

OneReason Technical Report

OneRec Team

Generative recommendation models in the OneRec family have been widely deployed in many real-world services, such as short-video, live-streaming, advertising, and e-commerce. However, these generative models can only benefit from the scaling advantage, while their reasoning ability is hard to activate, since we cannot construct meaningful Chain-of-Thought (CoT) sequences consisting of itemic tokens only. Inspired by the success of the reasoning-style “think before answer” paradigm in the LLM field, we conduct preliminary studies (i.e., OneRec-Think, OpenOneRec) to explore reasoning capability in generative recommendation. Nevertheless, we notice an unexpected phenomenon: the thinking mode does not show advantages over the non-thinking mode. Drawing insights from recent findings on CoT robustness in multi-modal language models, we argue that effective reasoning in recommendation rests on two factors: *perception*, the ability to ground itemic tokens in their underlying language semantics, and *cognition*, the ability to reorganize a user’s behavior sequence into coherent latent interest points. We therefore propose OneReason, which includes: (1) strong itemic token perception in pre-training, (2) a three-level cognition-enhanced CoT format for recommendation tasks in SFT, and (3) a specialize-then-unify training recipe in RL to enhance the thinking ability. Equipped with the above techniques, OneReason successfully achieves a thinking mode that outperforms the non-thinking mode across multiple real-world business benchmarks at Kuaishou. Moreover, we observe an interesting phenomenon: replacing unCoT recommendation data with CoT-supervised data can improve non-thinking inference in several domains. This suggests that some benefits of CoT supervision may transfer to direct decoding, although the current evidence is behavioral and does not by itself identify whether the effect comes from compression, reasoning, or their interaction. Finally, we systematically present a series of reasoning benchmarks and design insights, and we will open-source the OneReason-8B and OneReason-0.8B models to facilitate research on generative recommendation models.

arXiv:2606.06260v1 [cs.LG] 4 Jun 2026

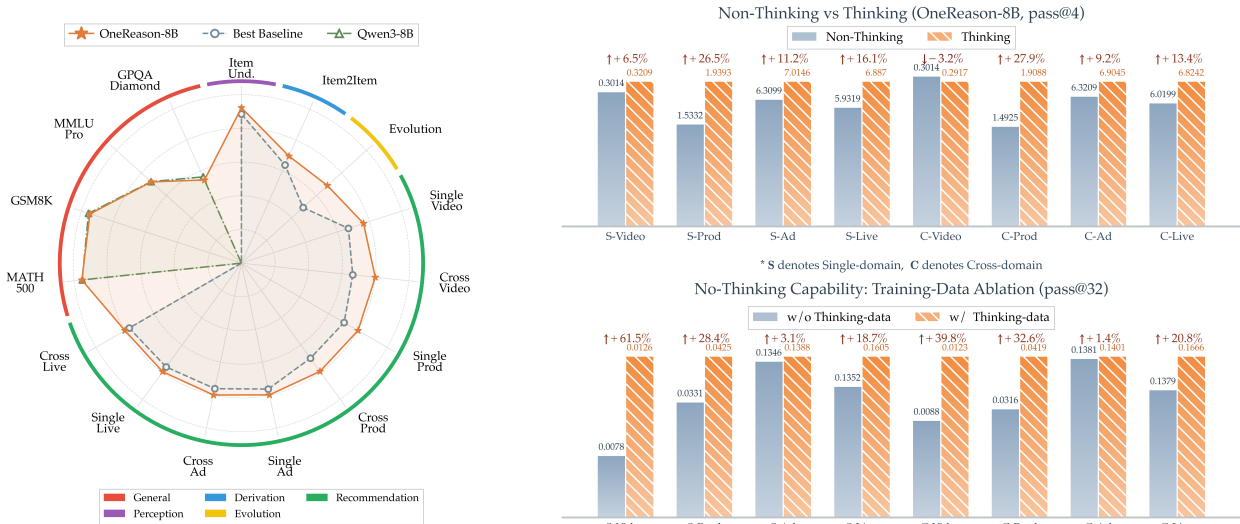


Figure 1 | Performance overview: (1) compared with Qwen3-8B, OneReason-8B maintains its general capabilities; (2) OneReason-8B exhibits SOTA performance across real-world recommendation benchmarks; (3) OneReason-8B showing promising “thinking” advantage to recommendation tasks.

Contents

1	Introduction	4
2	Reasoning Design Philosophy	7
3	Benchmark Design for Reasoning Recommendation	7
3.1	Task Taxonomy	8
3.2	Evaluation Protocols	10
3.3	General-Intelligence Sanity Check	11
4	Pre-Training Pipeline	11
4.1	Itemic Tokenizer	11
4.2	Pre-Training Data	11
4.3	Ablation Study on Recommendation Corpora	17
4.4	Training Recipe	18
5	SFT Pipeline	20
5.1	R0: Perception	20
5.2	R1: Derivation	21
5.3	R2: Evolution	22
5.4	R3: Recommendation	23
5.5	Itemic Instruction Data	29
5.6	General-Domain Data	30
6	RL Pipeline	31
6.1	Recommendation-oriented Reinforcement Learning	31
6.2	Rejection Sampling Fine-tuning	37
6.3	Multi-Teacher On-Policy Distillation	38
6.4	Comparison and Discussions	42
7	CoT Analysis Indicators	44
8	Experiments	50
8.1	Performance Comparison	50
8.2	Non-Thinking Gains from Thinking Supervision	52

9	Deployment	55
9.1	Industrial Scenario Adaptation	55
9.2	Industrial Scenario Online Deployment	56
9.3	Online Experiment	57
10	Related Works	57
11	Conclusion & Future Work	59
12	Author List	66
A	Deployment Details	67
A.1	Industrial Scenario Adaptation	67
A.2	The Overall Methodology of OneReason for OneRec	67
A.3	Experimental Analysis	68
A.4	Discussions in Industrial Deployment	69
B	Benchmark Details	70
B.1	Task Construction	70
B.2	Recommendation Data Statistics and Analysis	70
B.3	Comparison with Existing Benchmarks	71
B.4	Evaluation Metrics	71
B.5	Task Formats	74
C	Pre-Training Details	80
C.1	Positive–Negative Similarity Margin Study	80
C.2	Data Mixture for Pre-Training	81
C.3	Four-Granularity Pre-Training Samples	83
D	SFT Details	85
D.1	SFT Data and Itemic-Token Perception	85
D.2	Cognitive-Reasoning Data Construction	88
D.3	Recommendation CoT Construction	99
D.4	Itemic Instruction Details	106

1. Introduction

In recent years, Transformer-based large language models (LLMs) have pushed the frontier of machine intelligence to unprecedented levels. Driven by massive compute scale and advanced training methodologies, LLMs have come to exhibit two exciting properties that guide us to build more powerful models:

- **Scaling property in pre-training:** When model size, training data, and compute resources are scaled up, the convergence loss follows a regular, extrapolable decay trend (Kaplan et al., 2020; Hoffmann et al., 2022). As scaling proceeds, the evaluation scores across multiple benchmarks improve simultaneously, demonstrating the models’ remarkable semantic understanding and generalization capability.
- **Reasoning property in post-training:** Following pre-training, the post-training stage further unlocks and refines the model’s advanced capabilities via high-quality Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) (Ouyang et al., 2022; Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025). Driven by these post-training techniques, LLMs learn to follow CoT-style procedures (Kojima et al., 2023; Wei et al., 2023), forming a “think-before-answer” reasoning paradigm. This paradigm significantly enhances their capabilities to solve complex, in-depth problems, as exemplified by the OpenAI o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) series.

Inspired by the **Scaling** and **Reasoning** properties of LLMs, we are deeply curious about whether these two properties can be further leveraged to push the capability of industrial recommender systems. Driven by this motivation, we have started a series of exploration efforts. For the Scaling property, our OneRec V1/V2 series of generative models (Zhou et al., 2025b; c) have validated **the scaling property in industrial recommendation scenarios**, contributing significant business gains at Kuaishou. For the Reasoning property, we acknowledge that the recommendation-data-driven OneRec model cannot unlock the reasoning ability on its own. Since the OneRec model is trained exclusively on pure itemic sequential data, it has only learned from flat transition patterns without any underlying logical thinking traces. To elicit the reasoning capabilities, our recent models, OneRec-Think (Liu et al., 2025) and OpenOneRec (Zhou et al., 2026), have explored introducing interleaved itemic-text recommendation data and general-domain reasoning data. In these two models, we have successfully observed that **the “think-before-answer” pattern can be generalized to recommendation tasks**, allowing us to inspect the underlying logic of the model and to understand the user’s interests as well as the rationale behind the next-item prediction. However, we also notice an unexpected phenomenon: **the thinking mode does not show significant advantages on recommendation benchmarks compared with the non-thinking mode**.

To explore the root cause of this unexpected phenomenon, we turn to the multi-modal LLM (MLLM) literature, where a strikingly similar reasoning-mode brittleness has been observed. (Sun et al., 2026b) reveals that when textual and visual modalities are insufficiently aligned, models tend to mechanically “read” surface-level visual text rather than genuinely reason about the underlying visual semantics, providing a compelling indication that deep cross-modal alignment is a prerequisite for authentic thinking. Extending this insight, (Zhou et al., 2025a) systematically surveys the perception-to-cognition pathway in MLLMs and concludes that advanced interactive reasoning cannot emerge without first establishing robust perceptual alignment between modalities. Building on this alignment foundation, Jiang et al. (2025b) further demonstrates that, even with aligned modalities, reasoning robustness remains fragile unless the CoT traces themselves are carefully structured, specifically by adopting logically coherent, coarse-to-fine reasoning formats that prevent hallucinations and over-thinking.

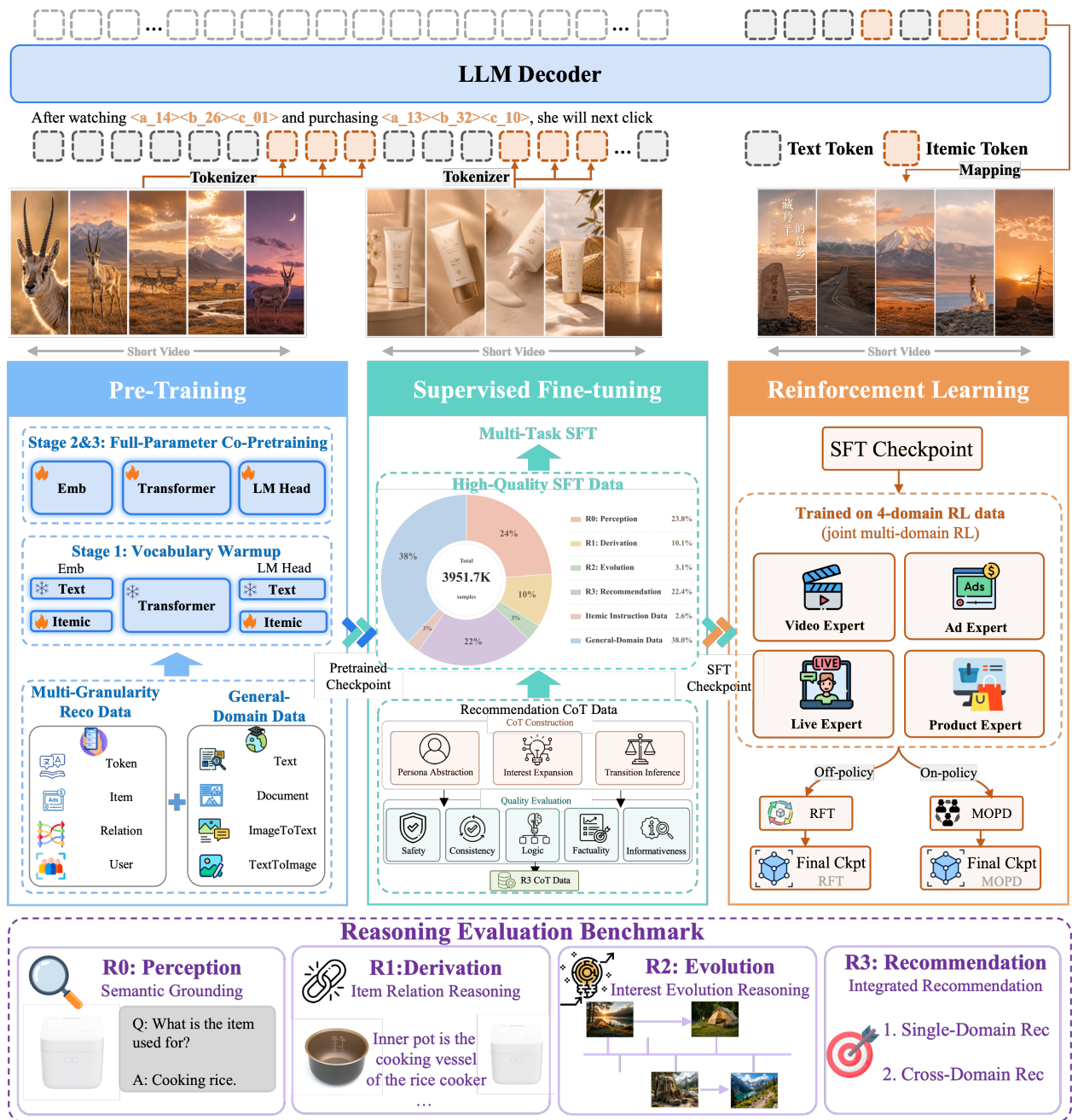
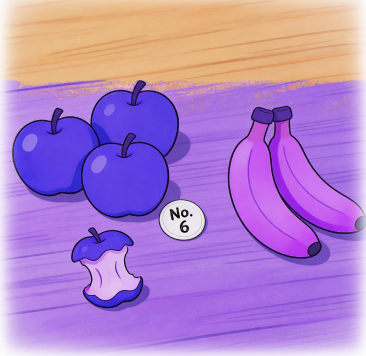


Figure 2 | The pre-training, SFT, RL, and benchmarking pipeline of OneReason.

Taken together, these works converge on two essential and complementary pillars for unlocking genuine reasoning capability: (i) **modality alignment in perception**, which grounds perception and prevents the model from operating on hollow surface signals; and (ii) **CoT quality in cognition**, which provides coherent coarse-to-fine reasoning trajectories and enables a faithful *think-before-answer* process once alignment is in place.

For generative recommendation, as illustrated in Figure 3, these insights carry a clear implication: enabling a model to truly *think* about user interests demands both (1) **deep alignment in perception** between itemic modality and natural-language modality, so that itemic tokens become referable, composable semantic units rather than opaque identifiers; and (2) **recommendation-specific and**

Question: How many complete fruits are there?



(a) Alignment Failure

✗ Answer: 6

"I see 4 apples and 2 bananas. So answer is 6."

Cause: Misperceived objects; weak visual grounding

(b) CoT Quality Failure

✗ Answer: 6

"3 apples + 2 bananas = 5. But apple core & No. 6 suggest 6 originally. Answer: 6."

Cause: Perception OK but faulty reasoning chain

(c) Robust Reasoning

✓ Answer: 5

"3 apples + 2 bananas = 5. Core & tag are irrelevant."

Key: Deep alignment + logical CoT

Figure 3 | Three failure/success modes of MLLM chain-of-thought reasoning given the same visual input. (a) Alignment failure: the model miscounts due to weak visual grounding. (b) CoT quality failure: correct perception but faulty reasoning drift. (c) Robust reasoning: deep alignment and logical CoT lead to the correct answer.

logically fluent CoT structures in cognition that are designed to scaffold deliberate reasoning over user histories and item properties on top of this alignment. These two dimensions serve as the cornerstones for building a powerful reasoning model, where we explicitly strengthen text-item alignment and devise tailored coarse-to-fine reasoning formats to make the model genuinely capable of recommendation-oriented thinking.

In light of the above, we propose OneReason, a recommendation foundation model with genuine reasoning capability. Specifically, for better **item perception ability**, we collect complex and diverse coarse-to-fine alignment pre-training corpora to align the newly inserted itemic tokens with text tokens in the same semantic space. For better **reasoning cognition ability**, we elaborate a standard thinking structure to ensure CoT quality and build the SFT corpora. Moreover, building on the model after pre-training and SFT, we find that the thinking mode still underperforms the non-thinking mode under multi-domain mixed RL, yet consistently surpasses it when RL is restricted to a single domain. This contrast motivates us to adopt a “specialize-then-unify” RL strategy - we first conduct domain-focused RL to fully unlock thinking-mode benefits within each domain, and then apply rejection sampling fine-tuning (Yuan et al., 2023) or Multi-Teacher On-Policy Distillation (Xiao et al., 2026) for cross-domain balancing and refinement. The overall training process of OneReason is visualized in Figure 2. Meanwhile, we also make an interesting observation: under an equal number of training tokens, adding recommendation CoT supervision can improve non-thinking-mode performance compared with training on pure unCoT data in several domains (see the bottom right panel of Figure 1). We treat this as behavioral evidence that some supervision signals from CoT traces may transfer to direct decoding, rather than as proof that CoT is universally necessary or that the mechanism has been identified.

To evaluate the reasoning ability of our model in recommendation scenarios, we introduce OneReason-Bench, a curated collection of reasoning-oriented recommendation benchmarks. Orga-

nized along multiple axes, this benchmark is designed to assess reasoning abilities in a comprehensive and structured manner, and has played a central role in guiding and validating our model development. As illustrated in Figure 1, our OneReason model achieves superior performance across this benchmark, demonstrating its robust reasoning capabilities. To facilitate research on generative recommendation models, we will open-source the OneReason-8B and OneReason-0.8B models and other fine-tuning-related materials to the public.

2. Reasoning Design Philosophy

Before describing the architecture and training recipes in detail, we first outline the design principle of recommendation reasoning: what is a good CoT trace in recommendation? In classical LLM reasoning tasks involving math, code, or symbolic logic, it is usually assumed that there exist intermediate steps that maintain task constraints and apply relevant world knowledge to logically answer the question step by step. Unlike classical reasoning tasks with a single correct answer, recommendation admits multiple plausible items at once, and the underlying user intent is never directly observable; it must be inferred from a long sequence of historical itemic tokens (where each complete sequence represents some item). Therefore, recommendation reasoning is abductive rather than deductive: it requires hypothesizing latent interest points, modeling how they evolve over time, and using them to justify why specific candidates fit the current context. In other words, a recommendation CoT should select relevant behaviors as hypothesized interest points, compress them into an interpretable preference, model the temporal transitions between these interests, and finally associate the inferred preference state with the recommended item. This distinction motivates our training and evaluation design. We decompose recommendation reasoning into four diagnostic capabilities along two dimensions: (1) **Perception (R0)**, which grounds itemic tokens in their semantic content; and (2) **Cognition (R1–R3)**, which builds progressively deeper reasoning on top of solid perception.

- **R0: Perception.** Interpreting itemic patterns via their explicit semantic content is the foundational ability; without it, user behavior is uninterpretable.
- **R1: Derivation.** Reasoning from individual item semantics to item-to-item relations via common-sense or knowledge-based associations; without it, latent interests cannot be extracted from noisy interaction histories.
- **R2: Evolution.** Reasoning over items of the same latent interest as a temporal process; without it, long-term, short-term, and periodic preferences cannot be modeled.
- **R3: Recommendation.** Reasoning coherently to produce high-quality decisions across all served domains (e.g., Short-Video, Live-Streaming).

The immediately following section operationalizes this hierarchy as an evaluation instrument, so that the trajectory from pre-training (Section 4) to post-training (Sections 5 and 6) to deployment (Section 9) can be measured at every step against the same diagnostic axes.

3. Benchmark Design for Reasoning Recommendation

Placing the benchmark before the model description is deliberate: the four-granularity pre-training (Section 4), the compression-then-reasoning SFT (Section 5), and the specialize-then-unify RL (Section 6) are each motivated by specific gaps revealed at the R0–R3 layers of OneReason-Bench. The benchmark therefore functions less as a leaderboard and more as a measurement protocol that justifies, monitors, and ultimately validates every downstream design choice.

While RecIF-Bench in OpenOneRec (Zhou et al., 2026) broadens the evaluation of recommendation

foundation models, its reasoning evaluation is still coarse-grained and not sufficiently diagnostic. To address this, we develop OneReason-Bench, extending RecIF-Bench toward multi-level reasoning.

3.1. Task Taxonomy

OneReason-Bench organizes reasoning-oriented benchmark tasks into a progressive hierarchy: Perception (R0), Derivation (R1), Evolution (R2), and Recommendation (R3). These reasoning layers progress from semantic grounding to relational and temporal reasoning, and finally to integrated recommendation tasks. Table 1 summarizes the taxonomy, and examples are shown in Figure 4.

Unified Task Formulation. All tasks are formulated as sequence generation $Y = \mathcal{F}(X)$, where X combines a task instruction I with a context C such as itemic patterns, user portraits, or interaction histories. The target Y can be an itemic pattern, an answer option, a natural-language response, or a structured evolution chain, allowing R0–R3 tasks to share a common generative evaluation protocol.

Table 1 | Task taxonomy of OneReason-Bench organized across four reasoning layers.

Layer	Task	Input (X)	Target (Y)	Metric
R0: Perception	Item Understanding	Item i	Item description	LLM-as-a-Judge
	Itemic Pattern Grounding	Item description	Item i	Pass@K, Recall@K
	Item QA	Item i + Answer options O^a	Correct option o^a	Accuracy
R1: Derivation	Item2Item	Source item i + Item options O^{item}	Correct option o^{item}	Accuracy
R2: Evolution	Evolution Action Selection	History \mathcal{H} + Topic t	Related actions \mathcal{A}	F1
	Evolution Topic Gen.	History \mathcal{H} + Topic t	Evolution chain \mathcal{E}_t	Action–Logic Score
	Evolution Direct Gen.	History \mathcal{H}	Evolution chain(s) $\{\mathcal{E}\}$	Multi-Chain Action–Logic Score
R3: Recommendation	Single-Domain Rec. (Video / Product / Ad / Live)	Portrait \mathcal{P} + History \mathcal{H}^d	Next items $\{i\}$	Pass@K, Recall@K
	Cross-Domain Rec. (Cross-Video / Cross-Product / Cross-Ad / Cross-Live)	Portrait \mathcal{P} + History \mathcal{H}	Next items $\{i\}$	Pass@K, Recall@K

3.1.1. R0: Perception

Perception grounds itemic patterns in explicit semantic content, providing the semantic basis for higher-level reasoning and recommendation (Zheng et al., 2024a; Zhou et al., 2026). We instantiate this layer with three task types:

- **Item Understanding:** given an item i , the model generates its natural-language description, decoding item semantics from itemic patterns.
- **Itemic Pattern Grounding:** given a natural-language description, the model grounds it to the corresponding item.
- **Item QA:** given an item i and content-grounded multiple-choice options O^a , the model selects the correct option, assessing whether it can use item semantics to answer targeted questions.

3.1.2. R1: Derivation

Building upon Perception, Derivation evaluates whether the model can reason from individual item semantics to item-to-item relationships under common-sense or knowledge-based associations.

- **Item2Item:** given a source item i and a set of candidate items O^{item} , the model selects the item that best matches the intended item-to-item relation.

