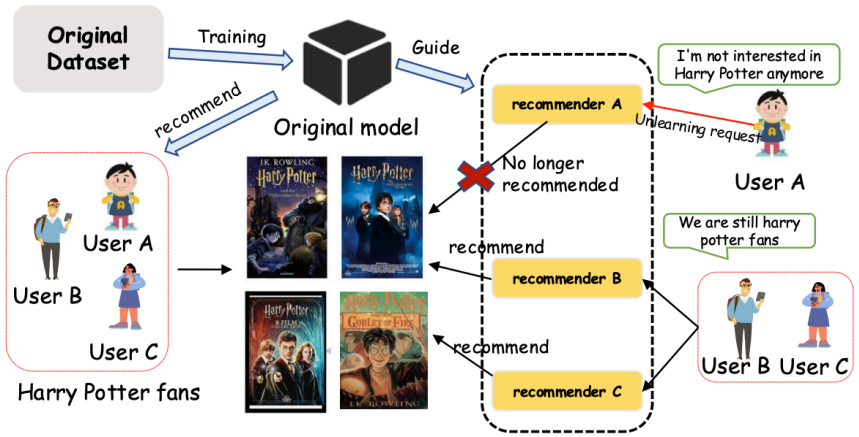
OASIS is a scalable, open-source social media simulator that incorporates large language model agents to realistically mimic the behavior of up to one million users on platforms like Twitter and Reddit [1].

[1] https://github.com/camel-ai/oasis

RAG-based Recommendation Unlearning Framework (Prompt+RAG)

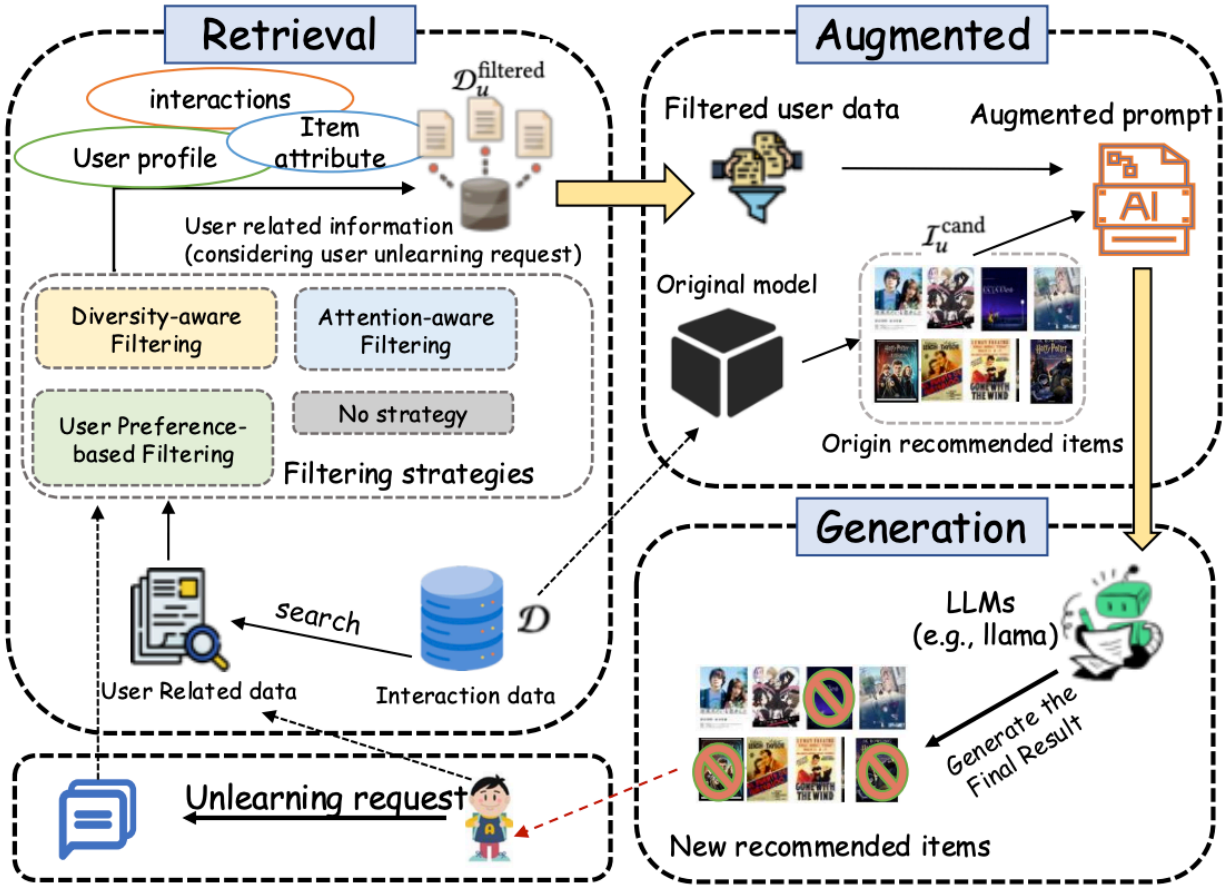
Core Idea: Addressing how to precisely erase the influence of specific user data without disturbing the overall system, avoiding the chain reactions caused by traditional parameter updates.