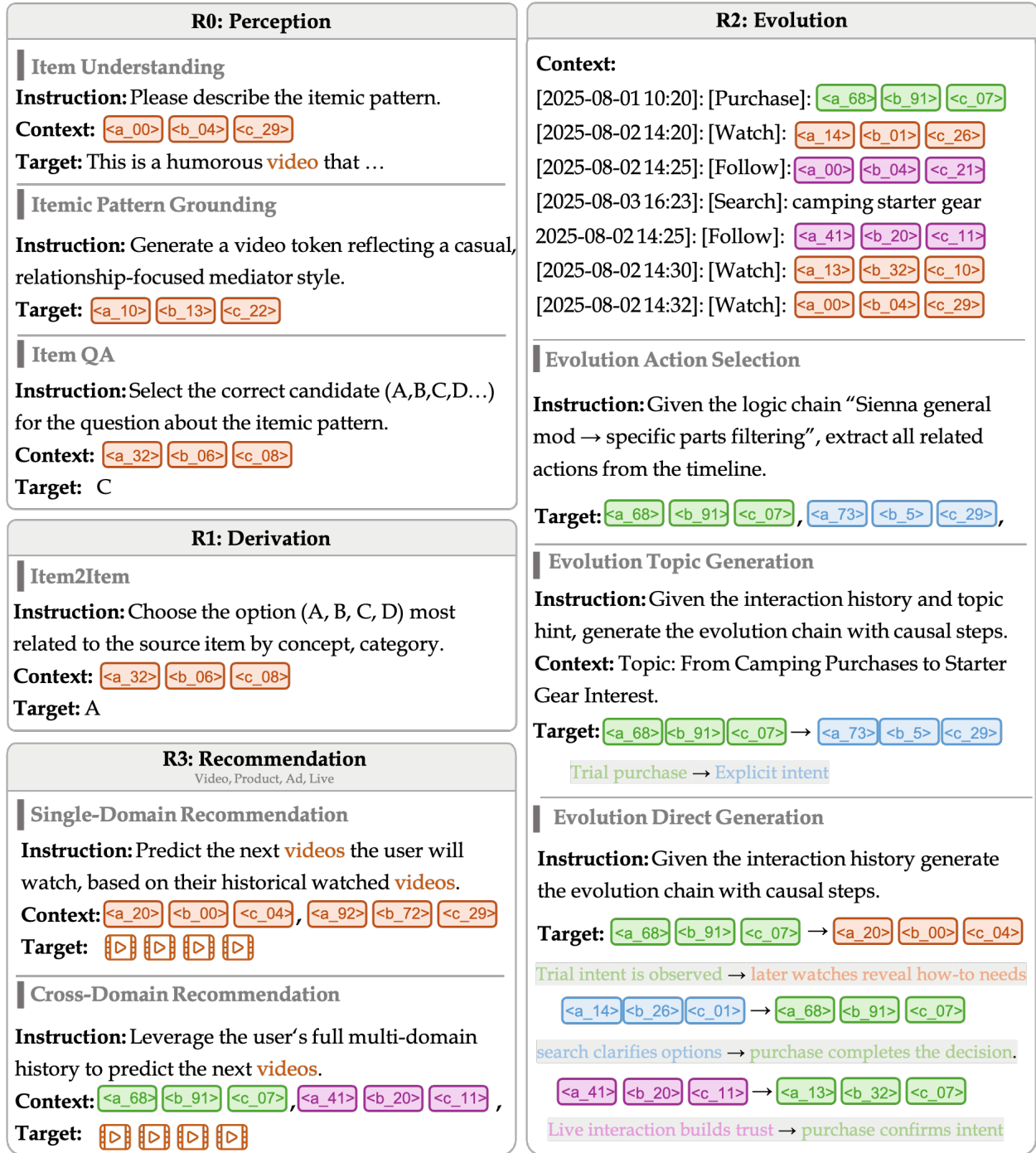


Figure 4 | Illustration of the task taxonomy. Each task is specified with its instruction, context, and target format.

3.1.3. R2: Evolution

Evolution evaluates whether the model can reason over user interests as a temporal process rather than a static preference profile. We design three task variants with increasing autonomy, from extracting topic-relevant actions to generating topic-guided or fully autonomous evolution chains:

- **Evolution Action Selection:** given a history \mathcal{H} and a specified interest topic t , the model

identifies the historical actions \mathcal{A} logically connected to that topic.

- **Evolution Topic Generation:** given a history \mathcal{H} and a specified interest topic t , the model generates the evolution chain \mathcal{E}_t describing how this interest developed.
- **Evolution Direct Generation:** given only the history \mathcal{H} , the model autonomously generates valid interest-evolution chains $\{\mathcal{E}\}$ without topic guidance.

This layer evaluates multi-hop, temporally structured reasoning about user behavior, providing a bridge from item-level reasoning to context-aware recommendation.

3.1.4. R3: Recommendation

While the preceding tasks focus on specific R0–R2 capabilities, real-world recommendation scenarios often require models to flexibly combine perception, derivation, and evolution abilities. We therefore introduce integrated recommendation tasks to evaluate the model’s overall reasoning capability in practical recommendation settings. We assess this across two settings:

- **Single-Domain Recommendation:** given a user portrait \mathcal{P} and the user’s interaction history \mathcal{H}^d within a target domain (short video, advertising, e-commerce, or live-streaming¹), the model predicts the next items the user will engage with.
- **Cross-Domain Recommendation:** given a user portrait \mathcal{P} , the user’s multi-domain interaction history \mathcal{H} , and a specified target domain d , the model predicts the next items in that domain.

Detailed construction procedures, R3 data statistics, and a comparison with existing recommendation benchmarks are provided in Appendix B.1 and B.2.

3.2. Evaluation Protocols

We employ a set of task-specific metrics to cover recommendation accuracy, QA correctness, interest-evolution reasoning, and open-ended generation quality.

Text Generation Metrics. For the Item Understanding task, we reuse the LLM-as-a-Judge protocol introduced in OpenOneRec (Zhou et al., 2026) without modifying its metric definition.

Choice and Selection Metrics. For single-answer choice tasks such as Item QA and Item2Item, we report **Accuracy**. For multi-answer selection tasks, including Evolution Action Selection, we report **F1** computed over the predicted and ground-truth sets.

Evolution Generation Metrics. Evolution Topic Generation and Evolution Direct Generation are evaluated by the **Action–Logic Score** family. This metric family combines action alignment between the generated and ground-truth evolution chains with logic alignment between the corresponding reasoning statements. The detailed implementation of these metrics is provided in Appendix B.4.2.

Recommendation Metrics. For recommendation and itemic pattern grounding tasks, we report **Pass@K** and **Recall@K**. Pass@K indicates whether at least one ground-truth target appears in the top-K generated outputs, while Recall@K measures the fraction of ground-truth targets recovered

¹For brevity, we denote short video as Video, advertising as Ad, e-commerce as Product, and live-streaming as Live.

within the top- K list. The generated itemic patterns are first decoded into item IDs, and metrics are then computed at the item granularity.

3.3. General-Intelligence Sanity Check

Following OpenOneRec (Zhou et al., 2026), we also retain a general-intelligence sanity-check suite, covering general-purpose tasks such as MMLU-Pro (Wang et al., 2024), to ensure that recommendation specialization does not come at the cost of general reasoning or instruction-following ability.

With the diagnostic axes (R0–R3) and the general-intelligence guard rail now in place, we are ready to describe the training stack that OneReason-Bench is designed to evaluate. The next section opens with the perception foundation, itemic tokenization and four-granularity pre-training data, that R0 directly probes and on which every higher-order capability is built.

4. Pre-Training Pipeline

This section details our pre-training methodology. We begin by describing the itemic tokenizer in Section 4.1. Section 4.2 then introduces our core contribution: a four-granularity data system that hierarchically integrates recommendation corpora with general-domain data to achieve deep modality alignment. Section 4.3 systematically ablates each granularity to quantify its specific contribution to the model’s overall performance. Finally, Section 4.4 outlines the three-stage training recipe that operationalizes this data design, from vocabulary warm-up to full-scale training.

4.1. Itemic Tokenizer

To obtain compact yet semantically grounded item embeddings, we jointly train a multimodal encoder and a decoder LLM on item-understanding tasks. The encoder integrates a Vision Transformer (ViT) (Radford et al., 2021) and an LLM initialized from the open-source Qwen3-VL (Qwen3 et al., 2025), together with an audio encoder (Vyas et al., 2026). It distills rich content signals—including cover images, video frames, text descriptions, and audio—into a compact dense embedding. To optimize the item embedding in an end-to-end manner, we prepend the embedding as a soft prefix to a separate decoder LLM, followed by the item’s text descriptions.

We quantize these embeddings via RQ-KMeans (Luo et al., 2025). We use a three-layer codebook, with each layer containing 8192 codes. Each item is thus represented by an itemic pattern consisting of one domain-aware begin token and three sub-tokens. For instance, an itemic pattern is formatted as:

```
<|domain_begin|><a_5028><b_6733><c_2559>,
```

where domain is one of video (video), prod (product), ad (ad), or living (live) for our four recommendation scenarios, or sid for general-domain multimodal data. Unlike OpenOneRec (Zhou et al., 2026), we drop the trailing end token. This reduces the overall context length consumed by itemic patterns, leaving more capacity for the reasoning traces that are essential to our thinking paradigm.

4.2. Pre-Training Data

Prior generative recommendation works, including OneRec-Think and OpenOneRec, organize pre-training data by task type—item captioning, user behavior sequences, and persona-text interleaving—and mix them for joint training. While this establishes a basic mapping between itemic patterns

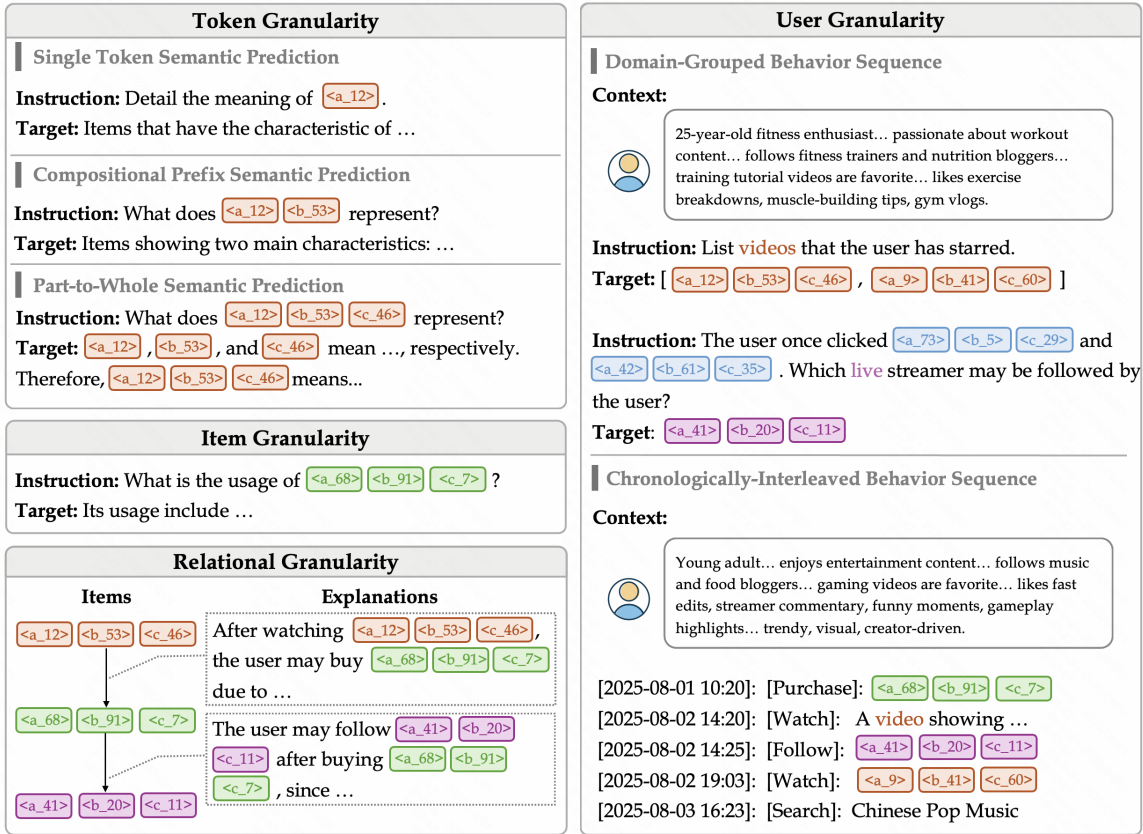


Figure 5 | Overview of our four-granularity pre-training recommendation corpora, ordered from fine-grained token-level semantics to coarse-grained user behavior modeling.

and natural language, it leaves three structural gaps. First, the semantic expression of items and users tends to be homogeneous, limiting the linguistic diversity the model encounters and weakening representational robustness. Second, the data organization stays at the task level without explicitly modeling the fine-grained semantic hierarchy inside itemic tokens or the relational logic across items. Third, user behavior modeling relies on a narrow conditioning paradigm in which a full user profile maps to a complete behavioral sequence, restricting the model’s ability to generalize across diverse recommendation contexts.

To address these gaps, we restructure the pre-training corpora into four progressive granularities, moving from microscopic token semantics to macroscopic user behavior. This hierarchical design systematically expands the alignment between the textual and itemic modalities along three axes: broader expression diversity; explicit structural reasoning from sub-token composition to inter-item relations; and richer conditioning patterns for user modeling. In the remainder of this section, we detail the construction of both the recommendation corpora and the general-domain corpora that jointly realize this vision.

4.2.1. Recommendation Corpora

To achieve a deep and seamless fusion between the textual and itemic modality, as shown in Figure 5, we structure our pre-training corpora into four progressive granularities:

- **Token Granularity** data focuses on individual or compositional sub-itemic-token semantics and achieves modality alignment at the finest granularity.

- **Item Granularity** data aligns itemic pattern with the textual modality. It pushes cross-modal fusion to the full item level and leverages multi-view tasks and diversified item sources to match complete itemic patterns with comprehensive textual descriptions.
- **Relational Granularity** data explains the connections between items via natural language. It summarizes the interest-transfer logic of associated items and enables collaborative signals to be encoded into text representations and model parameters.
- **User Granularity** data accomplishes thorough cross-modal fusion by enabling the model to understand and generate multi-modal content under unified, personalized user contexts.

To empirically validate the enhanced modality alignment afforded by our multi-granularity corpora, we randomly sample a set of items and extract embeddings for both their itemic patterns and textual captions. Each item is paired with its own caption to form a positive pair, while a caption randomly drawn from another item serves as the corresponding negative. Whereas prior work typically reports positive and negative pair similarities in isolation (Rusak et al., 2024; Lee et al., 2025), we additionally compute their difference, yielding a *positive-negative* similarity margin that more directly quantifies the cross-modal alignment. As shown in Figure 6, replacing the pre-training recommendation corpora of OpenOneRec with our proposed corpora consistently shifts the distribution of this margin to the right across all itemic domains, evidencing stronger cross-modal alignment. We defer the comprehensive methodology of this evaluation to Appendix C.

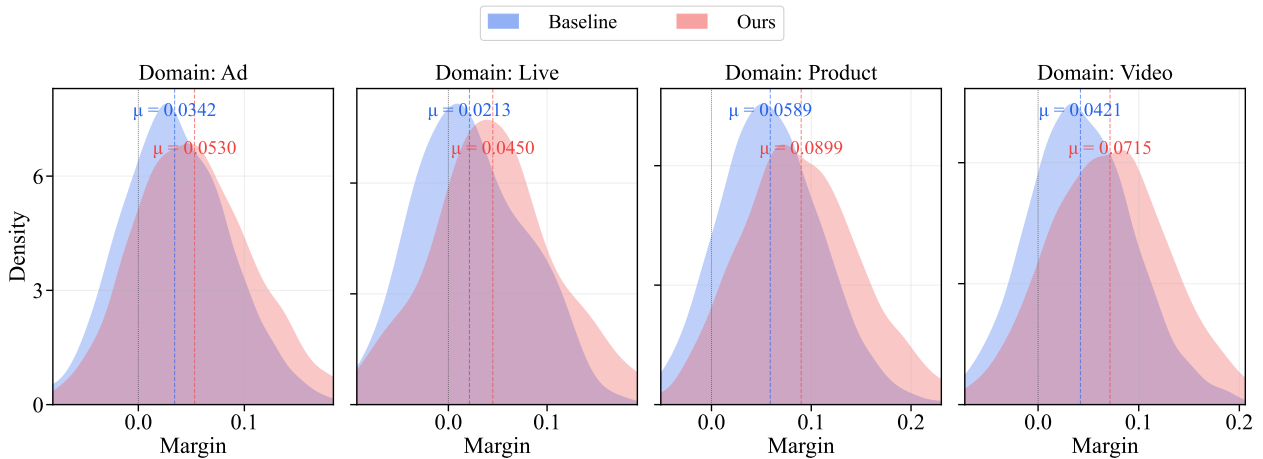


Figure 6 | Distribution of the *positive-negative* similarity margin across the four itemic domains. A rightward shift indicates stronger cross-modal alignment. Our corpora consistently outperform the reproduced OpenOneRec baseline.

In the remainder of this section, we describe the detailed construction methodology for the pre-training corpora at each of the four granularities. The detailed token ratios for each corpus are provided in Appendix C.2.

Token Granularity. Achieving precise alignment between language and itemic patterns requires modeling semantics across a continuous spectrum of granularities. Traditional methods (Zhou et al., 2026) align each itemic pattern with its holistic caption. At the opposite end, recent explorations (Feng et al., 2026) focus on single-token semantic prediction by pairing each individual sub-token with its semantic meaning. However, the crucial compositional process through which the semantics of these isolated sub-tokens hierarchically combine into the semantics of the full itemic pattern remains implicit and is not explicitly modeled.

To bridge this gap, our token-granularity data is designed to explicitly model the compositional process through which sub-token semantics are progressively integrated into complete itemic-pattern semantics. Building upon the basic premise of **single-token semantic alignment**, we further introduce two novel components that strengthen sub-token-level semantic understanding while explicitly modeling the hierarchical composition of semantics across granularities:

1. **Compositional Prefix Semantic Prediction.** We formulate tasks that require the model to predict the combined semantic meaning of a prefix sub-token pair $\langle a_xxxx \rangle \langle b_xxxx \rangle$. The data is constructed by leveraging an LLM to summarize the shared semantics of items that share the given prefix. This kind of data explicitly teaches the model how isolated semantics merge and interact at the intermediate sub-token-pair granularity. We further construct the reverse direction, **Prefix Itemic Token Grounding**, where the model must retrieve the correct prefix sub-token pair given a textual description of the combined semantics. This bidirectional formulation ensures that the model not only interprets what a prefix means but also maps semantic concepts back to the corresponding sub-token space.
2. **Part-to-Whole Semantic Prediction.** We structure the data to elicit a two-step semantic prediction process. The first step predicts the fine-grained semantic explanation for each individual sub-token within the complete sub-token sequence corresponding to a given item, while the second step subsequently synthesizes these discrete semantics into a cohesive, holistic caption for the item. The data thus enforces a structured, multi-step semantic prediction paradigm from parts to the whole.

Item Granularity. Similar to the *Itemic Dense Caption Data* introduced in OpenOneRec [Zhou et al. \(2026\)](#), which comprises a bi-directional mapping between itemic patterns and captions, our item-granularity data serve to connect itemic patterns with natural-language descriptions. However, our item-granularity data significantly advances this paradigm through two critical enhancements:

1. **Capacity-Aware Caption Coarse-Graining.** The captions used as item-granularity alignment targets are often packed with very fine-grained details, yet each item is represented by only three sub-tokens. These sub-tokens capture an item’s coarse semantics but cannot encode such fine details, so aligning them with an over-detailed caption forces the model to hallucinate and pollutes the semantic space. We therefore coarse-grain each caption before alignment, keeping only what three sub-tokens can reliably carry, via three operations. (i) We remove instance-specific noise that the item code cannot recover, such as on-screen OCR text, ASR lyrics, dates, and product model numbers. (ii) We map fine-grained continuous attributes onto coarse buckets—an exact price becomes a price band and an exact age becomes a semantic label such as *young adult*—and we drop decorative modifiers while keeping the underlying entity (e.g., “pink top with a brown skirt” becomes “top with a skirt”). (iii) We preserve the high-discrimination content skeleton, including the item category, brand, IP, and title names, core selling points, target audience, and key material. Since the noise distribution differs across business domains, we design a separate set of coarse-graining rules for video, product, live, and ad.
2. **Multi-Perspective Item QA.** While standard captioning builds an initial perception of an item, it often falls short of capturing its nuanced attributes. To deepen alignment, we further construct multi-perspective question-answering data that queries the model from various angles, including target-audience preferences, core product attractions, visual styles, and negative-feedback rationales. This task design elevates alignment from a singular descriptive mapping to a robust, multi-dimensional semantic interrogation of the itemic pattern.

Relational Granularity. In recommendation scenarios, semantic alignment remains incomplete without capturing the collaborative connections between items. Existing paradigms typically rely on implicit collaborative-filtering signals from raw user interaction sequences, leaving item-to-item transitions devoid of explicit and interpretable semantic grounding. To bridge this gap, our relational-granularity data explicitly aligns these collaborative signals with the textual space by constructing interleaved sequences of itemic patterns and natural-language explanations. Specifically, each training sample takes the following form:

$$\text{Itemic_Pattern}_0 \rightarrow \text{Textual_Explanation}_0 \rightarrow \text{Itemic_Pattern}_1 \rightarrow \dots \rightarrow \text{Itemic_Pattern}_n. \quad (1)$$

This data corpus encompasses both **direct item-to-item associations** (where $n = 1$) and **multi-item interest flows** (where $n > 1$).

We use the internal *search-after-play* table to construct **direct item-to-item associations**. Each row in this table links a consumed source video to a specific search result product. This connection is established when a user issues a search query after a valid playback and positively engages with a product from the search results. The accompanying natural-language explanation is generated by fusing the captions of the video and the product into predefined templates.

Compared with direct item-to-item associations, the **multi-item interest flows** not only feature greater lengths and encapsulate richer collaborative signals per sample, but also provide broader modality coverage, ensure inter-item non-triviality, and yield higher explanation quality.

1. To expand the **modality coverage**, we further include two data sources apart from the *search-after-play* table. The first source is the internal *TagNext CF Relations*, where collaborative video pairs are extracted from positive user behaviors. The second involves aggregating co-occurring item pairs within sliding windows over temporally ordered, multi-domain user behavior sequences. Specifically, an item pair is retained if a substantial number of users provide positive feedback on two items that appear within a specific window size in their temporal behavior sequence.
2. To improve the **inter-item non-triviality**, we employ a dedicated process. First, we build a cross-user global item graph based on the three aforementioned data sources and randomly sample item chains from it. The sampled item chains capture latent relations that may not be prominent within a single user’s behavior sequence. Since transitions between adjacent items in a sampled chain $[Item_0, Item_1, Item_2, \dots]$ may be somewhat intuitive, we further sample a sub-chain $[Item_0, Item_{\text{interval}}, Item_{2 \cdot \text{interval}}, \dots]$ to form the training sample. Sub-chains are discarded if the semantic similarity between any two adjacent item nodes within the sub-chain exceeds a certain threshold.
3. To boost the **explanation quality**, we leverage an LLM to generate natural-language explanations that articulate the underlying interest transitions from $Item_{i \cdot \text{interval}}$ to $Item_{(i+1) \cdot \text{interval}}$. To ground these explanations in real-world transition paths, the LLM is also provided with the semantic captions of the intermediate item nodes $[Item_{i \cdot \text{interval}+1}, \dots, Item_{(i+1) \cdot \text{interval}-1}]$.

User Granularity. As the final part of our pre-training recommendation corpora, the user-granularity data expands the cross-modal alignment to the complete sequential evolution of user interactions.

A key prerequisite for this alignment is high-quality user-side text. However, raw user profiles assembled from business logs typically suffer from noise, incoherence, and rigid template patterns. We therefore employ an open-source LLM to recaption each profile across diverse aspects and formats, transforming them into fluent and natural narratives while preserving the original factual content. Building on these purified profiles, we design two data formats that seamlessly integrate natural language with itemic patterns to enable deep modality alignment:

1. **Domain-Grouped Behavior Sequences.** Each sequence in this format is anchored by a textual user profile, followed by the itemic patterns derived from the user’s interaction history. The involved itemic patterns are grouped into sub-sequences by their domains. Rather than unfolding the domain-grouped subsequences as a flat narrative, we dynamically structure them through multi-turn QA dialogues. By using natural language to explicitly query specific domain behaviors—often conditioning the prediction in one domain on the history in another—the text acts as an active control signal. The model must comprehend the semantic constraints of the question and map them accurately into the itemic space, thereby solidifying a robust alignment between the textual and itemic modalities.
2. **Chronologically Interleaved Behavior Sequence.** This data format reconstructs the user behavior sequence strictly by timestamps, naturally interleaving interactions across all domains. This timeline-based restructuring enables the model to differentiate short-term behavioral bursts from long-term interest evolution. Most importantly, to further enforce cross-modal alignment, we probabilistically replace a subset of itemic patterns within this timeline with their corresponding textual captions. This creates an interleaved, hybrid timeline that forces the model to achieve deep semantic integration between continuous natural language and discrete itemic patterns within a unified sequential context.