> The Unlearning Dilemma:
>
> - Traditional methods cause **global parameter perturbations**.
> - Parameter updates trigger **performance degradation for non-target users**.
> - Balancing dynamic preference capture with **bias propagation control** is difficult.



> Traditional methods potentially degrade recommendations for others.

> In contrast, our method leverages RAG with LLMs to perform efficient and precise user–level unlearning without affecting unrelated users.



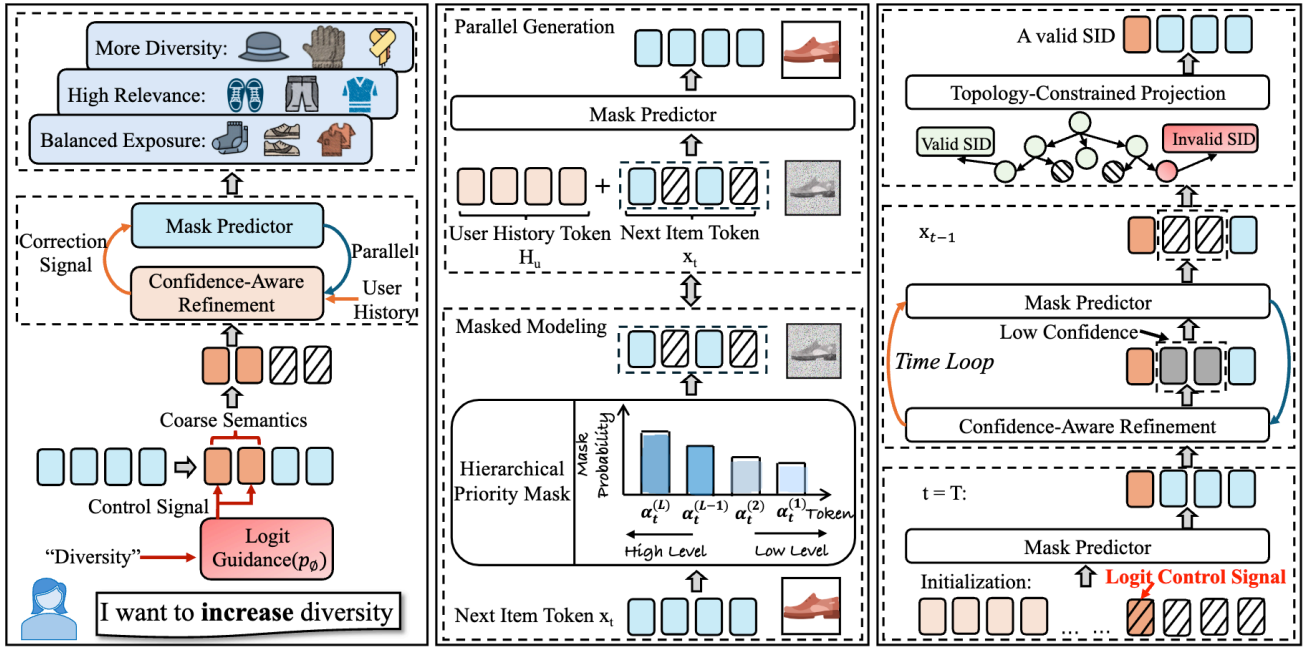
Haichao, Z., et al. "Customized Retrieval–Augmented Generation with LLM for Debiasing Recommendation Unlearning." The 25th International Conference on Data Mining (ICDM), 2025.

Semantic Decoding for LLM-Based Recommendation

Status Quo: LLM-based recommendation systems still use decoding strategies designed for NLP (e.g., Greedy Search, Beam Search) when generating results.

- **Contradiction**:
  1. **Different Goals**: NLP generates fluent **text sequences**, while RecSys only needs an accurate **Item ID**.
  2. **Semantic Redundancy**: Users may have equal interest in multiple **functionally similar** items (e.g., Bluetooth headphones of different brands). Traditional decoding treats these as independent options, diluting the model's "confidence."

❌ **Simply put: Using "essay writing" methods for "multiple-choice" questions is inefficient and misses the point.**



> ReGen reformulates generation as a ``mask–filling'' task, predicting all tokens simultaneously rather than one by one –-> "global view".
> This global view allows us to apply control signals directly to tokens with high–level semantics without retraining the model.
> To ensure the generated IDs map to real items, we introduce a topology–constrained tree, which iteratively corrects low–confidence predictions while strictly enforcing valid codebook paths.

5. Summary & Outlook

Expected Outcomes & Innovative Contributions

- **❝ Theoretical Innovation**
  Establish the first **quantitative modeling and prediction theory** for information cocoons.
- **❝ Technical Breakthrough**
  Develop a set of **user–controllable, real–time, natural language–supported** personalized recommendation algorithms.
- **❝ Social Value**
  Promote a healthier, more diverse, and inclusive information ecosystem.
  Provide technical support for **Controllable, Trustworthy, and Responsible Third–Gen AI**.
- **❝ Future Work**
  Smarter Interaction: Explore continuous user control based on multi–turn dialogue.

Thank You! Q&A

> Email: jia.wang02@xjtlu.edu.cn

> Open to academic collaborations