4.2.2. General-Domain Corpora

Text Corpora. Beyond the four-granularity recommendation corpora, we maintain a substantial proportion of general-domain text data throughout pre-training, spanning mathematical reasoning, code, scientific literature, medical text, and general instruction-following corpora. This mixture serves to preserve the model’s core capabilities in logical reasoning, structured knowledge understanding, and instruction adherence while it absorbs recommendation-specific knowledge.

Multimodal Corpora. Beyond text-only corpora, we further enrich pre-training with diverse multimodal data, which we convert into itemic-text format by tokenizing the images into itemic patterns. These data span four complementary sources: **image-text understanding pairs**, **text-image interleaved documents**, **image-editing triplets** (source image, editing instruction, target image), and **text-to-image generation data**. Across all four sources, we selectively retain only samples whose visual semantics operate at a coarse granularity—such as overall image style, dominant subject matter, and high-level scene composition—while discarding samples that hinge on fine-grained details that exceed the representational capacity of three sub-token codes. This deliberate diversity in data format, together with the coarse-granularity filter, jointly serves three purposes:

1. First, each source brings a distinct modality composition and linguistic style, collectively acting as a regularizer that prevents the model from overfitting to the homogeneous patterns of rule-synthesized recommendation corpora.
2. Second, these datasets are curated with high-quality semantic annotations, providing clean supervision signals for concept grounding and compositional visual reasoning.
3. Third, by mapping open-world visual concepts into itemic patterns, they function as external semantic anchors that expand the model’s knowledge boundary beyond the closed recommendation domain, thereby maintaining general multimodal-understanding capability throughout intensive recommendation training.

Data Composition Strategy. Since we continue pre-training from the instruction-tuned Qwen3 backbone (Qwen3 et al., 2025), a central concern is injecting domain knowledge without eroding its

pre-existing instruction-following ability and textual fluency (Parmar et al., 2024; Ovadia et al., 2025). To this end, and unlike OpenOneRec (Zhou et al., 2026), we adopt a unified next-token prediction objective and route noisy, template-synthesized business data into the loss-masked context, reserving the loss-bearing target for higher-quality text. Consequently, the effective training targets naturally exhibit a roughly 1:1 ratio of QA-style and free-form data: the former keeps the model aligned with instruction formats, while the latter anchors it in natural text distributions. The loss-masked context further allows the model to condition on domain signals without learning to reproduce their noise.

4.3. Ablation Study on Recommendation Corpora

4.3.1. Experimental Settings

We perform the ablation study on the 0.8B model with a fixed budget of 30B tokens, keeping the computational cost manageable while our main results are obtained on the 8B model. The five configurations are defined as follows:

- **Exp1 (Baseline):** the standard OpenOneRec dataset, comprising original item-caption data and user-granularity data (including user-behavior data and user-persona grounding data).
- **Exp2 (Exp1 + Token):** incorporates our token-granularity data into Exp1, proportionally downsampling the original mixture to maintain the budget.
- **Exp3 (Exp2 + Item):** replaces the baseline item-caption data in Exp2 with our item-granularity data.
- **Exp4 (Exp3 + Relational):** integrates our relational-granularity data into Exp3, rebalancing the overall mixture accordingly.
- **Exp5 (Exp4 + User):** substitutes the baseline user-granularity data in Exp4 with our user-granularity data.

Notably, the token and relational granularity data introduce novel capabilities absent from the baseline, whereas the item and user granularity data serve as upgraded replacements for their baseline counterparts. This sequential pipeline effectively isolates the marginal contribution of each granularity while ruling out any confounding effects from increased data volume.

4.3.2. Results

To systematically demonstrate the effectiveness of our multi-granularity data construction pipeline, we present the performance trajectory across OneReason-Bench tasks (Table 2). The results reveal a clear capability progression, confirming that each granularity level successfully injects its targeted compositional or semantic competencies into the model.

Exp1 → Exp2: Token Granularity. Adding token-granularity data sharply improves **R0** retrieval and understanding across all domains (e.g., `Itemic Pattern Grounding_prod` 2.42% → 5.81%; `Item Understanding_ad` 16.37% → 37.86%). **R1** `Item2Item QA` surges from 0.00% to 20.57%, confirming that token-level QA-style data unlocks the QA format absent in the baseline. **R2** temporarily drops to 0.00%, indicating disruption of higher-level reasoning.

Exp2 → Exp3: Item Granularity. Replacing the baseline’s captions with multi-perspective item data pushes several **R0** retrieval metrics to their peaks. **R2** recovers to 0.32%, and **R3** cross-domain recommendation improves (e.g., `Cross-Live` 2.32% → 3.49%). **R0** `Item Understanding` regresses on most domains, attributable to residual noise in business-sourced captions (see Table 3).

Table 2 | Each experiment builds on the previous one by adding a new granularity or replacing an existing component. We report pass@64 (%) for Itemic Pattern Grounding and cross-domain recommendation, and macro-weighted F1 (%) for the remaining tasks. All values are reported as percentages. The results show a consistent performance progression, with the largest gains in cross-domain recommendation emerging after User-granularity data is introduced.

Task		Exp1	Exp2	Exp3	Exp4	Exp5
R0: Itemic Pattern Grounding	ad	0.48	0.80	2.40	0.48	<u>1.92</u>
	live	0.40	<u>1.01</u>	1.21	1.21	1.21
	prod	2.42	5.81	<u>5.33</u>	2.91	<u>5.33</u>
	video	0.35	0.35	0.87	1.05	1.05
R0: Item Understanding	ad	16.37	37.86	31.65	29.69	<u>32.56</u>
	live	35.16	46.16	<u>39.27</u>	37.39	35.52
	prod	20.58	28.62	<u>29.51</u>	29.78	28.75
	video	22.56	36.76	30.77	28.68	<u>31.63</u>
R1: Item2Item QA		0.00	20.57	20.73	<u>25.65</u>	29.72
R2: Evolution Direct Gen	mixed	0.13	0.00	<u>0.32</u>	0.18	0.37
R3: Cross-Domain Rec.	ad	9.06	8.75	<u>9.54</u>	8.58	10.84
	live	2.29	2.32	<u>3.49</u>	3.25	8.56
	product	1.15	0.94	1.29	<u>1.54</u>	1.65
	video	0.65	0.74	0.72	<u>0.73</u>	0.66

Exp3 → Exp4: Relational Granularity. Integrating relational data lifts **R1** Item2Item QA to 25.65% and **R3** Cross-Prod to 1.54%. Some **R0** retrieval metrics decline (e.g., Itemic Pattern Grounding_prod 5.33% → 2.91%), reflecting a trade-off toward relational reasoning. **R2** dips slightly (0.32% → 0.18%), since item-to-item relational data does not directly benefit this user-side task.

Exp4 → Exp5: User Granularity. Substituting the baseline’s user data with our temporally-aware sequences pushes **R2** to its peak (0.37%) and drives **R3** cross-domain recommendation to its highest levels, most notably Cross-Live (3.25% → 8.56%) and Cross-Ad (8.58% → 10.84%). **R1** reaches 29.72%, and **R0** retrieval metrics recover to earlier peaks.

Key Insights. The ablation reveals three findings: (1) Token- and relational-granularity data expand the model’s skill set through augmentation, while item- and user-granularity data refine existing capabilities through replacement. (2) Capability trade-offs emerge during training—gains in relational reasoning may temporarily reduce static retrieval, but subsequent stages recover these capabilities. (3) User granularity acts as the ultimate integrator, contextualizing all prior knowledge within temporal behavior to achieve peak performance across diverse recommendation scenarios.

4.4. Training Recipe

Our training recipe, illustrated in Figure 7, unfolds in three stages. The core hyperparameters of each stage are summarized in Table 4.

Table 3 | A representative Item Understanding case in the video domain. For each model, we show its actual caption (excerpted; omissions are marked by ellipses) together with an English analysis. The baseline (Exp1) invents a drama title and degenerates into repetitive, contentless lines; after the token-granularity data is added (Exp2), the model stays faithful to the genre; once the item-granularity data is further incorporated (Exp3), the model fabricates a concrete, wrong drama title and re-frames the clip as a medical drama. Fabricated or incorrect content is marked in red.

Prompt: 请解读以下视频 token 并还原视频内容: < video_begin ><a_8164><b_6973><c_7395>			
Ground Truth	Exp1 (Baseline)	Exp2 (Exp1 + Token)	Exp3 (Exp2 + Item)
[Reference]	[Repetitive Hallucination]	[Faithful Genre]	[Fabricated Title]
<p>这是一段都市言情剧情短视频...女主角池晚作为神医救醒了病危的霍老爷子...霍老爷子为报恩极力撮合池晚与小儿子霍南洲结婚...前未婚夫霍司寒决定与池晚退婚、迎娶白月光池娇...两份婚书同时送到池家...最终池晚选择了霍南洲的婚书。</p> <p><i>Analysis:</i> a modern urban-romance drama about an arranged-marriage conflict within a wealthy family.</p>	<p>这段视频是关于一部名为《我爱的他》的电视剧的...通过对话和场景切换展现角色之间复杂的情感纠葛...角色之间的对话如“我爱的他”、“我爱的他”...揭示了复杂的情感纠葛和关系变化。</p> <p><i>Analysis:</i> invents a drama title and degenerates into repeated, contentless lines, capturing only a vague emotional conflict and nothing specific to the plot.</p>	<p>这些短视频主要属于剧情类短剧与影视解说...强剧情冲突与反转: 聚焦人物关系的复杂纠葛 (如豪门恩怨、复仇、误会)...鲜明的人物标签 (如“霸道总裁”、“灰姑娘”) ...面向喜爱现代都市情感、家庭伦理剧的受众...</p> <p><i>Analysis:</i> correct at the genre level (modern urban romance / family-conflict short drama) and consistent with the ground truth, though generic and not tied to the specific story.</p>	<p>这是一个关于《医祖凌云传》的故事...女性角色身着白色上衣展现专业形象, 男性角色穿着西装...画面中出现文字“医祖凌云传”...还出现了手术台、医疗器械等元素...兴趣点在于吸引喜爱《医祖凌云传》和关注医疗题材的观众。</p> <p><i>Analysis:</i> after item-caption data is added, the model fabricates a concrete, wrong title and re-frames the clip as a medical drama, contradicting the actual urban-romance plot; the hallucination worsens.</p>

Table 4 | Summary of training stages.

Configuration	Stage 1	Stage 2	Stage 3
Trainable parameters	Extended vocab. + LM head	All	All
Learning rate	$2 \times 10^{-4} \rightarrow 1 \times 10^{-4}$	$1 \times 10^{-4} \rightarrow 1 \times 10^{-5}$	$1 \times 10^{-5} \rightarrow 1 \times 10^{-6}$
Token budget	110B	449B	19B

All stages employ sample packing to improve sequence utilization, producing packed sequences of 48K tokens. The key distinction across stages lies in the maximum length of individual samples within each packed sequence. Stage 1 and Stage 2 restrict individual samples to 4K tokens to prioritize training throughput. In Stage 1, only the newly introduced embeddings and the corresponding LM head weights are trained, while the rest of the backbone remains frozen, allowing the newly

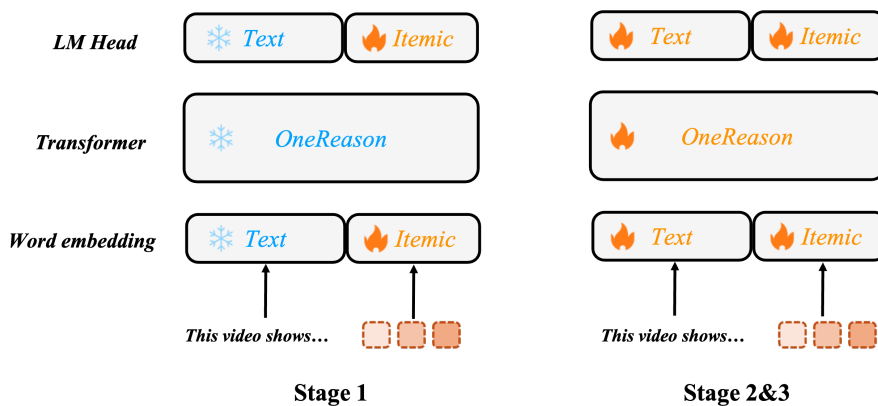


Figure 7 | The three-stage pre-training recipe. Stage 1 only tunes the incremental itemic-pattern-related parameters, while Stages 2 and 3 tune all parameters.

introduced itemic-pattern embeddings to settle into the semantic space without disturbing the pre-trained weights. Stage 2 unfreezes all parameters for full joint training, enabling the model to absorb recommendation knowledge from the four-granularity data mixture. Stage 3 lifts the per-sample length limit to 32K tokens, allowing the model to train on full user histories with long-range behavioral dependencies.

5. SFT Pipeline

Itemic–text alignment pre-training gives the model a necessary perception foundation: it can associate itemic tokens with natural-language content and has absorbed collaborative signals from user behavior. SFT builds on this foundation by cultivating recommendation cognition, namely the ability to operate over grounded itemic tokens under instruction format. We organize this stage around the same R0–R3 hierarchy, but with two complementary training axes. The *compression axis* teaches the model to reduce long and noisy user histories into typed persona states and compact interest-evolution motifs, so that the later reasoning step can compare a small set of evidence-backed hypotheses rather than a raw behavior log. The *reasoning axis* teaches dynamic operations: R1 derives one-hop item-to-item bridges, R2 tracks temporal interest evolution, and R3 performs transition judgement for next-interaction prediction. Section 5.1 introduces R0 perception supervision; Section 5.2 and Section 5.3 describe R1 derivation and R2 evolution reasoning; Section 5.4 presents R3 recommendation supervision, where compression and transition judgement are composed into recommendation CoT; and Section 5.5 and Section 5.6 cover itemic-instruction and general-domain data for instruction robustness.

5.1. R0: Perception

Building on the four-granularity alignment established during pre-training, we further refine the data and carry out high-quality CoT capability construction. R0 provides the perception layer for the SFT stage. Its role is to make itemic tokens readable, controllable, and semantically usable before the model is asked to perform higher-level recommendation reasoning. The caption data covers four major content domains, short videos, e-commerce products, live streamings, and advertising. QA data is narrower: it covers short-video and e-commerce product domains, while live-streaming and advertising contribute caption-style grounding only. R0 therefore contains two complementary task families. The first is **bidirectional mapping between itemic tokens and natural-language descriptions**, which reinforces item↔language grounding under instruction-style supervision. The second is **question-answering (QA) over platform content**, which expands perception beyond caption reconstruction to attribute recognition, audience understanding, negative-feedback analysis, and open-ended semantic comprehension for the two QA-covered domains.

Since the pre-training stage has already established the hierarchical itemic-token space, R0 does not introduce a new tokenizer or repeat the full token-granularity construction. Instead, it turns this hierarchy into explicit SFT supervision. Table 5 shows this difference with the same video itemic tokens: pre-training provides a direct itemic-token-to-caption mapping, whereas R0 asks the model to explain the hierarchical itemic tokens before producing the caption. In this case, the CoT first uses the coarse token to identify a character-centered or ACG-style content type, then uses the middle token to narrow the evidence toward young-female visual presentation, and finally uses the fine token to support the conclusion about personal style and aesthetic display. The key difference is therefore not the target caption itself, but the supervised verbalization of how coarse-to-fine itemic-token semantics support that caption. More bilingual R0 perception examples are provided in Appendix D.1.1.

The caption data includes both thinking and non-thinking variants. In the thinking format, the

model first explains how each itemic token contributes to the final item meaning and then produces the final caption. This explicit coarse-to-fine trace teaches itemic-token-to-text grounding rather than recommendation decision-making: it supplies the semantic substrate later reused by R1 derivation, R2 evolution, and R3 recommendation reasoning. The unCoT variant, in parallel, supervises direct itemic-token-to-caption generation, preserving generation quality and efficiency when explicit reasoning is unnecessary.

Table 5 | Representative R0 case comparing item-caption pre-training with instruction-style SFT perception supervision.

Pre-training itemic-token-to-caption mapping

User: <|video_begin|><a_7300><b_7894><c_4541>

Assistant: 视频为一段以二次元角色扮演为主题的短视频，主要展示一位身着黑色紧身战斗服、金发的女性角色在铁轨上的摆拍画面，以及一位身穿黑色吊带裙的女性在停车场奔跑的片段。视频配有“甜妹”“萌妹子”等标签，整体风格偏向二次元美学与个人形象展示。

Signal: pre-training provides direct itemic-token-to-caption grounding. It teaches the model to associate the itemic tokens with the caption, but does not explicitly verbalize how the individual itemic tokens support the description.

SFT instruction-style perception

User: 视频token <|video_begin|><a_7300><b_7894><c_4541> 包含哪些视频内容?

Assistant: <think> 从<a_7300> 来看，该视频属于以角色或人物形象展示为核心的内容类型，常见于二次元文化、Cosplay 或情感化美学展示领域；结合<b_7894>，可以进一步确定主角是年轻女性，内容围绕美女展示、时尚穿搭或Cosplay 展开；再融合<c_4541>，视频在视觉呈现上具有较强个人风格和审美导向。三层语义交集指向以年轻女性角色形象为核心、融合二次元或Cosplay 元素、注重视觉美学与情感氛围营造的展示类短视频。</think>

视频为一段以二次元角色扮演为主题的短视频，主要展示金发女性角色摆拍、女性奔跑等片段，配合“甜妹”“萌妹子”等标签，吸引喜爱美女欣赏和视觉审美的年轻观众。

Signal: the same itemic-token evidence is converted into instruction-followable perception; the CoT makes the coarse-to-fine grounding explicit before the final caption.

QA supervision is retained as an instruction-style perception supplement: it extends caption grounding from itemic-token decoding to content-aware question answering over structured attributes and open-ended semantics in the video and product domains. Together, caption and QA data convert the aligned itemic-token perception acquired during pre-training into instruction-followable perception ability. In total, R0 contains approximately **941K** samples, including about **682K** caption samples and **259K** QA samples.

5.2. R1: Derivation

R1 targets **cross-item relation derivation**. It extends R0 from understanding an individual item to explaining why one item may naturally lead to another. The intended capability is not generic item similarity. Instead, R1 teaches the smallest useful *source-to-follow-up bridge*: given a source item during student SFT, the model should generate and justify a plausible follow-up itemic-token answer by identifying the need, clue, scenario, constraint, or refinement on the source side. Source-destination pairs are used during teacher-side data filtering and consistency checking; the student training input remains source-only, and the source prompt tokens are not counted as target loss. This local bridge-building ability later becomes part of R3 recommendation CoT, where the model must connect observed behavior, textual evidence, and target-domain decisions rather than rely on surface textual overlap.

R1 is constructed from the same raw candidate pools used in relational pre-training, including TagNext CF relations and after-play-search data. The difference is the supervision standard. Relational pre-training exposes the model to large-scale collaborative transitions, whereas R1 requires

each retained pair to support explicit, one-hop, evidence-grounded reasoning. After deduplication, diversification, and embedding-similarity filtering, a locally deployed flagship LLM assigns an explicit-relevance judgement to each pair: *explicitly related*, *explicitly unrelated*, or *uncertain*. We retain only pairs whose relation can be directly supported by visible item evidence, such as item metadata and dense captions, and remove weak pairs whose connection depends mainly on broad assumptions such as audience overlap, popularity, or platform co-occurrence. The full judgement prompt and its English translation are provided in Appendix D.2.1.

Table 6 | Representative R1 case comparing message-style relational pre-training with one-hop SFT derivation supervision.

Pre-training one-hop data

User: 商品<|prod_begin|><a_2094><b_4972><c_1852>与什么视频内容最相关?

Assistant: <|video_begin|><a_1776><b_5416><c_9172>

SFT one-hop reasoning

User: 请推荐一条与<|video_begin|><a_1303><b_7205><c_8005> 主题接近、语义相关的视频。

Assistant: <think> 我注意到源视频虽然主要展示金刚板等建材的价格与材质，但其描述的温馨家居场景和具体的平米计价方式，精准触达了用户对于装修成本控制的深层关注。这种对硬装材料性价比的敏感，往往会自然激发用户对于整体空间软装搭配的联想。当用户开始计算地板或板材预算时，他们已经在构建一个完整的居住画面，此时床品、地毯等软装不仅是装饰，更是需要纳入同一笔账目的重要支出。基于这种由硬到软的推导逻辑，用户接下来的浏览轨迹很可能滑向家居软装搭配、床品选购或整体空间色彩规划内容。</think>

<|video_begin|><a_0474><b_3348><c_1453>

Signal: SFT turns the pair into an explicit source-to-target derivation. The target itemic tokens are the answer, while the reasoning must justify the bridge from source-side evidence.

Table 6 contrasts R1 with relational pre-training. Relational pre-training presents the source item and the related target item in a message-style association. R1 instead keeps a single source itemic-token sequence as the instruction input and supervises the answer side with an explicit bridge followed by the target itemic tokens, making the local relation judgement directly learnable.

For high-confidence candidates, the flagship LLM further extracts compact bridge variables, including source-side need, bridge type, abstract bridge, continuation direction, reason seed, and confidence. These variables separate the relation explanation from the destination item’s surface content: the destination item is used only for consistency checking, while the generated explanation must be written from the source-side evidence and bridge variables. This design reduces target leakage and prevents the model from learning to copy target-only entities or restate the destination description as the rationale. The generation prompt is shown in Appendix D.2.2, and bilingual sanitized examples are provided in Appendix D.2.3.

This process distills the raw candidate pools into approximately 358K TagNext CF pairs and 388K after-play-search pairs, from which we sample a refined R1 training set of roughly 400K examples. Overall, R1 converts implicit item-to-item signals into concise derivation supervision: it preserves the collaborative signal introduced during pre-training, but upgrades it into explicit bridge reasoning that can be reused by R3 when connecting user evidence to the next-interest direction.

5.3. R2: Evolution

R2 targets **user interest evolution modeling**. It extends recommendation cognition from static preference understanding to temporally structured interpretation of behavior. While R0 grounds item semantics and R1 derives explicit relations between item pairs, R2 asks how intent develops

over time: which behaviors act as triggers, which later behaviors refine or correct earlier ones, and how multiple events accumulate into an interpretable trajectory. This capability is essential when the recommendation decision depends not only on what the user interacted with, but also on how the user’s underlying need became more specific.

We construct R2 from full-domain user behavior timelines, where heterogeneous actions are organized chronologically. Rather than treating a timeline as a flat context, a strong LLM first identifies key evolution nodes that indicate meaningful shifts, refinements, or closures in user intent. The extraction emphasizes coherent progression, topic continuity, and nontrivial cognitive increment, converting raw behavior streams into candidate evolution chains with explicit temporal structure. In this sense, R2 is the SFT counterpart of user-level pre-training: pre-training exposes the model to long behavior contexts, while R2 teaches it to explain which parts of those contexts form valid interest-evolution paths.

To improve reliability, candidate chains are filtered by a separate LLM-as-a-Judge quality-evaluation stage. This stage checks whether each proposed trajectory is supported by the original timeline and whether its transitions form a plausible, self-consistent reasoning path. The diagnostic criteria cover order sensitivity, cognitive increment, trigger-source evidence, strong causal-style exclusivity, evidence closure, and no-mind-reading constraints. In practice, this filtering removes pseudo-logical chains, random-consumption chains, same-category drift, and weakly grounded transitions before downstream task construction. The main extraction prompt and its English translation are provided in Appendix D.2.4; the complete quality-control prompt contains the full set of eleven checks used for second-pass evaluation.

Based on the validated evolution chains, we construct three SFT task families aligned with the R2 benchmark taxonomy:

- **Evolution Action Selection:** Given a user history \mathcal{H} and a specified interest topic or candidate chain description, the model identifies the related historical actions \mathcal{A} , encouraging discriminative reasoning over temporal structure.
- **Evolution Topic Generation:** Given a user history \mathcal{H} and a specified topic t , the model generates the corresponding evolution chain \mathcal{E}_t , focusing on topic-conditioned temporal reasoning.
- **Evolution Direct Generation:** Given only the user history \mathcal{H} , the model directly generates valid evolution chains $\{\mathcal{E}\}$ without topic guidance, requiring the highest degree of autonomy.

During data construction, these three families are organized as *selection*, *topic-generation*, and *direct-generation* samples, respectively. Some samples include CoT rationales that make the temporal-transition judgement explicit, while direct-answer samples are kept for response efficiency and format robustness. Representative sanitized examples for all three training formats are provided in Appendix D.2.5.

The final R2 dataset contains approximately **130K** training instances. Overall, R2 serves as the temporal cognition layer of the SFT mixture. By transforming user timelines into validated evolution supervision under multiple task formats, it teaches the model to interpret user intent as an evolving process rather than a static profile. This temporal scaffold is later reused by R3, where recommendation traces must compress noisy histories, compare candidate interest directions, and decide which direction is best supported by the user’s recent and long-term behavioral evidence.

5.4. R3: Recommendation

R3 is the recommendation-level composition layer of the SFT mixture. R0–R2 train dedicated capabilities separately, whereas R3 brings them together in the final next-interaction prediction

setting. Its contribution is not another standalone reasoning task, but a trace-construction protocol that places R1-style derivation and R2-style temporal-evolution reasoning over a two-axis compressed user state. The R3 block contains both thinking samples and non-thinking direct-answer samples: the former teach structured recommendation cognition, while the latter preserve direct mapping from user context to target itemic tokens and support non-thinking format alignment. This subsection focuses on the construction and evaluation of the CoT traces.

5.4.1. Recommendation Reasoning Trace Construction

Each R3 synthesis sample is built from three ingredients: a user profile \mathcal{P}_u , a chronological interaction timeline \mathcal{H}_u , and the ground-truth subsequent interaction y_u^+ . The target interaction is used only as the supervision answer and is explicitly prohibited from appearing directly in the reasoning trace. The trace must therefore reason from the observed profile and behavior evidence as if the answer were unknown. During data construction, the teacher first writes the trace over natural-language user evidence and marks referenced historical items with their original domain ids, such as video, product, live-author, or ad ids, so that every concrete evidence reference remains machine-matchable. A subsequent post-processing stage joins this raw trace with the itemic-token-converted user context, replaces each matched historical id with its itemic tokens, removes unmatched references, and writes the final assistant message as a <think> trace followed by the target itemic-token answer. Thus, the training trace seen by the model is not a text-only rationale: it interleaves natural-language reasoning with itemic-token evidence inside the CoT, while keeping the target itemic tokens outside the reasoning span.

The central design principle is to treat recommendation CoT as a **two-axis compression plus transition judgement** process. The first axis is persona-type compression, which summarizes stable preference priors, demographic or life-stage clues, temporal rhythm, content preference, price sensitivity, interaction depth, and possible shared-device ambiguity. The second axis is interest-evolution-type compression, which identifies recent trajectory motifs such as triggering search, need expansion, parameter narrowing, scenario continuation, saturation-driven substitution, cross-domain echo, or closure from browsing to purchase. These compressed states reduce long-history noise and provide typed variables that the subsequent reasoning stage can manipulate.

Compression in this setting is task-directed understanding, not a generic summary of the observed history. It preserves evidence useful for next-interaction prediction and suppresses routine, accidental, or weakly supported behavior. This design is motivated by the information bottleneck view of representation learning, where useful representations preserve prediction-relevant information while suppressing input noise (Tishby and Zaslavsky, 2015); we use this as a design motivation rather than a claim about universal deep-network dynamics (Saxe et al., 2018). It is also consistent with work connecting language modeling and compression (Delétang et al., 2024) and with the view that reusable abstractions compress long primitive derivations into higher-level concepts (Aksenov et al., 2026). In our setting, the practical implication is that a model capable of compressing a noisy user history into an appropriate persona state and evolution motif has already performed a substantial part of recommendation cognition.

After compression, the trace performs transition judgement over the candidate hypotheses. This stage reuses two earlier SFT abilities. From R1, it inherits local bridge construction: the trace should explain how an earlier item, query, or behavior can naturally lead to a follow-up need. From R2, it inherits temporal evolution judgement: the trace should decide whether a candidate direction is supported by order, continuity, refinement, or closure in the behavior timeline. We use *causal-style* only as a data-construction heuristic: the teacher prefers evidence-supported triggers, refinements, and scenario continuations over shallow co-occurrence, without claiming identifiable

causal effects from observational logs. This heuristic is aligned with causal recommendation studies that emphasize separating genuine user interest from confounding factors or conformity effects, as well as counterfactual explainable recommendation methods that seek explanations with strong decision influence (Tan et al., 2021; Zheng et al., 2021; Luo et al., 2024).

Concretely, we adopt a three-stage protocol: **Persona Abstraction**, **Interest Expansion**, and **Transition Inference**. **Persona Abstraction** performs the persona-type compression axis. **Interest Expansion** starts from interest-evolution-type compression and expands the most informative behavior signals into a small set of evidence-grounded hypotheses. **Transition Inference** compares these hypotheses under evidence strength, recency, temporal continuity, persona compatibility, target-domain compatibility, and leakage control, then commits to the most plausible next-interest direction.

Table 7 | Representative persona-abstraction cases for recommendation reasoning. Personas are used only as soft priors; final decisions must be supported by user behavior.

Persona type	Chinese reasoning case and English analysis
Middle-aged homemaker	<p>由【晚饭前连续看家常菜教程、收纳改造、孩子作业辅导】推测用户偏向中年家庭主妇，并且最明显的需求是家人日常饮食、家居整理和孩子学习支持。后续更应关注实用、低门槛、能直接服务家庭场景的内容，而不是泛泛推荐生活方式内容。</p> <p><i>Analysis:</i> the abstraction compresses several household-oriented signals into a family-consumption prior, but the next recommendation still has to follow the concrete need revealed by recent cooking, storage, or tutoring behaviors.</p>
Live-shopping enthusiast	<p>由【长时间停留在带货直播间、反复点击购物车、关注限时优惠】推测用户对直播购物偏好较强，并且最明显的兴趣集中在高性价比日用品、食品补货和家居小件。推理时需要优先判断她信任的主播、常购品类和促销节点，避免只把她归为普通购物用户。</p> <p><i>Analysis:</i> the persona prior highlights promotion sensitivity and trusted live-room categories, so the trace should reason about the next purchase scenario rather than merely predicting generic e-commerce interest.</p>
Shared-device user	<p>由【上午看少儿动画、晚饭前看家常菜、深夜又集中看体育赛事】推测该账号可能存在多人共用手机。此时画像不能直接合并成一个稳定用户，而要按时段和近期行为拆分，优先识别当前活跃使用者，避免被其他家庭成员的历史行为带偏。</p> <p><i>Analysis:</i> the abstraction is used as a noise-control mechanism: conflicting behavior clusters are treated as different possible users, and the final transition should rely on the current stable segment rather than the whole mixed history.</p>

At a conceptual level, the protocol can be written as

$$C_u = \text{Abstract}(\mathcal{P}_u, \mathcal{H}_u), \quad \mathcal{Z}_u = \text{Expand}(C_u, \mathcal{H}_u), \quad (2)$$

$$z_u^* = \text{Infer}_{\text{trans}}(\mathcal{Z}_u | C_u, \mathcal{H}_u, d) = \arg \max_{z \in \mathcal{Z}_u} s(z | C_u, \mathcal{H}_u, d), \quad (3)$$

where C_u denotes the compressed user state, \mathcal{Z}_u denotes the expanded set of candidate interest hypotheses, d denotes the target domain, $\text{Infer}_{\text{trans}}(\cdot)$ denotes the final transition-judgment step, and $s(\cdot)$ is a conceptual scoring function that jointly considers evidence strength, recency, temporal continuity, persona compatibility, and target-domain compatibility. The final CoT trace is written as a concise realization of this process with text and itemic-token references interleaved, while the target itemic tokens are emitted only after the reasoning trace. Representative sanitized training samples of the final recommendation CoT format are provided in Appendix D.3.3, and the complete bilingual construction prompt is provided in Appendix D.3.1.

Persona Abstraction The goal of **Persona Abstraction** is to compress sparse and noisy behavioral evidence into a compact, interpretable prior. Given the user profile and interaction history, the teacher model identifies the most plausible user archetype, or summarizes a customized profile when the user does not fit any predefined category. The abstraction is a soft, typed user state rather than a deterministic label: it narrows the set of likely interest directions, but the final inference must still be grounded in observed behaviors. This stage reduces the search space for transition judgement by converting a high-dimensional behavior log into a small number of preference factors, such as family consumption needs, game-skill improvement, live-shopping sensitivity, food preference, fitness goals, or shared-device ambiguity. Table 7 gives representative Chinese persona-abstraction cases, together with English analyses of how each abstraction constrains recommendation reasoning.

Interest Expansion After **Persona Abstraction**, the model performs **Interest Expansion** so that the trace does not commit too early to a single explanation. This stage begins with interest-evolution-type compression: it identifies typed motifs from recent trajectory evidence, such as a search-triggered need, a refinement from broad category to concrete parameter, a scenario continuation from content to product, or a cross-domain echo between live, product, ad, and video behaviors. These motifs are then opened into a small set of candidate next-interest hypotheses. In this way, **Interest Expansion** remains broad enough to represent recommendation uncertainty, but compact enough to prevent the trace from being overwhelmed by raw behavior tokens.

We use four heuristics to keep this expansion broad but evidence-grounded:

- **Behavior motivation.** The trace should explain the intent behind informative behaviors, especially search queries, because they often directly reveal the user’s current interest and real need.
- **Temporal sensitivity.** The trace should respect event order and explain how later behaviors may evolve from earlier ones, rather than treating the history as an unordered set.
- **Transition coverage.** When the user appears to shift from one topic to another, the expansion should preserve plausible before–after directions as candidates, while leaving the final judgement to **Transition Inference**. If a later behavior abandons all previous conditions without a logical bridge, it should be treated as weakly related or unrelated.
- **Noise attenuation.** Routine replenishment, random consumption, repeated exposure, and one-off actions are down-weighted because they are usually weak signals for stable future recommendation intent.

Concretely, **Interest Expansion** requires the reasoning trace to reach the finest reliable granularity supported by evidence. If the history supports interest in a specific mobile game, the trace should not stop at the generic gaming category; if the user repeatedly engages with a particular opera genre, the trace should mention that genre rather than only “traditional culture.” This design keeps the CoT trace fine-grained and evidence-grounded while preserving multiple plausible future directions before the final decision is made.

To calibrate how broad this stage should be, we vary the interest expansion width n during SFT trace construction and evaluate the resulting think-mode recommendation quality across the four domains. Figure 8 shows a consistent pattern: keeping the expansion compact ($n \in \{1, 3, 5\}$) is consistently better than broader widths such as 10 or 20. This result fits the compression view above. Once the candidate set becomes too large, the trace starts to carry many weak branches, which dilutes the strongest transition signal and makes the later compare-and-commit step noisier. A small hypothesis set is therefore not a loss of reasoning capacity; rather, it is the point where uncertainty

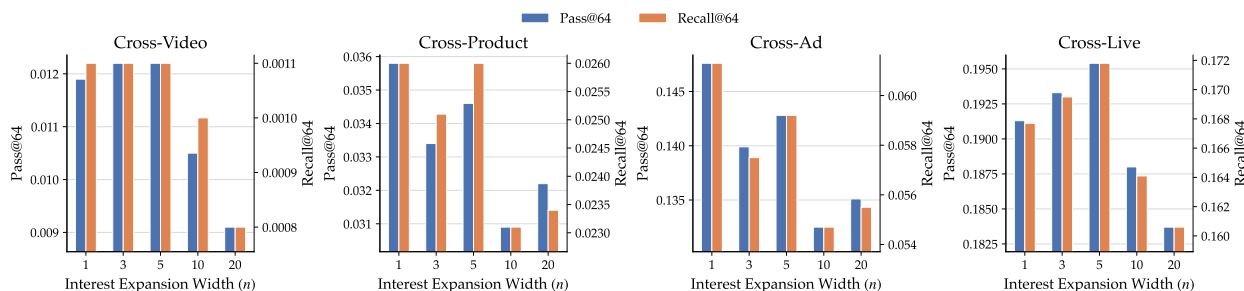


Figure 8 | Effect of interest expansion width n during Interest Expansion. Across the cross-video, cross-product, cross-ad, and cross-live domains, compact expansion with $n \in \{1, 3, 5\}$ consistently outperforms broader widths such as 10 or 20 on think-mode Pass@64 and Recall@64.

is preserved without letting the reasoning path fragment. In practice, we treat Interest Expansion as a controlled reasoning decompression step: it briefly opens the compressed trajectory into a few plausible directions, then hands a still-compact hypothesis space to **Transition Inference**.

Table 8 | R3 recommendation CoT case showing Interest Expansion and Transition Inference.

Interest Expansion

由【中年女性、家庭日用服装购买、关注穿搭/健康/房产家居】推测用户偏向家庭实用消费型用户，近期最明显的兴趣是家居清洁与子女衣着。从近期行为可发散出多个兴趣点：A. 家居清洁维护，表现为多次购买<|prod_begin|><a_662><b_7813><c_7791>等清洁产品，并长播<|video_begin|><a_192><b_6421><c_6240>等教程；B. 子女衣着，表现为搜索“青少年加长版的裤子新款男”并购买<|prod_begin|><a_6266><b_6810><c_2477>；C. 健康养生，表现为关注<|living_begin|><a_3241><b_7205><c_1272>、点击<|prod_begin|><a_6655><b_5260><c_329>及长播<|video_begin|><a_5695><b_1638><c_6707>；D. 零食/食品，表现为搜索“馋嘴小零食”并购买<|prod_begin|><a_6488><b_6398><c_7103>；E. 个人衣着与护肤，表现为购买<|prod_begin|><a_8099><b_2904><c_503>并长播<|video_begin|><a_1409><b_4885><c_5059>；F. 房产与民生资讯，表现为完播民生、时政类内容，并浏览<|video_begin|><a_4946><b_5797><c_6364>。

Analysis: the trace does not collapse the user into one label. It keeps several evidence-backed branches, each tied to concrete itemic-token evidence and a specific intent.

Transition Inference

综合权衡，A(家居清洁)和B(子女衣着)是当前最直接、最可能延续的强需求，因为它们同时具备搜索、点击、购买和长播等多种证据，且都服务于家庭场景中的即时问题解决。C(健康养生)和D(零食/食品)也有明确行为支撑，但更像阶段性补充需求，通常会间歇触发；E(个人衣着与护肤)与用户长期画像一致，但近期行动密度弱于A/B；F(房产与民生资讯)主要是信息浏览，不足以作为最强后续方向。因此，后续推荐应优先围绕家居清洁技巧、清洁产品测评、青少年男装选购以及实用的家庭生活小窍门展开，同时少量保留健康养生和食品补货作为辅助方向。

Target: <|video_begin|><a_6166><b_1595><c_1850>

Analysis: the final step chooses the branches with both recency and action density, then emits the target itemic tokens only after the reasoning trace.

Transition Inference The final stage compares the expanded hypotheses against the full context and commits to the most likely subsequent interaction direction. It reuses R1 by requiring a local bridge from antecedent evidence to follow-up need, and reuses R2 by checking whether the candidate direction is supported by temporal order, continuity, refinement, or behavioral closure. The teacher weighs each hypothesis by evidence strength, recency, continuity, persona compatibility, target-domain compatibility, and leakage risk. This compare-and-commit step prevents the trace from merely listing

possible interests without making a decision useful for recommendation.

The following constraints make the final judgement decision-oriented and leakage-safe:

- **Evidence priority.** Candidate directions supported by repeated actions, active searches, or recent dense interactions receive higher weight; those supported only by isolated or accidental behaviors are down-weighted.
- **Transition bridge.** The trace should explain the bridge from earlier evidence to later intent. Valid bridges include feedback, bottleneck, cognitive refinement, scenario continuation, parameter narrowing, or need completion; simple topical similarity is insufficient.
- **Granularity calibration.** The final judgement should be as specific as the evidence allows. If the evidence supports a concrete game, opera genre, function, product attribute, or content style, the trace should not stop at a coarse category; if the evidence is insufficient, the trace should avoid over-specific claims.
- **Conflict resolution.** When multiple directions are plausible, the trace compares their fit with the persona prior, recent behaviors, long-term interests, and target domain, then chooses the most likely direction.
- **Leakage control.** Although the target interaction is visible during data construction, the reasoning must be written as if the answer is unknown. It must not copy the target title, expose the target item id or itemic token, or use target-only entities as evidence.

A compact training trace in Table 8 shows how *Interest Expansion* keeps several grounded directions active, while *Transition Inference* commits to the strongest continuation.

Through *Persona Abstraction*, *Interest Expansion*, and *Transition Inference*, the synthesized CoT becomes compact, fine-grained, and decision-oriented. The first two stages denoise the behavior history and organize the hypothesis space, while the final stage compares candidate transitions by reusing R1 bridge reasoning and R2 evolution reasoning. This structure allows SFT to learn not only the final target itemic-token prediction, but also the behavioral logic that connects user evidence to recommendation decisions.

5.4.2. Quality Evaluation of Reasoning Traces

The quality of recommendation reasoning traces is evaluated along five complementary dimensions, each highlighting a different aspect of useful recommendation cognition: they assess whether the trace avoids direct answer leakage, aligns with the intended target, demonstrates logically transferable reasoning over user behavior, accurately reflects observed history, and provides actionable guidance for subsequent recommendations.

- **Safety.** Checks whether the CoT trace directly exposes target content, such as item id, itemic token, title, or any entity that could uniquely identify the target. Safe traces avoid direct answer leakage while remaining informative.
- **Consistency.** Evaluates whether the trace’s conclusions, preference direction, and behavioral explanations align with the intended recommendation target. A consistent trace supports correct downstream predictions.
- **Logic.** Measures whether the trace demonstrates transferable and generalizable reasoning over user behavior rather than surface-level concatenation, forced associations, or memory of a single instance.
- **Factuality.** Ensures that the trace accurately reflects the observed user history. It checks for fabricated behaviors, incorrect attributions, exaggerated trends, temporal confusions, or unsupported statements.

- **Informativeness.** Assesses whether the reasoning provides useful and specific recommendation guidance without leaking the target. It should summarize interest directions, scenarios, functionality, style, attributes, or exclusions that are actionable for recommendation.

Together, these dimensions provide a compact diagnostic view of trace quality. Safety constrains leakage, consistency checks target alignment, logic checks reasoning transferability, factuality checks history grounding, and informativeness checks whether the trace provides clear guidance for recommendations. The evaluation therefore distinguishes traces that are merely fluent from traces that express useful recommendation cognition. Figure 9 summarizes the resulting score distribution and mean-score profile across these five dimensions.

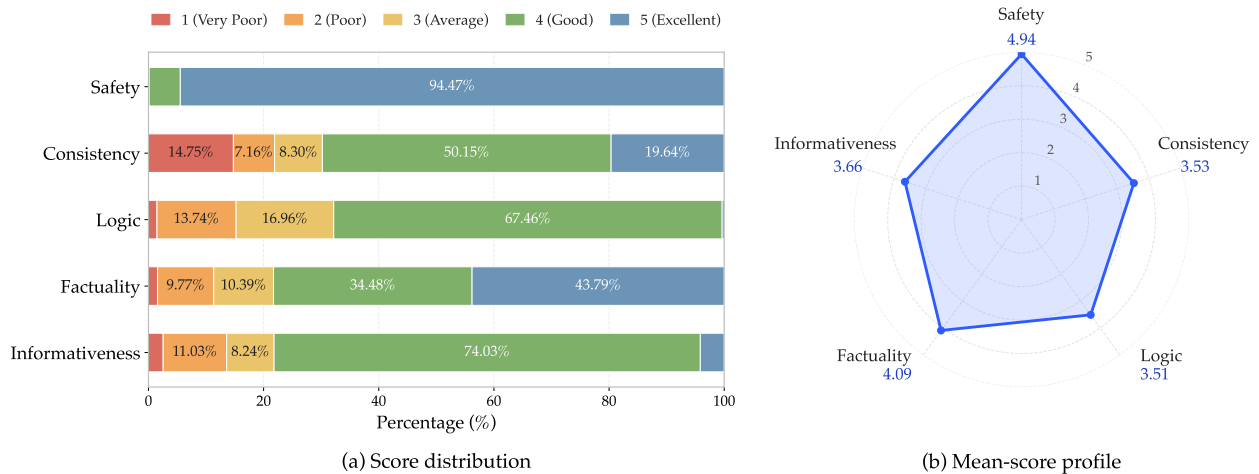


Figure 9 | Score distribution and mean-score profile for reasoning-trace quality across safety, consistency, logic, factuality, and informativeness dimensions.

Representative unreasonable traces identified by this quality-evaluation strategy are shown in Appendix D.3.2 (Table 28), where problematic statements in the generated CoT are marked in red.

5.5. Itemic Instruction Data

Itemic Instruction Data improves instruction-following behavior when itemic tokens appear explicitly in the input context. It complements the compression–reasoning backbone by ensuring that itemic-token-rich inputs remain instruction-controllable, rather than being automatically interpreted as recommendation-reasoning prompts. We observe that pretrained models often behave unstably in such settings: instead of following the user’s instruction, the model may ignore the requested task, explain the meaning of specific itemic tokens, or implicitly reinterpret the input as a user-interest analysis problem. Such behavior limits usability in practical recommendation-oriented interactions.

To address this issue, we construct a specialized instruction-tuning dataset centered on **itemic-token-grounded task execution**. Data generation begins with a predefined inventory of task tags, each corresponding to a representative instruction type in recommendation and content understanding scenarios. Rather than focusing only on itemic-token decoding, the tag space spans conversion, retrieval, matching, comparison, generation, editing, summarization, and explicit instruction-control tasks. For each tag, we generate suitable instructions and pair them with sampled user-side material, so that each example requires the model to complete a concrete task under inputs containing itemic tokens. Representative bilingual training cases are shown in Appendix D.4.1. The main task groups are organized as follows:

- **Conversion and Editing.** This group teaches the model to convert between itemic tokens and natural language, rewrite text-token pairs, replace itemic tokens with textual descriptions, or reorder token sequences under instruction constraints. A representative instruction is: “Convert the following product itemic token into a short product introduction.”
- **Retrieval and Selection.** This group focuses on retrieving itemic tokens from structured user contexts according to time, behavior, attribute, or interaction-order constraints. A representative instruction is: “Find all video itemic tokens the user interacted with during a given time window.”
- **Matching and Classification.** This group trains the model to match natural-language descriptions with candidate itemic tokens, select the most suitable token-description pair, handle fuzzy matching, or classify the scene represented by itemic tokens. A representative instruction is: “Which itemic token best matches the description ‘noise-canceling Bluetooth earphones’? Output only the token.”
- **Comparison and Aggregation.** This group requires the model to compare itemic tokens or token groups, perform set-level operations, and summarize token sequences into compact natural-language descriptions. A representative instruction is: “Compare the functional differences between these two product itemic tokens in three sentences.”
- **Generation and Verification.** This group asks the model to generate descriptions, promotional copy, constrained summaries, or verification results based on itemic-token-linked content. A representative instruction is: “Write a 50-word promotional copy for the following product itemic token.”
- **Instruction Control.** This group explicitly trains the model to follow the user’s instruction even when itemic tokens appear as distracting context. It includes ignoring irrelevant tokens, conditional execution, text-preserving rewriting, and format-constrained response generation. A representative instruction is: “Ignore all itemic tokens below and answer: what are the main evaluation metrics for recommender systems?”

A key design principle is that itemic tokens are treated as **grounding signals** rather than as the final target of explanation. The objective is not to make the model interpret token semantics in isolation, but to train it to use itemic-token-linked context correctly while remaining faithful to the user’s instruction. This improves robustness in mixed natural-language and token-based inputs. Among the task groups above, the instruction-control cluster is especially important because it directly teaches the model to prioritize user intent even when itemic-token-rich context could otherwise distract or mislead generation.

To ensure data quality, generated instructions are evaluated for usability before large-scale expansion, and final samples are checked with rule-based consistency tests to verify that itemic-token references in the input and output match the associated material. Samples with unsupported token references, weak task grounding, or unclear intent are removed. The final Itemic Token Instruction dataset contains approximately **103.0K** examples. Overall, this dataset serves as a behavior-alignment layer for itemic-token-aware interactions, improving the model’s ability to follow instructions accurately in recommendation scenarios involving structured content identifiers.

5.6. General-Domain Data

To preserve general instruction-following ability and mitigate over-specialization during recommendation-oriented SFT, we incorporate the publicly available **Step-3.5-Flash-SFT** dataset released by Step-Fun (Huang et al., 2026) as the general-domain data source. We apply conservative preprocessing to the raw dataset: samples containing tool-use traces are discarded to avoid introducing environment-interaction behaviors outside the scope of our SFT setting; only samples following the standard three-role dialogue schema, namely *system*, *user*, and *assistant*, are retained to ensure a homogeneous

conversational format; and samples with malformed structures, missing fields, or corrupted formats are removed to further improve data quality. After preprocessing, the resulting corpus contains approximately 1.5 M high-quality SFT samples, which are mixed into the training data to maintain broad instruction-following capability and stabilize general-purpose response behavior under domain-specific fine-tuning. General-domain data therefore serves as a stabilizer around recommendation cognition, preventing the model from overfitting to itemic-token reasoning formats.

Overall, this SFT stage builds structured recommendation cognition on top of aligned itemic perception: R0 grounds itemic tokens, R1 supplies evidence-grounded derivation, R2 supplies temporal-evolution supervision, and R3 composes these abilities for next-interaction prediction. Because the R3 traces are still teacher-generated, they are not yet optimized by recommendation outcome feedback. The next stage closes this loop with recommendation-oriented RL: domain-specific RL first discovers stronger recommendation behaviors within each domain, and RFT/MOPD then consolidate these specialized policies into a unified four-domain recommendation foundation model.

6. RL Pipeline

After the SFT stage, the model has acquired general semantic understanding, instruction-following ability, and an initial capability for recommendation-oriented reasoning. However, SFT mainly enables the model to imitate teacher-generated reasoning trajectories, and the resulting performance is still largely bounded by the capability of the teacher model. Since the model has already been exposed to large-scale recommendation data and has developed strong recommendation capabilities, a natural next step is to examine whether the model can go beyond imitation and further improve its recommendation reasoning through self-exploration. Reinforcement learning (RL) has recently become a promising direction for further improving recommendation models beyond supervised training (Tan et al., 2025; He et al., 2026; Zhou et al., 2026). However, for a foundation recommendation model that supports diverse tasks across domains, directly applying RL to mixed-domain data may cause cross-domain interference, since different domains often involve distinct user intents, item semantics, and reward landscapes.

To mitigate this issue, we adopt a “specialize-then-unify” strategy. As illustrated in Figure 10, we first optimize the model separately on each recommendation domain through RL, resulting in four teacher models that specialize in domain-specific recommendation reasoning. We then investigate two approaches to consolidate these domain-specific capabilities into one unified model: rejection sampling fine-tuning (RFT), which refines the model on verified high-quality reasoning trajectories sampled from domain-specialized teachers; and Multi-Teacher On-Policy Distillation (MOPD), which distills multiple domain-specific policies into the student model on its own generated trajectories.

6.1. Recommendation-oriented Reinforcement Learning

We first optimize the model with GRPO (Shao et al., 2024) on each recommendation domain separately, where the model generates trajectories and receives outcome feedback based on the correctness of its prediction.

Recommendation-oriented GRPO. For a user u with context (query) q , the model takes it as input and samples a group of recommendation rollouts, where each rollout consists of a reasoning trace and generated itemic tokens, as illustrated in Figure 11. Formally, let

$$\mathcal{G}_u = \{(\text{CoT}_{u,i}, \mathbf{c}_{u,i})\}_{i=1}^G \quad (4)$$

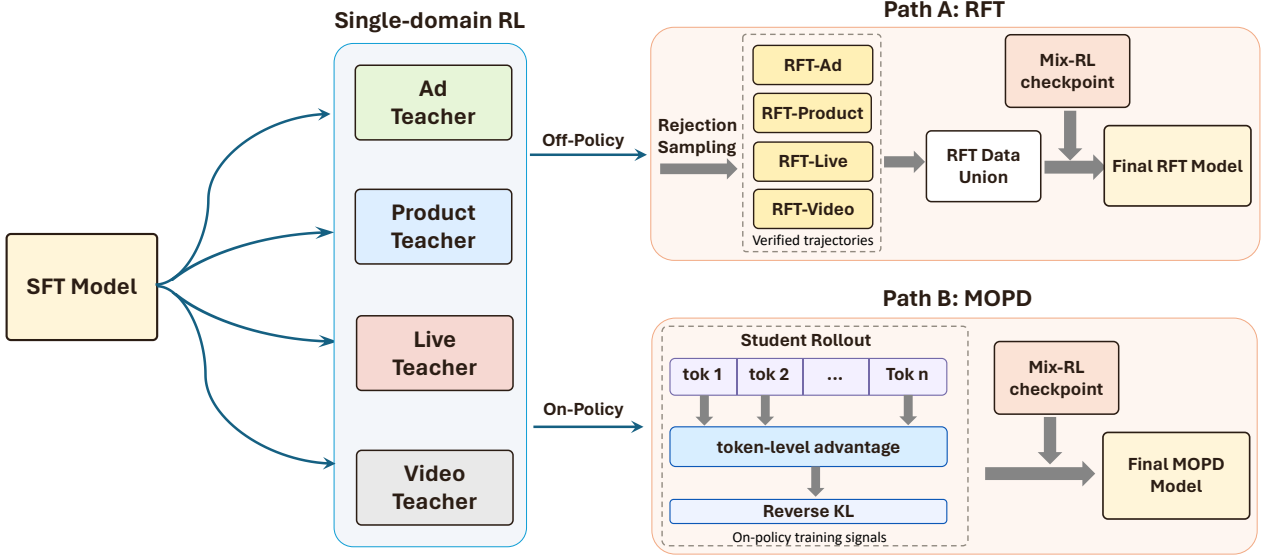


Figure 10 | Overview of the “specialize-then-unify” pipeline.

denote the group of G rollouts sampled for user u , where $\text{CoT}_{u,i}$ is the generated reasoning trace and $\mathbf{c}_{u,i}$ is the corresponding itemic tokens. Each rollout is evaluated by an outcome reward $R_{u,i}$, which measures whether the generated itemic tokens match the target recommendation. GRPO then normalizes rewards within the same group to obtain the relative advantage:

$$\hat{A}_{u,i} = \frac{R_{u,i} - \text{mean}(\{R_{u,k}\}_{k=1}^G)}{\text{std}(\{R_{u,k}\}_{k=1}^G) + \delta}, \quad (5)$$

where δ is a small constant added for numerical stability. The model is optimized to increase the probability of high-advantage rollouts, where the generated reasoning trace leads to itemic tokens that match the ground-truth items.

Rollout Design. While GRPO serves as a powerful tool to enhance the effectiveness of recommendation reasoning, directly applying it to recommendation tasks still faces practical challenges. The nature of recommendation tasks, where the model predicts user interests from a massive candidate space, leads to an inherently low hit rate and requires a sufficiently large rollout size to obtain useful reward signals. This is particularly costly for reasoning-augmented recommendation, where each recommendation is preceded by a reasoning trajectory.

To address this issue, we design a two-stage rollout strategy tailored to recommendation-oriented RL. As illustrated in Figure 11, the central idea is to amortize the cost of reasoning over multiple recommendations. For each user u , we first sample N reasoning traces $\{\text{CoT}_{u,i}\}_{i=1}^N$. Conditioned on each reasoning trace $\text{CoT}_{u,i}$, the model then generates K itemic token sequences in parallel, $\{\mathbf{c}_{u,i,j}\}_{j=1}^K$, where

$$\mathbf{c}_{u,i,j} = [\langle \text{domain_begin} \rangle, c_{u,i,j}^{(1)}, c_{u,i,j}^{(2)}, c_{u,i,j}^{(3)}]. \quad (6)$$

This produces $N \times K$ effective recommendation rollouts while generating only N reasoning traces. Since the additional expansion is restricted to the short itemic tokens, the reward coverage is significantly improved with limited extra computation and memory overhead.

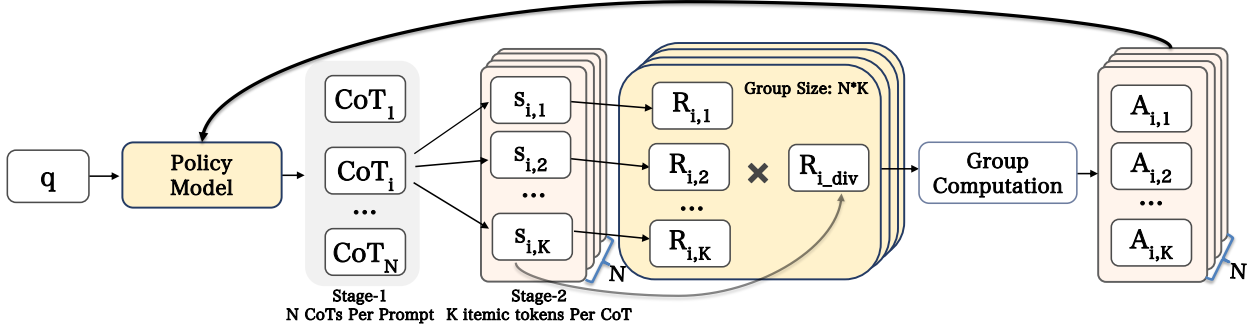


Figure 11 | Overview of recommendation-oriented GRPO.

Reward Design. Recommendation tasks focus on set-wise optimization, where the model should retrieve a set of relevant items rather than only optimize a single prediction. Since user interests are often multi-faceted, an ideal reasoning trace should cover diverse relevant items that satisfy user interests. Therefore, given the $N \times K$ rollouts produced by the two-stage decoding procedure, we assign each rollout a mixed reward that considers both item-level accuracy and the diversity of recommendations induced by each reasoning trace. For each rollout $(\text{CoT}_{u,i}, \mathbf{c}_{u,i,j})$, the corresponding reward is defined as

$$R_{u,i,j} = R_{\text{rule}}(\mathbf{c}_{u,i,j}) \cdot R_{\text{div}}(\text{CoT}_{u,i}), \quad (7)$$

where R_{rule} provides the main accuracy signal based on the predicted itemic tokens, while R_{div} evaluates the diversity of the recommendations induced by the same reasoning trace. The rule-based reward evaluates whether the predicted itemic tokens hit the ground-truth set C_u^+ :

$$R_{\text{rule}}(\mathbf{c}_{u,i,j}) = \mathbf{I}[\mathbf{c}_{u,i,j} \in C_u^+]. \quad (8)$$

Recommendation aims to generate a set of items for each user, whose interests may span multiple categories. Therefore, a desirable reasoning trace should guide the model toward diverse types of recommended items. We encourage such diversity by computing an additional reward over the K itemic token sequences generated from the same reasoning trace. Specifically, we focus on the first digit of the itemic tokens, since the first sub-token largely determines the subsequent decoding process and captures coarse-grained item categories. Let $m_i^{(1)}$ denote the number of distinct first-digit values among $\{c_{u,i,j}^{(1)}\}_{j=1}^K$. The diversity factor is defined as

$$R_{\text{div}}(\text{CoT}_{u,i}) = \frac{\max(0, m_i^{(1)} - 1)}{K - 1}. \quad (9)$$

Encouraging diversity at the first digit promotes broader category-level coverage, which naturally leads to more diverse item predictions. The final reward $R_{u,i,j}$ thus assigns greater reward to trajectories that both match the target item and lead to better recommendation diversity. The diversity reward is defined at the reasoning-trace level, since all K itemic token sequences generated from the same reasoning trace share the same diversity factor R_{div} .

Optimization. GRPO uses clipping to constrain the importance ratio between the updated policy and the behavior policy, which stabilizes policy updates. For each generated token o_t , the importance ratio is defined as

$$r_t(\theta) = \frac{\pi_\theta(o_t | o_{<t})}{\pi_{\text{old}}(o_t | o_{<t})}. \quad (10)$$

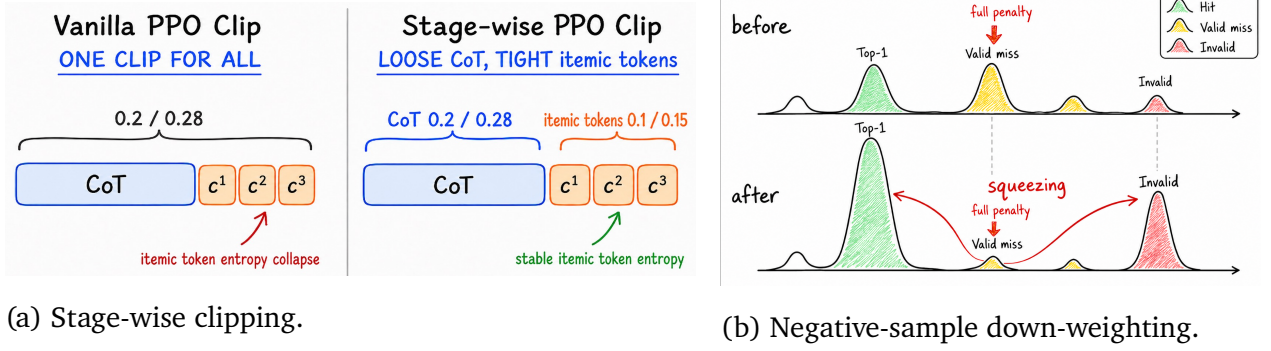


Figure 12 | Illustrations of two optimization-stabilizing strategies.

In reasoning-augmented recommendation, however, reasoning tokens and itemic tokens exhibit substantially different behaviors. Reasoning traces are long and mainly support exploration, whereas the final itemic token sequence is short but directly determines the recommendation reward. Using the same clip range for both stages can make the distribution of itemic tokens change too aggressively and cause entropy collapse. We therefore apply stage-wise clipping, with a looser clip range for reasoning tokens and a tighter clip range for itemic tokens, as illustrated in Figure 12 (a). Specifically, we use two stage-specific clipping ranges,

$$\epsilon_{\text{CoT}} = (\epsilon_{\text{CoT}}^-, \epsilon_{\text{CoT}}^+), \quad \epsilon_{\text{item}} = (\epsilon_{\text{item}}^-, \epsilon_{\text{item}}^+), \quad (11)$$

where ϵ_{CoT} is looser than ϵ_{item} . For each token o_t , the clipping range is selected according to its stage:

$$\bar{r}_t(\theta) = \text{clip}(r_t(\theta), 1 - \epsilon^-(t), 1 + \epsilon^+(t)), \quad \epsilon(t) = \begin{cases} \epsilon_{\text{CoT}}, & o_t \in \text{CoT}, \\ \epsilon_{\text{item}}, & o_t \in \mathbf{c}, \end{cases} \quad (12)$$

where $\epsilon(t) = (\epsilon^-(t), \epsilon^+(t))$. This preserves reasoning exploration while stabilizing updates on the final recommendation tokens.

To further stabilize GRPO under sparse recommendation rewards, a sample-level importance re-weighting scheme is applied to the actor loss, as illustrated in Figure 12 (b). Since only a small fraction of rollouts hit the ground-truth, most rollouts receive zero reward and contribute negative relative advantages. Without re-weighting, these numerous non-hit samples may dominate the aggregated gradient and push the policy toward overly conservative updates (Gao et al., 2025; Ren and Sutherland, 2025). After the group-relative advantage is computed, each rollout is assigned a scalar weight according to its hit status:

$$w_{u,i,j} = \begin{cases} 1.0, & \text{if } R_{\text{rule}}(\mathbf{c}_{u,i,j}) = 1, \\ \beta, & \text{otherwise,} \end{cases} \quad (13)$$

where $\beta < 1$ controls the contribution of non-hit rollouts. The weight is applied to the whole response-level actor loss of the rollout ($\text{CoT}_{u,i}, \mathbf{c}_{u,i,j}$), while the advantage estimation remains unchanged. This preserves the full gradient contribution from successful rollouts and down-weights non-hit rollouts, reducing the dominance of negative-advantage samples without changing the GRPO objective form.

Ablation Study. We first examine the effect of the two-stage rollout design by comparing GRPO with the full recommendation-oriented optimization against an ablated version that removes the two-stage rollout under the same effective rollout size. As shown in Figure 13, we evaluate both

training efficiency and recommendation performance. The blue curves correspond to GRPO with full optimization, while the orange curves correspond to the ablated variant. For performance, we evaluate the model on the Cross-Live Recommendation task using Recall@1 and Recall@8 under both thinking and non-thinking inference modes.

The two-stage rollout substantially reduces per-step training time by reusing each generated reasoning trace across multiple itemic token sequences, avoiding repeated generation of long reasoning prefixes. Meanwhile, the performance results show consistent gains, with particularly clear improvements under the thinking mode. These results indicate that expanding multiple recommendations from the same reasoning trace provides denser reward feedback for each reasoning path, enabling GRPO to more effectively optimize recommendation-oriented reasoning.

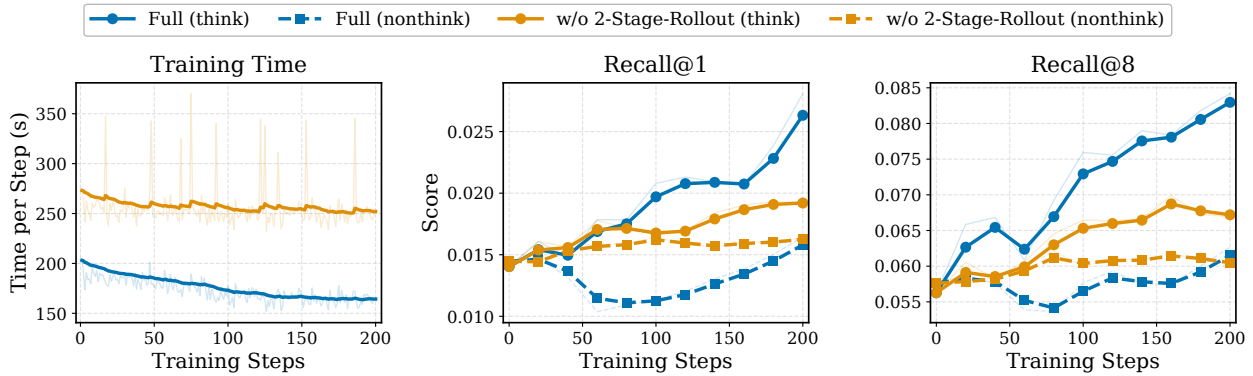


Figure 13 | Ablation study on rollout optimization. Two-stage rollout improves training efficiency while achieving stronger recommendation performance.

We further evaluate the effect of the diversity reward using the same ablation setting, where the ablated version removes R_{div} from the full reward design. The methods are evaluated on Cross-Live with Recall@1, Recall@8, and Recall@32 under both thinking and non-thinking inference modes. The blue curves denote the fully optimized method, and the orange curves denote the ablated version. Encouraging diversity among the subsequent itemic token sequences from the same reasoning trace helps the model retrieve a broader set of potential items. As a result, as shown in Figure 14, the diversity reward improves the effectiveness of reasoning, with more pronounced gains when Recall@K is evaluated at larger K.

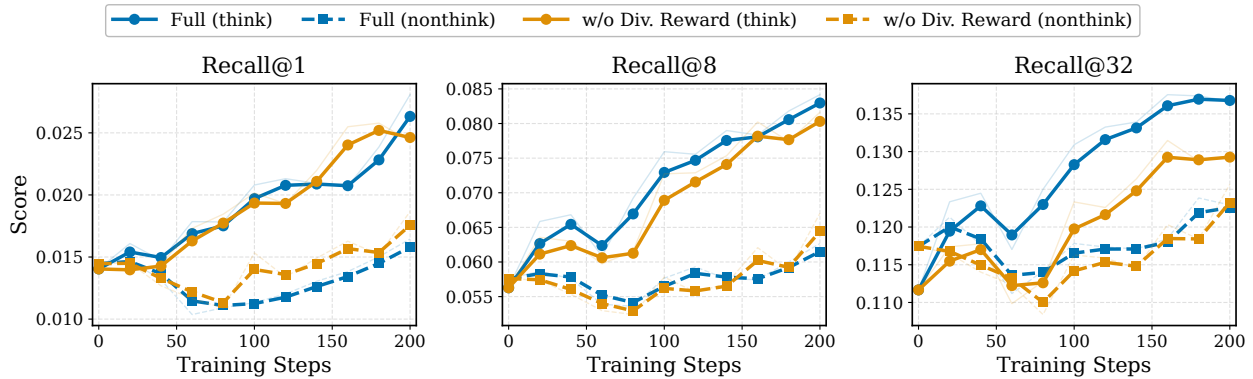


Figure 14 | Ablation study on reward optimization. Diversity reward strengthens thinking-based recommendation, with recall@K improvements more significant at larger K.

Finally, we analyze the two optimization-side stabilizers: stage-wise clipping and negative-sample down-weighting. For stage-wise clipping, the ablated version uses the same clipping range for reasoning tokens and itemic tokens. For negative-sample down-weighting, the ablated version removes the sample-level weight $w_{u,i,j}$. Figure 15 shows the ablation results of stage-wise clipping, and Figure 16 presents the ablation of negative-sample down-weighting. Both designs contribute to more stable training and better performance. Removing stage-wise clipping weakens the recommendation performance, especially the reasoning effectiveness on Recall@8 and Recall@32, indicating that reasoning tokens and itemic tokens benefit from different trust regions during policy optimization. Removing negative-sample down-weighting makes GRPO optimization less stable under sparse recommendation rewards, as abundant non-hit rollouts are assigned the same weight as successful ones. As shown in Figure 16, the ablated model suffers from a higher risk of training collapse, as reflected by a higher tendency to generate invalid itemic tokens and the resulting degradation in recommendation performance.

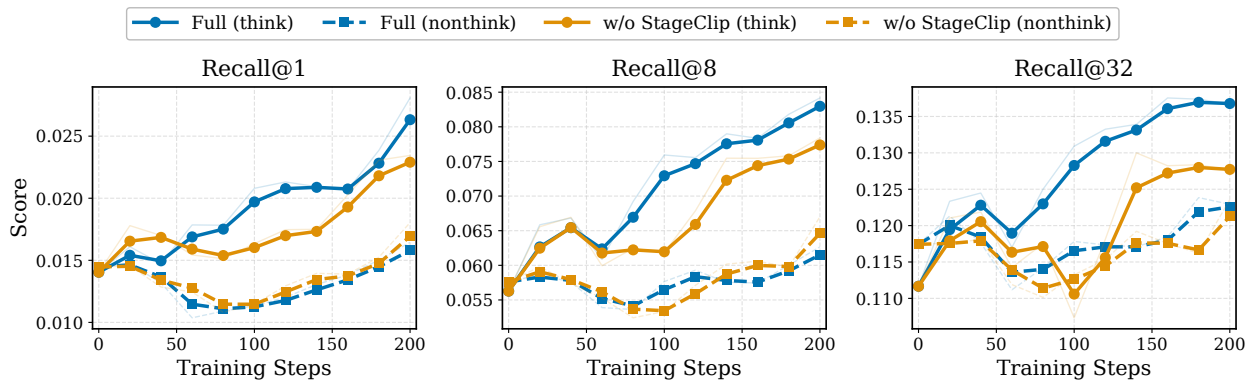


Figure 15 | Ablation study on optimization strategies. Stage-wise clipping consistently improves recommendation performance.

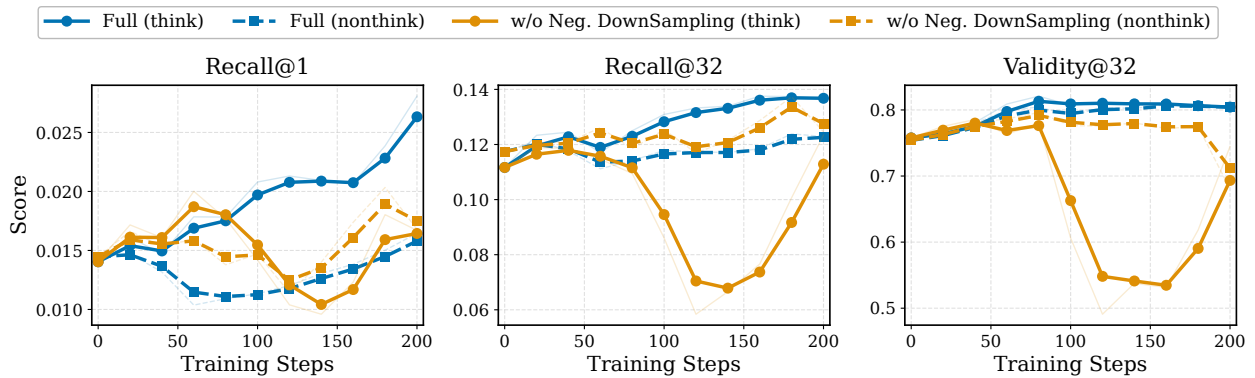


Figure 16 | Ablation study on optimization strategies. Negative-sample down-weighting helps stabilize RL training.

GRPO sharpens reasoning online but suffers from low sample efficiency on rare-but-correct trajectories. Rejection sampling fine-tuning (Section 6.2) is a natural offline complement: it harvests the small set of self-discovered high-quality rollouts and re-injects them as dense supervision, consolidating broadly useful behaviors that GRPO alone tends to forget between updates.

6.2. Rejection Sampling Fine-tuning

We first explore rejection sampling fine-tuning (RFT) to consolidate the knowledge of domain-specialized teacher models. RFT transfers domain-specific reasoning ability by selecting verified successful trajectories from each teacher and continuing supervised training on the aggregated trajectory set.

RFT workflow. While domain-specific GRPO improves the model through outcome-driven exploration, its policy optimization also sharpens the output distribution and makes the model concentrate on a small set of high-confidence items (Chen et al., 2026c). This can improve top-ranked predictions, but not fully aligned with the goal of recommendation, where the model is expected to predict a set of potential items that cover diverse user interests. As a result, RL optimization alone may provide limited improvement for the final recommendation task, especially when the model needs to retrieve a broader set of relevant items.

To complement GRPO, we introduce rejection sampling fine-tuning (RFT) as the consolidation method. Each domain-specialized GRPO model is used as a teacher to generate reasoning-augmented recommendation trajectories. For each user u , we retain trajectories whose predicted item matches the ground-truth set C_u^+ , filter out low-quality or inconsistent reasoning traces, and merge the remaining verified trajectories across the four domains. This yields an aggregated rejection-sampled dataset

$$\mathcal{D}_{\text{RFT}} = \{(x_u, \text{CoT}_u, \mathbf{c}_u)\}, \quad (14)$$

where x_u denotes the user context, CoT_u is the verified reasoning trace, and \mathbf{c}_u is the corresponding itemic tokens.

Before RFT, we first obtain a Mix-RL checkpoint by applying GRPO to mixed data covering R2 evolution and R3 recommendation tasks from all domains until convergence. Initialized from this checkpoint, the unified model is further trained on \mathcal{D}_{RFT} with the standard next-token prediction objective:

$$\mathcal{L}_{\text{RFT}} = -\mathbb{E}_{(x_u, \text{CoT}_u, \mathbf{c}_u) \sim \mathcal{D}_{\text{RFT}}} \sum_{t=1}^{|y_u|} \log \pi_{\theta}(y_{u,t} | x_u, y_{u,<t}), \quad y_u = [\text{CoT}_u; \mathbf{c}_u]. \quad (15)$$

In this way, RFT transfers domain-specific reasoning patterns discovered by RL into a single model while maintaining a smoother training objective based on next-token prediction.

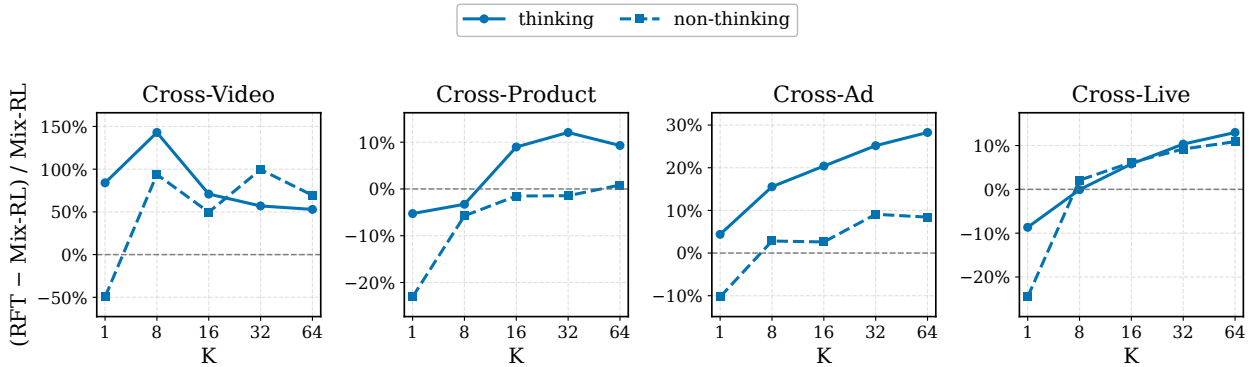


Figure 17 | Relative gains of RFT over Mix-RL across Recall@K. RFT generally shows larger improvements as K increases.

Performance Analysis. To analyze the effectiveness of RFT, we compare the performance of RFT against mixed-domain RL as the baseline method across four recommendation domains. We report the relative improvement of RFT over Mix-RL under Recall@ K , where $K \in \{1, 8, 16, 32, 64\}$. This evaluation investigates whether RFT improves only the top-ranked prediction or provides broader gains when the model is asked to retrieve a larger set of relevant items.

As shown in Figure 17, RFT does not always improve Recall@1, especially on Cross-Product and Cross-Live, where the relative improvements at small K are often negative or close to zero. However, as K increases, RFT consistently shows stronger advantages across domains. The improvement becomes particularly clear on most domains at larger K , achieving large positive gains across all evaluated cutoffs. This trend suggests that RFT is less focused on sharpening a single top prediction and more effective at improving candidate coverage. Such behavior is better aligned with the goal of recommendation, where the model should retrieve a set of relevant items that can cover diverse user interests rather than only optimize the first prediction.

6.3. Multi-Teacher On-Policy Distillation

This section introduces multi-teacher on-policy distillation (DeepSeek-AI, 2026; Dou et al., 2026; Xiao et al., 2026; Yang et al., 2026) to integrate specialized capabilities from four domains into a unified student model in an on-policy manner. We adopt a Monte Carlo RL formulation (Li et al., 2026), i.e., the log-probability ratio is evaluated *only* on the token sampled from the student itself at each step. This converts the optimization target into an expectation over sampled trajectories. The teacher is therefore only queried for a single-point log-probability at the sampled token, with no requirement to provide a full vocabulary distribution. The effectiveness of this RL-based formulation stems from the OneReason training pipeline, where the student model has already acquired foundational capabilities across all target domains. As a result, the student’s policy distribution does not exhibit significant deviation from the teacher’s policy. This distributional proximity naturally bounds the variance of the single-point log-probability evaluations. Therefore, we can circumvent the severe high-variance instability that typically plagues sparse RL updates, achieving significantly higher computational efficiency while maintaining stable training.

MOPD Objective. Formally, let μ_θ denote the student’s behavior policy deployed in the inference engine for trajectory generation, and π_θ denote the policy being actively optimized in the training engine. For a given prompt x , we first sample a complete rollout $y = (\mathbf{CoT}; \mathbf{c}) \sim \mu_\theta(\cdot | x)$. Based on the specific capability domain required by x , we route the prompt to the corresponding domain-expert teacher, denoted as π_{domain_i} . At each timestep t of the auto-regressive two-stage rollout, we define the token-level distillation advantage under a reverse-KL objective as:

$$\hat{A}_{\text{MOPD},t} = \text{sg} \left[\log \pi_{\text{domain}_i}(y_t | x, y_{<t}) - \log \pi_\theta(y_t | x, y_{<t}) \right], \quad (16)$$

Since the trajectories are generated by the behavior policy μ_θ but optimized under the target policy π_θ , we introduce truncated importance weighting $w_t(\theta)$ to correct for this off-policy discrepancy:

$$w_t(\theta) = \begin{cases} \text{sg} \left[\frac{\pi_\theta(y_t | x, y_{<t})}{\mu_\theta(y_t | x, y_{<t})} \right], & \epsilon_{\text{low}} \leq \frac{\pi_\theta(y_t | x, y_{<t})}{\mu_\theta(y_t | x, y_{<t})} \leq \epsilon_{\text{high}}, \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

where $\text{sg}[\cdot]$ denotes stop-gradient. We then define the surrogate loss of MOPD as:

$$\mathcal{L}_{\text{MOPD}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \mu_\theta(\cdot | x)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} w_t(\theta) \hat{A}_{\text{MOPD},t} \log \pi_\theta(y_t | x, y_{<t}) \right]. \quad (18)$$

Information-Gain-Aware Trajectory Filtering

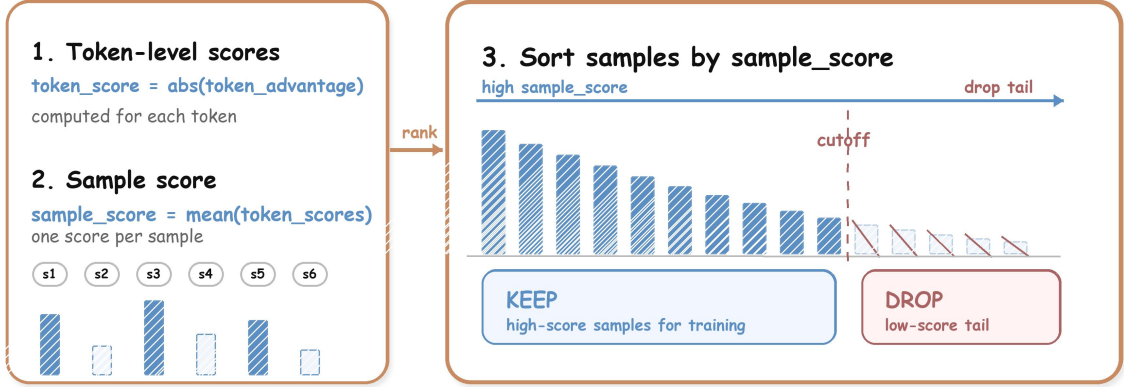


Figure 18 | Overview of the information-gain-aware trajectory filtering mechanism.

Optimization. Generative recommendation inherently struggles with massive item candidate spaces and highly sparse user-interest signals. In on-policy training, this translates to an exceptionally low hit rate during trajectory rollouts. When applying MOPD to such environments, treating all sampled rollouts uniformly leads to a severe gradient-dilution problem. Specifically, the student model quickly aligns with the teachers on generic, popular “head” items, resulting in near-zero token-level advantages ($|\hat{A}_{\text{MOPD},t}| \approx 0$). Consequently, the rare yet crucial domain-specific insights for “tail” items—where the specialized teachers’ expertise is most valuable (characterized by high $|\hat{A}_{\text{MOPD},t}|$)—are entirely overwhelmed by redundant gradients from already-aligned or noisy samples. To address this problem, we adopt an information-gain-aware trajectory filtering mechanism (Chen et al., 2026b), as shown in Figure 18. Mathematically, the absolute advantage $|\hat{A}_{\text{MOPD},t}|$ quantifies the instantaneous information gain provided by the teachers. We dynamically discard low-information trajectories, which forces the student to focus on the most dissimilar and informative rollouts. It is crucial for preventing policy performance degradation in environments with inherently low hit rates.

Specifically, we utilize the token-level advantage gap as a label-free indicator of information gain. To quantify the overall deviation from the teacher, we compute the token-average absolute gap for each trajectory y_j as its informativeness score $s(y_j)$:

$$s(y_j) = \frac{1}{T_i} \sum_{t=1}^{T_i} |\hat{A}_{\text{MOPD},j,t}|. \quad (19)$$

This absolute scoring naturally captures the bilateral correction demand: it identifies both instances where the student is overly confident in an incorrect path ($\hat{A}_{\text{MOPD},j,t} \ll 0$) and those where it fails to capture the teacher’s domain-specific insights ($\hat{A}_{\text{MOPD},j,t} \gg 0$).

Given a training batch \mathcal{B} and a target information-gain ratio ρ (e.g., $\rho = 0.8$), we sort all sampled rollouts in descending order based on their informativeness scores, yielding the sequence $x_{(1)}, x_{(2)}, \dots, x_{(|\mathcal{B}|)}$. To effectively filter out the redundant, already-aligned rollouts, we determine the smallest prefix length M that covers the target information-gain ratio ρ :

$$M = \min \left\{ m : \frac{\sum_{j=1}^m s(y_{(j)})}{\sum_{j=1}^{|\mathcal{B}|} s(y_{(j)})} \geq \rho \right\}. \quad (20)$$

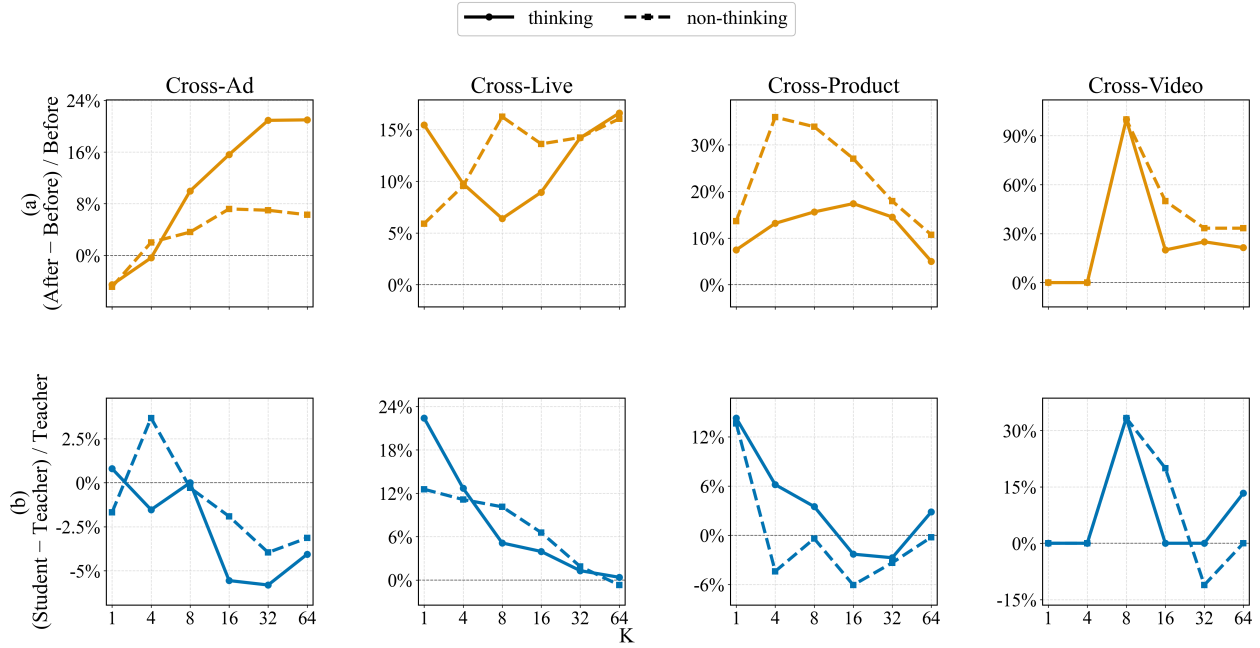


Figure 19 | Relative Recall@K changes of MOPD across four domains. (a) reports the improvement of the student before and after MOPD, computed as $(\text{After} - \text{Before}) / \text{Before}$. (b) reports the gap between the student after MOPD and the teacher, computed as $(\text{Student} - \text{Teacher}) / \text{Teacher}$. Positive values indicate higher Recall@K for the first term in each comparison. Solid and dashed lines denote thinking and non-thinking evaluation settings, respectively. Overall, MOPD consistently improves the student while keeping its performance close to the teacher across domains.

The active subset retained for optimization is then defined as $\mathcal{S} = \{y_{(1)}, \dots, y_{(M)}\}$. By discarding the low-information rollouts, we ensure that the optimization budget is concentrated exclusively on the most challenging and informative instances. The final MOPD objective is subsequently computed solely over this filtered subset \mathcal{S} .

Performance Analysis. As illustrated in Figure 19 (a), the proposed training paradigm yields strictly positive performance gains across all evaluated domains. Notably, the relative improvement expands as K increases in almost all domains. This demonstrates that MOPD effectively revitalizes the student model’s long-tail ranking capability within massive candidate spaces. Furthermore, the *thinking* mode exhibits a steeper growth trajectory at larger K compared with the *non-thinking* mode in the Ad and Live domains. This suggests that the explicit reasoning process significantly broadens the search boundaries, enabling the retrieval of hard-to-find tail items that intuition alone (non-thinking) might miss. Figure 19 (b) shows that the student aligns perfectly with or surpasses the teachers when $K \leq 16$, but a gap remains at $K > 16$. This remaining gap is a fundamental theoretical limit driven by the reverse-KL objective. Since the reverse-KL objective is inherently *mode-seeking*, it flawlessly captures the teacher’s primary probability peaks at $K \leq 16$. While our active filtering method maximizes the extraction of long-tail information from the teachers, perfectly replicating the entire distribution of the teachers remains challenging.

Ablation Study. We first analyze the effect of the information-gain-aware trajectory filtering mechanism. Figure 21 illustrates the policy gradient (PG) loss. MOPD without filtering (blue) exhibits larger variance and a steep initial drop, indicating unstable optimization and overfitting to redundant “head”

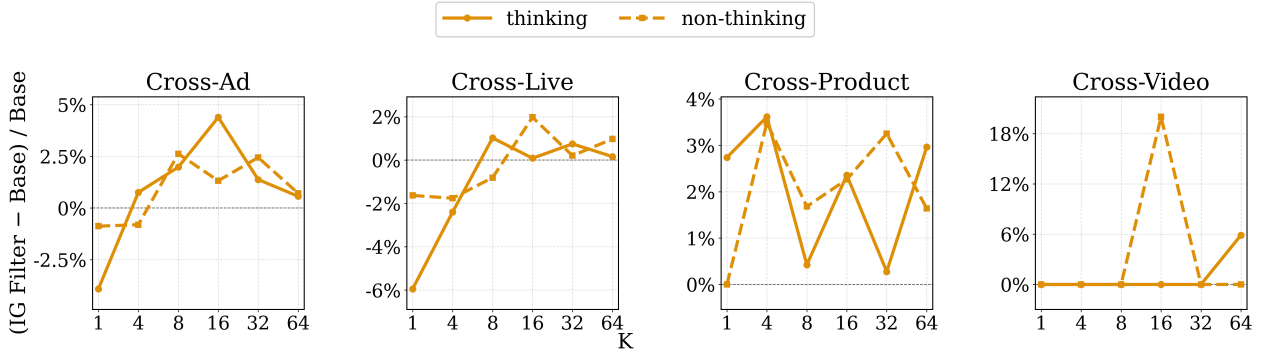


Figure 20 | Effect of the information-gain-aware (IG) filter on MOPD performance across recommendation domains. We report the relative Recall@K change between MOPD with IG filtering and the base MOPD setting, computed as (IG Filter – Base)/Base. Positive values indicate that IG filtering improves Recall@K over the base setting. Solid and dashed lines denote thinking and non-thinking evaluation settings, respectively. Overall, IG filtering brings consistent gains in most domains and evaluation settings, especially at larger values of K.

items. In contrast, MOPD with filtering (orange) significantly suppresses gradient noise, enabling a smoother and more stable convergence by exclusively focusing on high-information samples.

The relative Recall@K curves in Figure 20 provide a direct empirical validation of our core motivation, i.e., gradient budget reallocation based on trajectory-level information gain. The optimization budget is successfully concentrated on high-divergence, high-information samples ($s(y_j) \gg 0$). This focused gradient flow directly translates into the steep, upward performance trajectory starting at $K \geq 8$ and culminating in positive gains at $K = 64$. By reallocating the training focus from redundant easy items to challenging long-tail items, our filtering mechanism effectively mitigates the gradient dilution bottleneck, allowing the unified student to capture the specialized long-tail capabilities of the expert teachers under a highly constrained parameter budget.

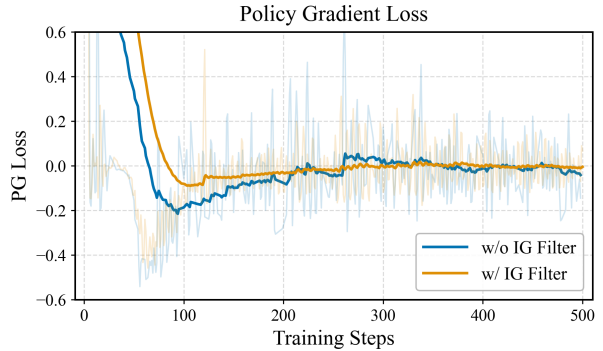


Figure 21 | Policy-gradient loss during MOPD training with and without IG filtering. The IG filter removes low-information trajectories based on token-level advantage statistics, leading to a smoother and more stable policy-gradient signal.

Further, we investigate the impact of different student selection, including SFT, SFT→RFT, and SFT→Mix-RL. As shown in Figure 22, all three students can benefit from MOPD training. When student and teacher distributions are roughly aligned, even with initial differences in student levels, MOPD can consistently improve student performance to a level near or even slightly surpassing that of the teachers. This indicates that the teacher’s upper bound limits the extent of improvement in MOPD. Specifically, even weaker students can catch up with the teacher, but it is difficult for them to surpass the teacher significantly.

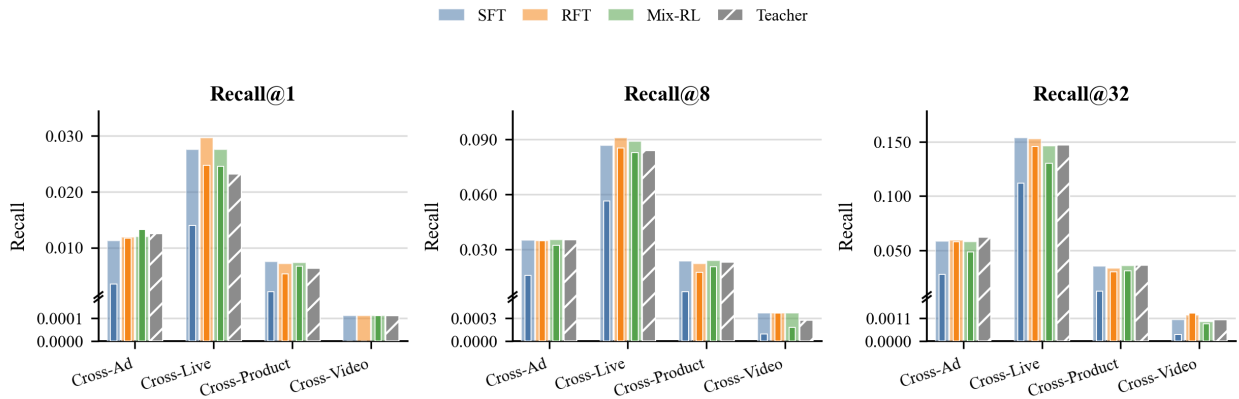


Figure 22 | Comparison of different student models before and after MOPD training. Dark bars denote the performance before MOPD, while light bars denote the performance after MOPD. This suggests that MOPD improvements are bounded by the teacher: weaker students can catch up, but rarely surpass it substantially.

6.4. Comparison and Discussions

In this section, we compare the recommendation accuracy of five optimization strategies: the SFT checkpoint, mixed-domain RL, domain-specific RL, RFT, and MOPD. The evaluation is conducted under the cross-domain recommendation setting, where we report Recall@K for the Cross-Video, Cross-Product, Cross-Ad, and Cross-Live tasks. This comparison allows us to examine whether the model learns better reasoning through self-exploration, whether the “specialize-then-unify” strategy alleviates cross-domain interference, and how different teacher-knowledge integration methods affect the final performance.

As shown in Table 9, all post-SFT optimization methods substantially improve over the SFT baseline on all domains and metrics, indicating that learning from self-exploration can further enhance recommendation performance beyond supervised imitation. However, directly mixing all domains during RL does not consistently achieve the best performance. Compared with domain-specific RL, mixed-domain RL generally yields weaker performance, suggesting that heterogeneous domains can introduce conflicting optimization signals. The two knowledge-integration methods then provide complementary benefits. RFT achieves strong and stable gains, especially on Cross-Video, showing that verified successful trajectories can effectively consolidate domain-specific reasoning patterns. MOPD performs competitively and obtains the best results in several settings, indicating that policy-level distillation can better preserve teacher preferences in some domains. Overall, these results support the effectiveness of the “specialize-then-unify” strategy: domain-specific RL first discovers effective reasoning behaviors, while subsequent knowledge integration transfers these domain-specialized capabilities into a unified model.

We further compare the impact of two integration paradigms, MOPD and RFT. We have several findings:

- MOPD and RFT exhibit domain-specific strengths. MOPD achieves superior performance in Product and Live domains, even significantly outperforming the Live teacher. Conversely, RFT consistently yields marginal and stable advantages over MOPD, particularly at larger K values in Video and Ad domains.
- RFT consistently guarantees that the thinking mode outperforms the non-thinking mode across

Table 9 | Performance comparison (in %) of different optimization strategies. Both RFT and MOPD effectively integrate domain-specific knowledge and outperform Mix-RL across different domains.

Domain	Metric	Mode	SFT	Mix-RL	Single-RL	RFT	MOPD
Cross-Video	Recall@1	thinking	0.00	<u>0.01</u>	0.01	0.01	0.01
		non-thinking	0.01	<u>0.01</u>	0.01	0.01	0.01
	Recall@8	thinking	0.01	0.02	0.03	0.05	<u>0.04</u>
		non-thinking	0.02	0.02	0.03	0.04	<u>0.03</u>
	Recall@16	thinking	0.02	0.05	0.06	0.08	<u>0.06</u>
		non-thinking	0.03	0.04	0.05	0.06	<u>0.05</u>
	Recall@32	thinking	0.03	0.08	<u>0.10</u>	0.13	0.10
		non-thinking	0.06	0.06	<u>0.09</u>	0.12	0.08
	Recall@64	thinking	0.06	0.14	0.15	0.24	<u>0.18</u>
		non-thinking	0.11	0.12	0.16	0.19	<u>0.16</u>
Cross-Product	Recall@1	thinking	0.21	0.67	0.63	0.67	<u>0.67</u>
		non-thinking	0.26	0.66	<u>0.66</u>	0.55	0.76
	Recall@8	thinking	0.69	2.05	<u>2.29</u>	2.06	2.36
		non-thinking	1.14	1.83	2.46	1.82	<u>2.42</u>
	Recall@16	thinking	0.94	2.53	3.04	2.83	<u>2.98</u>
		non-thinking	1.69	2.44	3.30	2.63	<u>3.05</u>
	Recall@32	thinking	1.22	3.10	3.65	3.54	<u>3.62</u>
		non-thinking	2.32	3.18	3.88	3.24	<u>3.80</u>
	Recall@64	thinking	1.65	3.79	3.87	4.19	<u>4.11</u>
		non-thinking	2.96	3.83	<u>4.25</u>	3.96	4.34
Cross-Ad	Recall@1	thinking	0.35	<u>1.32</u>	1.25	1.34	1.23
		non-thinking	1.07	1.23	<u>1.19</u>	1.10	1.14
	Recall@8	thinking	1.56	3.21	3.53	3.67	<u>3.67</u>
		non-thinking	3.02	3.33	<u>3.46</u>	3.41	3.53
	Recall@16	thinking	2.11	3.97	4.86	<u>4.80</u>	4.76
		non-thinking	4.16	4.31	4.71	4.46	<u>4.58</u>
	Recall@32	thinking	2.78	4.83	6.20	<u>6.14</u>	5.94
		non-thinking	5.37	5.44	6.06	<u>5.84</u>	5.84
	Recall@64	thinking	3.41	5.86	<u>7.39</u>	7.50	7.19
		non-thinking	6.49	6.68	7.33	<u>7.26</u>	7.13
Cross-Live	Recall@1	thinking	1.40	2.46	2.32	<u>2.53</u>	2.67
		non-thinking	1.45	<u>2.37</u>	2.23	1.96	2.41
	Recall@8	thinking	5.63	8.29	8.39	<u>8.70</u>	8.82
		non-thinking	5.76	7.39	7.80	<u>7.82</u>	8.58
	Recall@16	thinking	8.18	10.63	11.14	11.49	<u>11.46</u>
		non-thinking	8.34	9.99	<u>10.65</u>	10.63	11.18
	Recall@32	thinking	11.17	13.03	<u>14.69</u>	14.63	14.82
		non-thinking	11.74	12.86	<u>14.42</u>	14.25	14.71
	Recall@64	thinking	14.32	16.03	<u>18.62</u>	18.35	18.89
		non-thinking	15.52	16.09	18.80	18.17	<u>18.66</u>

all domains and K values. In contrast, the improvements of MOPD in the thinking and non-thinking modes are highly synchronized, with the non-thinking mode achieving competitive gains.

This divergence is driven by the extreme reward sparsity of recommendation tasks and the distinct sampling dynamics of the two paradigms:

- **RFT:** In a massive item candidate space, successful reasoning paths are extremely sparse. RFT bypasses this exploration bottleneck offline by executing multiple rollouts and strictly retaining only those that yield correct outcomes. This rejection-sampling process distills only “golden” reasoning paths in which the CoT is logically aligned with the correct recommendation, thereby securing a stable reasoning benefit (*thinking* > *non-thinking*).
- **MOPD:** Conversely, MOPD relies on on-policy distillation, continuously aligning the student’s generated distributions with the teachers across all rollouts (regardless of whether the CoT perfectly hits the target). Because the student processes these rollouts within a shared parameter space, this continuous distillation calibrates the foundational prompt-to-item semantic embeddings holistically. As a result, the nonthinking capability itself is massively enhanced. Therefore, MOPD produces a synchronized uplift: it improves the *thinking* capability while simultaneously and profoundly strengthening the underlying “intuition” (*non-thinking*). However, successful CoTs are extremely sparse. Without strict off-policy filtering of correct CoT, the on-policy process inevitably absorbs noisy or ungrounded CoT steps from the weak priors. This noise might prevent the thinking mode from fully unlocking its theoretical superiority to consistently eclipse the massively enhanced non-thinking baseline.

7. CoT Analysis Indicators

Beyond target accuracy, the practical utility of recommendation reasoning traces depends on whether the CoT genuinely contributes to the final itemic pattern prediction, remains faithful to the observed user history, respects the structural constraints of the itemic token space, and follows a coherent intent-evolution trajectory. Recent evidence suggests that free-form reasoning in itemic token-based recommendation foundation models may introduce a general-subspace prior, where textual inertia dominates inference and dilutes critical itemic pattern evidence (Zhang et al., 2026b). Motivated by this observation, we design a four-criterion diagnostic framework to evaluate the quality of generated CoT traces. We organise them along two orthogonal axes that together cover the full diagnostic surface of a recommendation reasoning trace. The four criteria populate this 2×2 matrix exactly once each (Table 10).

- **Symbolic vs. Probabilistic** (*what kind of check?*). A *symbolic* check inspects the discrete set of itemic pattern referenced inside the CoT against a static reference set; it requires only string parsing and set operations, runs without a model forward pass and without access to y_{GT} . A *probabilistic* check inspects how the CoT shifts the conditional log-likelihood of the target y_{GT} ; it requires both a model forward pass and the ground-truth label, and is therefore strictly costlier but directly measures predictive effect.
- **Local vs. Global** (*against which reference scope?*). A *local* check compares against a narrow, instance-bounded reference (the current user’s history, or the immediately preceding reasoning prefix). A *global* check compares against a wide, corpus-bounded or trace-bounded reference (the full item catalog, or the entire CoT versus the no-CoT baseline).

Table 10 | The four-criterion diagnostic framework organised on a 2×2 matrix: *Symbolic vs. Probabilistic* (kind of check) \times *Local vs. Global* (reference scope). Within each row, the local cell strictly strengthens the global cell (Property P1). Across rows, symbolic checks are cheap/label-free while probabilistic checks are forward-pass/label-bound (Property P2). The four cells are jointly necessary; each guards against a distinct failure mode.

	Global (corpus-/trace-bounded)	Local (instance-bounded)
Probabilistic	C1: ΔLL $\log p(y_{GT} \mathbf{x}, c) - \log p(y_{GT} \mathbf{x})$	C2: ℓ_t progression $\ell_t - \ell_{t-1}$
Symbolic	C3: γ_{legal} $\mathcal{S}(c) \cap \mathcal{V}_{\text{item}}$	C4: $\gamma_{\text{hist} \text{legal}}$ $(\mathcal{S}(c) \cap \mathcal{V}_{\text{item}}) \cap \mathcal{S}(\mathbf{x}_{\text{hist}})$

Criterion 1: CoT Likelihood Gain. We first quantify whether the generated CoT contributes positively to the prediction of the ground-truth target. Inspired by the analysis of CoT-induced distributional shift in (Zhang et al., 2026b), we define the CoT likelihood gain as:

$$\Delta LL = \log p(y_{GT} | \mathbf{x}, c) - \log p(y_{GT} | \mathbf{x}), \quad (21)$$

where \mathbf{x} denotes the input context, including the user profile and interaction history, c denotes the generated CoT, and y_{GT} is the ground-truth target itemic pattern. A positive ΔLL indicates that conditioning on the CoT increases the model’s likelihood of the correct target, whereas a negative value suggests that the CoT distracts the model from the history-grounded itemic pattern evidence.

As shown in Figure 23, for each model-domain pair, we rigorously compute ΔLL by contrasting the target log-likelihood under the thinking mode, where prediction is conditioned on the generated CoT, with that under the non-thinking mode, where the target is predicted directly from the input context without CoT. We then report the average per-sample ΔLL values of the SFT and RFT models across four domains. The SFT model consistently yields negative ΔLL values in all domains, indicating that its generated CoT tends to distract the model from the ground-truth target prediction. In contrast, the RFT model achieves positive ΔLL values across all four domains, demonstrating that the RFT stage significantly improves the ability of the generated CoT to contribute positively to the prediction of the ground-truth target.

Criterion 2: Log-Likelihood Progression Along the CoT. A high-quality reasoning trace should progressively strengthen the model’s confidence in the ground-truth target as the reasoning process unfolds. To characterize this intent-evolution behavior, we examine how the target likelihood changes along the CoT.

To obtain intermediate reasoning units, we use DeepSeek (DeepSeek-AI, 2026) to segment the original CoT into a sequence of semantically coherent reasoning segments:

$$c = \{c_1, c_2, \dots, c_T\}. \quad (22)$$

For each prefix of the segmented CoT, we compute the conditional log-likelihood of the ground-truth target:

$$\ell_t = \log p(y_{GT} | \mathbf{x}, c_1, \dots, c_t), \quad t = 1, \dots, T. \quad (23)$$

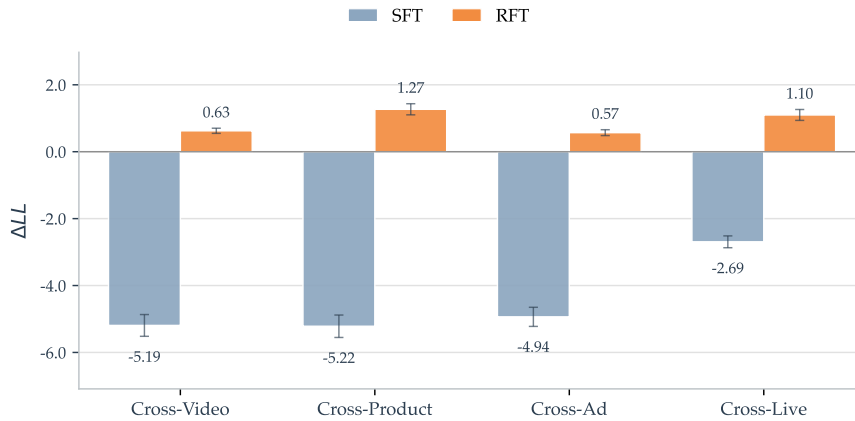


Figure 23 | For each model-domain pair, ΔLL is computed as the difference between the log-likelihood of the ground-truth target under thinking mode and that under non-thinking mode. Positive values indicate that the generated CoT increases the model’s confidence in the ground-truth target, whereas negative values suggest that the CoT distracts the prediction process. Across all four domains, the SFT model exhibits consistently negative ΔLL , while the RFT model achieves positive ΔLL , demonstrating that RFT substantially enhances the effectiveness and utility of thinking reasoning relative to the non-thinking setting.

As shown in Figure 24, in the thinking mode setting, progressively incorporating segmented CoT prefixes consistently improves the conditional log-likelihood of the ground-truth target for both the SFT and RFT models. Compared with the thinking mode prediction without CoT, the likelihood trajectories of all four domains rise substantially as more reasoning segments are added, indicating that the generated CoT provides incremental predictive evidence rather than serving as a merely post-hoc explanation. This trend is observed in both optimization stages: the SFT model already benefits from CoT prefixes, while the RFT model further exhibits stronger and more stable likelihood improvements across domains. Although a few domains show minor fluctuations at later prefixes, the overall upward progression demonstrates that, under the thinking mode inference setting, the current models can effectively exploit their generated CoT to strengthen confidence in the ground-truth target.

While both SFT and RFT benefit from progressively incorporating CoT prefixes, RFT exhibits a much earlier saturation behavior, reaching near-optimal likelihood after only a few reasoning segments. This suggests that not all generated reasoning tokens contribute equally to prediction, and that a substantial portion of the predictive gain can be achieved with a shortened CoT, leaving room for future exploration of reasoning compression or adaptive early-stopping mechanisms.

Criterion 3: Item Legality. Every itemic pattern referenced in the CoT should also correspond to a valid identifier in the item corpus. Let $\mathcal{V}_{\text{item}}$ denote the complete item catalog. We define the item legality ratio as:

$$\gamma_{\text{legal}} = \frac{|\mathcal{S}(c) \cap \mathcal{V}_{\text{item}}|}{|\mathcal{S}(c)|}, \quad (24)$$

Invalid itemic patterns indicate that the reasoning trace contains structurally illegal item identifiers, which may introduce noise into subsequent itemic token prediction. As shown in Table 11, γ_{legal} already saturates at 1.00 for both SFT and RFT across all four domains, providing a clean basis on which the conditional history-grounding measure of Criterion 4 can be evaluated.

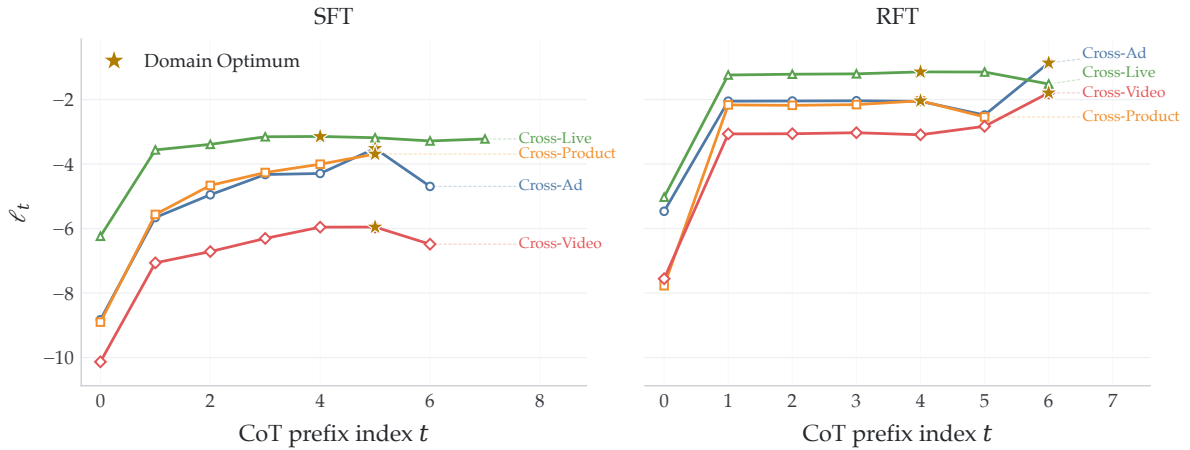


Figure 24 | ℓ_t denotes the conditional log-likelihood of the ground-truth target given the input context and the first t reasoning segments. CoT prefix index $t=0$ denote the thinking mode prediction without CoT, while brown stars indicate the peak likelihood achieved along the reasoning trajectory for each domain. Across all domains, progressively incorporating CoT prefixes consistently improves the likelihood relative to the no-CoT thinking mode, demonstrating that the generated CoT contributes useful predictive information rather than serving as a purely post-hoc explanation. Moreover, compared with SFT, the RFT model typically reaches its optimal likelihood at substantially earlier reasoning prefixes, suggesting that predictive evidence is concentrated into a smaller portion of the reasoning process and revealing potential opportunities for CoT compression and adaptive early stopping.

Criterion 4: History Item Reference Validity (conditioned on item legality). Since our prompt explicitly constrains the CoT to reason only based on the observed user history, every *legal* itemic pattern mentioned in the CoT should further correspond to an item that actually appears in the user’s interaction history. Differently from a raw history-grounding ratio over $\mathcal{S}(c)$, we condition the measurement on item legality being satisfied: we first restrict the referenced set to its legal subset $\mathcal{S}(c) \cap \mathcal{V}_{\text{item}}$, and then measure the fraction of this subset that is supported by the user’s history $\mathcal{S}(\mathbf{x}_{\text{hist}})$. This explicitly disentangles the orthogonal failure mode of *history-violating hallucination* from the upstream failure of *illegal-identifier generation* captured by Criterion 3. Formally,

$$\gamma_{\text{hist}|\text{legal}} = \frac{|(\mathcal{S}(c) \cap \mathcal{V}_{\text{item}}) \cap \mathcal{S}(\mathbf{x}_{\text{hist}})|}{|\mathcal{S}(c) \cap \mathcal{V}_{\text{item}}|}, \quad (25)$$

where $\mathcal{S}(c)$ denotes the set of itemic patterns extracted from the CoT and $\mathcal{S}(\mathbf{x}_{\text{hist}})$ the set of itemic patterns appearing in the interaction history. A low $\gamma_{\text{hist}|\text{legal}}$ indicates that, even after illegal references have been filtered out, the CoT still introduces itemic patterns unsupported by the observed history, i.e. prompt-violating reasoning that hallucinates beyond the user’s actual behaviour.

Qualitative Illustration: SFT vs. RFT on the Same User. To complement the aggregate diagnostics in Criterion 1–Criterion 4, we present a representative user for whom the SFT and RFT models produce markedly different reasoning traces over the *same* input context. The user is an 18–23 year-old male student whose recent interaction history is dominated by tactical-shooter livestreams, smartphone reviews, and snack purchases; the ground-truth next item is a video about *equipment in Delta Force*, a recently released tactical-shooter title that the user has only *clicked once via an in-feed advertisement*—i.e., a target that requires *cross-IP generalisation* rather than mere historical replay. Table 12 and Table 13 present the two CoT traces verbatim, organised along the same PERSONA ABSTRACTION → INTEREST EXPANSION → TRANSITION INFERENCE structure for direct comparison.

Table 11 | Cross-domain γ_{legal} and $\gamma_{\text{hist}|\text{legal}}$ of itemic pattern references in generated CoTs (pid_pass@32). All numbers are in %; best per metric-domain pair is in **bold**. **Takeaways.** (i) γ_{legal} saturates at 100% for both SFT and RFT across all four domains, indicating that legality is no longer a bottleneck after RFT. (ii) On Cross-Video and Cross-Ad, RFT tightens $\gamma_{\text{hist}|\text{legal}}$ over SFT (+2.50 pt and +4.27 pt respectively). (iii) On Cross-Live and Cross-Product, RFT instead exhibits a moderate drop in $\gamma_{\text{hist}|\text{legal}}$ (-1.18 pt and -14.59 pt). This drop is consistent with Table 9, where the relative gain of RFT over Mix-RL on Live and Product is also smaller than that on Video and Ad, suggesting that the symbolic CoT diagnostics are well aligned with downstream recommendation metrics. Tightening history grounding on Live and Product is left as future optimisation.

Metric	Method	Video	Ad	Live	Product
$\gamma_{\text{legal}} \uparrow$	SFT	100.00	100.00	100.00	100.00
	RFT	100.00	100.00	100.00	100.00
$\gamma_{\text{hist} \text{legal}} \uparrow$	SFT	97.50	94.93	100.00	97.92
	RFT	100.00	99.20	98.82	83.33

Table 12 | SFT recommendation CoT case: interest stays at the surface IP level (*Peace Elite*) and the final transition lands on the wrong IP family.

Persona Abstraction

用户是一名18–23岁的男性中学生，有伴侣，从历史行为来看，他属于典型的“游戏直播观众”与“Z世代潮流青年”的混合体。最直接的证据是：他近期大量观看、评论和打赏《和平精英》、《王者荣耀》等射击与MOBA类游戏直播如<|living_begin|><a_1659><b_7342><c_6278>、<|living_begin|><a_3713><b_4342><c_6278>，并频繁点击和长播相关游戏攻略视频如<|video_begin|><a_1985><b_581><c_1414>；同时，他对数码产品（尤其是手机）表现出浓厚兴趣，多次点击手机评测与对比内容如<|video_begin|><a_7328><b_7546><c_698>，并购买过动态壁纸等个性化商品如<|prod_begin|><a_2877><b_4537><c_6971>。因此，他的核心兴趣高度集中在战术竞技类游戏（特别是《和平精英》及其衍生模式“地铁逃生”）、高性价比数码产品（尤其是手机）以及由此衍生的游戏装备交易/账号服务上。

Analysis: this first segment is persona abstraction: it compresses the raw history into a user state and dominant preference factors, rather than expanding future hypotheses.

Interest Expansion

基于此进行可能性发散：**A.** 用户最可能继续深入关注《和平精英》的游戏内容更新或装备获取技巧，因为这是其长期稳定投入的核心领域；**B.** 由他对高性能手机的兴趣延伸至最新款机型的评测或促销信息；**C.** 其对短剧类型[复仇_女性成长]等有一定消费记录，在碎片时间可能被类似强情节的娱乐内容吸引。

Analysis: the SFT trace stays at the surface IP layer—every branch is named by an existing IP the user has already heavily consumed (*Peace Elite*, smartphone reviews, short dramas), with no abstraction to a latent category axis.

Transition Inference

权衡来看，**A**的可能性最大且最直接。用户近期互动密集围绕《和平精英》，包括在<|ad_begin|><a_5687><b_6727><c_2789>点击该游戏广告、长播多个赛季更新视频如<|video_begin|><a_5632><b_1206><c_6057>，并首次打赏了提供装备交易服务的游戏主播<|living_begin|><a_1790><b_1523><c_7951>。这些行为表明他正处于对该游戏操作细节和资源获取的高度活跃期。**B**方向虽然相关，但更偏向于间歇性消费需求；**C**方向则与其主要互动轨迹关联较弱。因此，后续交互极有可能仍是围绕《绝地求生》或《和平精英》这类战术竞技游戏的具体玩法展示、新赛季活动介绍或高水平玩家操作集锦展开。

Target: 《三角洲行动》装备讲解/装备玩法视频 (cross-IP target, only one prior ad-click exposure in history).

Analysis: the conclusion is an in-IP replay of the user’s strongest historical signal. The correct genre is captured, but the target IP family is missed—consistent with the negative ΔLL regime of SFT.

In contrast, the RFT model, faced with the *identical* input context, produces a reasoning trace that opens with a catalogue-grounded quantitative prior and lifts the candidate space from a single IP to a latent category axis, as shown in Table 13. The critical gain comes from R1-derived interest evolution: instead of replaying *Peace Elite* or *PUBG* as surface names, RFT bridges them to the tactical-shooter category and then to *Delta Force* through a weak but history-supported ad-click signal.

Linking the case to the four criteria. This single example concretely instantiates all four quantitative findings. The SFT trace lands on the wrong IP family and would lower the target likelihood, whereas the RFT trace commits to *Delta Force* early in its candidate enumeration and merely defends the choice afterwards—mirroring the positive ΔLL in Criterion 1 and the earlier ℓ_t saturation in Criterion 2. Moreover, the RFT trace anchors its reasoning to a catalogue-grounded statistic ($[GAMING]=69.57\%$) rather than free-form qualifiers, while both traces reference only history-supported items, consistent with the higher γ_{legal} (Criterion 3) and saturated γ_{hist} (Criterion 4) of RFT in Table 11.

Table 13 | RFT recommendation CoT case: reasoning is anchored to a catalogue-grounded quantitative prior and generalises across IPs, landing on the ground-truth title (*Delta Force*).

Persona Abstraction

用户是一名18–23岁的男性中学生，属于“游戏直播观众”和“二次元爱好者”的混合体，但游戏兴趣远大于二次元。证据包括：他关注的博主类型以【游戏】占69.57%为主；历史记录中大量出现《王者荣耀》、《和平精英》等电竞内容及主播互动；购买商品也多为电竞周边、手机配件和零食。具体来说，他最深度的兴趣集中在《王者荣耀》赛事/攻略、射击类手游（如和平精英）以及性价比数码产品上。

Analysis: the trace opens with a *quantised, catalogue-grounded statistic* (69.57%) rather than free-form qualifiers, so the persona prior is already grounded in catalogue evidence.

Interest Expansion

从近期行为看，用户最可能关心的兴趣点有以下几个方向：**A.** 《绝地求生》/战术竞技类游戏的新玩法或装备，因为他在`<|video_begin|><a_1985><b_3589><c_2398>` 评论了PUBG相关视频，并在`<|ad_begin|><a_1481><b_2918><c_7187>` 点击了三角洲行动的游戏广告；**B.** iQOO手机系列的最新动态与性能对比，这是他长期且稳定的兴趣，在`<|video_begin|><a_447><b_4491><c_889>` 点赞了iQOO帧率测试视频，并多次点击手机促销广告；**C.** 辣条等怀旧零食的测评与优惠信息，他在`<|prod_begin|><a_8138><b_1801><c_208>` 等多次点击辣条商品，并在`<|video_begin|><a_801><b_4733><c_6583>` 收藏了关于食品价格揭秘的内容。

Analysis: unlike SFT, which remains around *PUBG/Peace Elite*, the RFT trace lifts the candidate axis to “tactical shooter” and, through R1-style relation evolution, keeps *Delta Force* as a cross-IP successor supported by the weak ad-click evidence.

Transition Inference

综合权衡来看，**A**的可能性最大且最为即时。用户对射击游戏的热情不仅限于《和平精英》，已从《和平精英》延伸到同类战术竞技玩法（如绝地求生意外关联），而“三角洲行动”作为热门新游恰好承接了他的这一需求。**B**是他的基础盘会持续存在，但近期没有新的手机发布节点刺激其产生强烈的探索性消费或内容互动峰值；**C**属于间歇性消费需求，在完成几次点击后短期内重复推荐同类商品转化率会下降。

Target: 《三角洲行动》装备讲解/装备玩法视频 (cross-IP target, only *one* prior ad-click exposure in history).

Analysis: the early commitment to “Delta Force / tactical shooter” in the expansion step is merely *defended* here, mirroring the earlier ℓ_t saturation of RFT; the cross-IP target is reached because R1-style relation evolution operates over a category axis, not a single IP name.

Crucially, the difference between the two traces is not lexical polish but *abstraction level*: SFT reasons over surface IP names (*Peace Elite* → more *Peace Elite*), whereas RFT reasons over a latent category axis (“tactical shooter”) and is therefore able to bridge from observed evidence (*Peace Elite*, *PUBG* comment, *Delta Force* ad click) to a target IP that has only been *weakly* exposed in history. This category-level generalisation is what the 2×2 diagnostic framework in Table 10 is designed to detect, and it is what RFT systematically delivers.

8. Experiments

8.1. Performance Comparison

Baselines. We compare OneReason with representative baselines on four cross-domain recommendation tasks. We group the evaluated models by how they represent items:

- **ID-Based:** Items are represented by conventional item IDs. This group includes standard sequential recommendation models, namely SASRec (Kang and McAuley, 2018) and HSTU (Zhai et al., 2024).
- **Text-Based:** Items are represented by natural-language item dense captions. This group evaluates general-purpose LLMs under a zero-shot protocol on all tasks, with thinking mode enabled where supported. For recommendation, the LLMs generate a potentially relevant item caption, which is then encoded by Qwen3-Embedding-8B (Zhang et al., 2025b) along with the candidate item captions, and recommendations are obtained through approximate nearest-neighbor (ANN) retrieval.
- **Itemic Token-Based:** Items are represented by itemic tokens. In addition to OneReason variants, we evaluate TIGER (Rajput et al., 2023) and LC-Rec (Zheng et al., 2024b). For LC-Rec, we report three variants: LC-Rec-SFT-Only-8B, LC-Rec-SFT-Only-14B, and LC-Rec-PT-SFT-8B (initialized from Qwen3-8B, Qwen3-14B (Qwen3 et al., 2025), and our OneReason pre-training checkpoint). Following the original LC-Rec setting, its SFT data covers only recommendation and caption tasks.

Table 14 | Baselines on cross-domain recommendation (in %). Best results are **bolded**; second-best results are underlined.

Category	Model	Cross-Video		Cross-Product		Cross-Ad		Cross-Live	
		Pass@64	Recall@64*	Pass@64	Recall@64	Pass@64	Recall@64	Pass@64	Recall@64
ID-Based	SASRec	0.03	0.01	0.31	0.25	1.04	0.37	1.76	0.40
	HSTU	0.10	0.01	0.32	0.24	2.79	0.78	2.32	2.14
Text-Based	Qwen3-8B	0.05	0.01	0.15	0.12	0.48	0.09	2.10	1.85
	Qwen3-32B	0.33	0.03	0.84	0.63	1.21	0.30	5.64	5.10
	Qwen3-235B-A22B	0.24	0.02	0.64	0.49	0.77	0.19	5.10	4.66
	Deepseek-V3.2	0.11	0.01	0.38	0.31	0.62	0.13	3.46	3.12
	Claude-Opus-4.6	0.14	0.01	0.23	0.17	0.50	0.11	3.02	2.66
	Gemini-3-Preview	0.29	0.03	0.74	0.59	1.22	0.27	3.92	3.44
	GPT-4o-mini	0.19	0.02	0.73	0.55	1.21	0.28	4.01	3.57
	GPT-5.4	0.24	0.02	1.43	1.15	1.64	0.43	7.20	6.38
Itemic Token-Based	TIGER	0.88	0.07	0.21	0.17	7.65	2.39	2.32	1.78
	LC-Rec-SFT-Only-8B	0.22	0.02	0.06	0.05	2.83	0.67	0.89	0.71
	LC-Rec-SFT-Only-14B	0.20	0.01	1.03	0.73	5.99	1.94	3.76	3.09
	LC-Rec-PT-SFT-8B	1.49	0.13	3.95	3.00	15.85	6.55	19.32	16.70
	OneReason SFT non-thinking	1.33	0.11	3.94	2.96	15.73	6.49	18.05	15.52
	OneReason SFT thinking	0.71	0.06	2.18	1.65	9.16	3.41	16.43	14.32
	OneReason RFT non-thinking	<u>2.08</u>	<u>0.19</u>	<u>5.20</u>	<u>3.96</u>	<u>17.56</u>	<u>7.26</u>	<u>21.01</u>	<u>18.17</u>
	OneReason RFT thinking	2.41	0.24	5.47	4.19	17.78	7.50	21.10	18.35

* Video Recall@64 is notably low due to the largest target set (avg. 13.92 items) making full recall harder.

Recommendation Performance. Based on the evaluation results in Table 14, we highlight the primary reasons for the limited performance of ID-based and text-based baselines:

- **Cold-Start Sensitivity:** Traditional ID-based models struggle in cross-domain evaluation because 33.69% of target item IDs are unseen during training, causing severe cold-start issues; by

contrast, only 11.55% of target itemic patterns are unseen, as reported in Appendix B.2.

- **Limits of Text-Based LLM Recommendation:** More advanced general-purpose LLMs do not necessarily produce better recommendations, suggesting that recommendation ability is not reliably correlated with general intelligence or model scale. Their underperformance stems from the absence of collaborative signals and the errors introduced by ANN-based caption-to-item retrieval. In contrast, OneReason avoids this additional retrieval step by directly decoding itemic tokens. Pointwise scoring or better retrieval may yield stronger recommendations; we leave their study to the community.
- **Effect of OneReason Pre-Training:** The LC-Rec results show that our OneReason pre-training checkpoint provides a strong semantic foundation for itemic-token recommendation. Compared with the LC-Rec-SFT-Only variants, LC-Rec-PT-SFT-8B achieves a clear performance jump across all four domains, demonstrating the effectiveness of our OneReason pre-training strategy.

Non-Recommendation Performance. Table 15 evaluates R0–R2 capabilities on OneReason-Bench. Two findings stand out:

- **RFT thinking helps higher-level reasoning.** RFT thinking lags non-thinking on R0, consistent with perception-task overthinking (Jiang et al., 2025a), but improves average R1–R2 performance. Together with the recommendation gains in Table 14, this suggests that our RL improves recommendation while strengthening higher-level reasoning.
- **OneReason makes compact itemic tokens competitive.** Text-based LLMs are given richer inputs, whereas OneReason uses compact itemic tokens. Even under this input-information disadvantage, OneReason surpasses several larger advanced LLMs on parts of the R2 suite. This suggests that OneReason can perform efficient and effective long-sequence user modeling.

Table 15 | Performance comparison on R0–R2 tasks in OneReason-Bench (in %). For R0 tasks, results are macro-averaged over all domains. The grounding task is reported by Pass@64. For the R2 tasks evaluated by action-logic score, matching for text-based models uses 10-gram soft matching instead of exact matching. Best results are **bolded**; second-best results are underlined.

Category	Model	R0: Perception			R1: Derivation	R2: Evolution		
		Item Und.	Ground.	QA	I2I	Select.	Topic Gen.	Direct Gen.
Text-Based	Qwen3-8B	-	-	-	-	40.70	25.49	8.60
	Qwen3-32B	-	-	-	-	51.96	28.05	7.73
	Deepseek-V3.2	-	-	-	-	<u>57.18</u>	27.13	11.32
	Claude-Opus-4.6	-	-	-	-	56.84	17.16	13.46
	Gemini-3-Preview	-	-	-	-	56.83	33.68	14.76
	GPT-5.4	-	-	-	-	58.92	41.41	17.61
Itemic Token-Based	LC-Rec-SFT-Only-8B	22.98	0.00	0.40	3.43	0.00	0.00	0.00
	LC-Rec-SFT-Only-14B	26.48	0.00	56.45	16.21	0.00	0.00	0.00
	LC-Rec-PT-SFT-8B	35.41	<u>5.21</u>	63.90	25.54	3.32	8.60	4.46
	OneReason SFT non-thinking	<u>36.84</u>	3.95	<u>66.55</u>	<u>28.36</u>	35.07	33.87	15.42
	OneReason SFT thinking	36.91	1.06	64.60	23.88	32.18	31.60	14.31
	OneReason RFT non-thinking	36.82	5.24	67.25	23.99	38.92	39.33	<u>20.31</u>
	OneReason RFT thinking	36.78	1.35	65.65	28.60	42.42	<u>39.57</u>	21.23

General-Intelligence Sanity Check. Table 16 shows that in thinking mode, OneReason preserves its Qwen3-8B backbone’s general reasoning and instruction-following abilities on these four representative benchmarks after recommendation-oriented training, providing preliminary evidence that

recommendation-oriented training does not catastrophically degrade general capabilities. In contrast, LC-Rec variants suffer substantial degradation across all benchmarks.

Table 16 | General benchmark performance comparison (in %). Best results are **bolded**; second-best results are underlined.

Category	Model	MMLU-Pro	GPQA-Diamond	MATH-500	GSM8K
Text-Based	Qwen3-8B non-thinking	66.32	39.90	83.80	93.03
	Qwen3-8B thinking	72.35	56.06	95.20	95.68
Itemic Token-Based	LC-Rec-SFT-Only-8B	9.73	17.17	41.80	14.03
	LC-Rec-SFT-Only-14B	45.35	37.37	60.20	52.54
	LC-Rec-PT-SFT-8B	39.72	35.86	81.00	51.55
	OneReason SFT non-thinking	61.74	33.33	91.40	93.86
	OneReason SFT thinking	71.01	51.52	95.60	<u>95.00</u>
	OneReason RFT non-thinking	59.57	37.88	91.40	94.31
	OneReason RFT thinking	<u>72.08</u>	<u>54.04</u>	<u>95.40</u>	94.69

8.2. Non-Thinking Gains from Thinking Supervision

Although the recommendation CoT data in Section 5.4.1 is constructed to provide compact, evidence-grounded recommendation reasoning rather than to directly optimize non-thinking decoding, we observe an important downstream phenomenon: thinking supervision can improve direct recommendation even when explicit reasoning traces are suppressed at inference time. We study this phenomenon by comparing CoT and unCoT recommendation supervision under controlled training budgets. For clarity, we use **unCoT** to denote recommendation samples without explicit reasoning traces, where the model directly predicts the target itemic tokens from the user context. In contrast, **CoT** samples contain a trace before the final target, generated by the three-stage compression-reasoning protocol in Section 5.4.1: persona abstraction, interest expansion, and transition inference.

We first conduct a controlled token-aligned experiment. We compare two settings: (i) continued SFT on 100K unCoT samples; and (ii) continued SFT on a mixture of 40K CoT samples and 50K unCoT samples. Since CoT samples are longer than unCoT samples, the latter two SFT settings are aligned to the same training token budget of 0.25B tokens. All models are evaluated using the same non-thinking decoding format, where the model directly predicts target itemic tokens without generating a reasoning trace.

Table 17 | Token-aligned comparison of non-thinking recommendation performance (in %). The 100K unCoT setting and the 40K CoT + 50K unCoT setting use the same training budget of 0.25B tokens. For each domain-metric pair, the better setting is highlighted in bold.

Metric	Setting	Cross-Video	Cross-Product	Cross-Ad	Cross-Live
Pass@32	100K unCoT	0.88	3.16	13.81	13.79
	CoT + unCoT	1.17	3.62	13.57	15.99
Recall@32	100K unCoT	0.08	2.33	5.57	11.93
	CoT + unCoT	0.12	2.74	5.50	13.98
Pass@64	100K unCoT	1.64	4.38	16.08	18.12
	CoT + unCoT	1.95	4.86	15.84	20.32
Recall@64	100K unCoT	0.18	3.33	6.72	15.59
	CoT + unCoT	0.21	3.67	6.68	17.69

Table 17 shows that, under the same token budget, replacing part of the unCoT data with CoT data improves non-thinking performance on Cross-Video, Cross-Product, and Cross-Live. For example, the mixed setting improves Pass@64 from 1.64% to 1.95% on Cross-Video, from 4.38% to 4.86% on Cross-Product, and from 18.12% to 20.32% on Cross-Live. Cross-Ad is the main exception, where the mixed setting is slightly worse than unCoT-only training. This suggests that the benefit of CoT supervision is not uniform across domains, and motivates a finer-grained analysis of the CoT/unCoT mixture ratio.

We then conduct a sample-count-controlled mixture sweep. Under a fixed budget of 100K recommendation samples, we vary the ratio between CoT and unCoT samples. Let α denote the unCoT ratio:

$$\alpha = \frac{N_{\text{unCoT}}}{N_{\text{CoT}} + N_{\text{unCoT}}}, \quad 1 - \alpha = \frac{N_{\text{CoT}}}{N_{\text{CoT}} + N_{\text{unCoT}}}. \quad (26)$$

For each mixture ratio, we train a model from the same starting checkpoint and evaluate it with non-thinking decoding. Figure 25 reports the relative Pass@32 gain over the 100% unCoT baseline, defined as:

$$G_d(\alpha) = \frac{\text{Pass@32}_d(\theta_\alpha) - \text{Pass@32}_d(\theta_{1.0})}{\text{Pass@32}_d(\theta_{1.0})}, \quad (27)$$

where d denotes the target domain, θ_α is the model trained with unCoT ratio α , and $\theta_{1.0}$ is the model trained with 100% unCoT data.

Figure 25 shows that the response curve is not simply monotonic. Instead, most domains exhibit a clear intermediate optimum. Cross-Video reaches its best result near a balanced mixture, Cross-Product prefers a more CoT-heavy mixture, and Cross-Live also benefits from a balanced or moderately CoT-rich mixture. Cross-Ad, in contrast, shows a much flatter curve, and its best point lies closer to the unCoT side. These results support the conclusion from the token-aligned experiment: CoT supervision can improve non-thinking inference in most domains, but the best mixture depends on the domain.

We interpret this phenomenon as a balance between two types of supervision. unCoT data matches the non-thinking inference format and directly teaches the model to map user context to target item tokens. CoT data provides an additional training signal with two components: a *compression* signal, which teaches persona and trajectory abstraction from long behavior histories, and a *reasoning* signal,

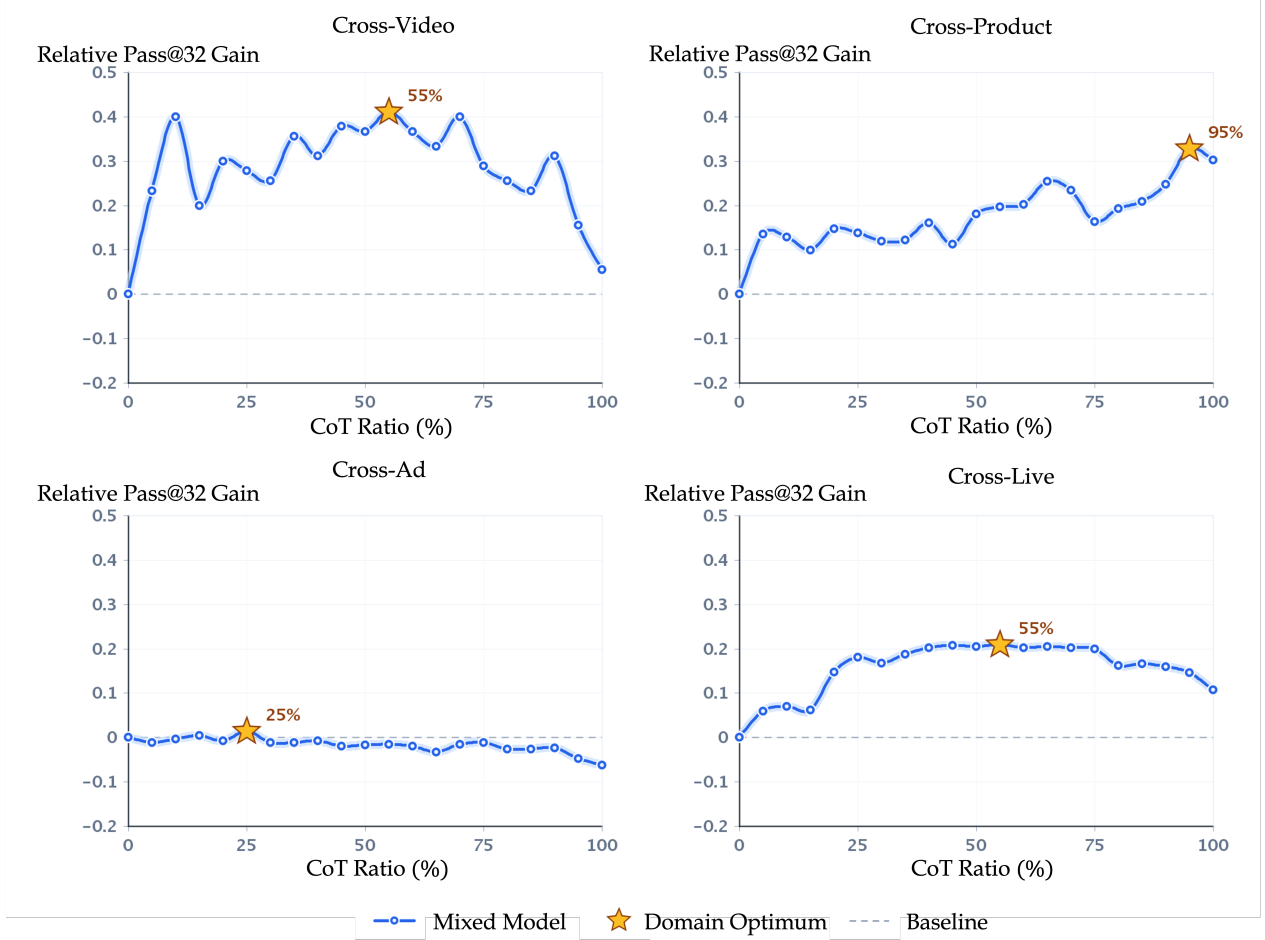


Figure 25 | Non-thinking recommendation gains across different CoT/unCoT mixture ratios. All models are evaluated with non-thinking decoding. The horizontal axis denotes the CoT ratio ($1 - \alpha$), and the vertical axis reports the relative Pass@32 gain over the 100% unCoT baseline. Stars mark the best mixture ratio for each domain.

which teaches transition judgement from the compressed state and supporting evidence to target-domain decisions. During non-thinking decoding, the model does not explicitly generate the trace. A plausible hypothesis is that part of the compact user-state construction and evidence-selection bias learned from CoT supervision is still reflected in the model parameters. The current experiment supports this hypothesis at the behavioral level, but it does not identify the exact contribution of compression versus reasoning.

However, more CoT data is not always better. CoT traces are longer than unCoT answers, so excessive CoT training may weaken the density of final-target supervision. CoT traces may also contain noise or over-explanation, and the training format may become less aligned with non-thinking decoding. Therefore, the best mixture is obtained when the reasoning signal from CoT and the direct target signal from unCoT are properly balanced.

Formally, we use the following conceptual decomposition to describe the non-thinking gain in domain d under unCoT ratio α :

$$G_d(\alpha) = B_d + A_d^{\text{uncot}}(\alpha) + I_d^{\text{cot}}(1 - \alpha) - C_d^{\text{trace}}(1 - \alpha) - C_d^{\text{format}}(\alpha, 1 - \alpha), \quad (28)$$

where B_d denotes the base capability in domain d , A_d^{uncot} denotes the direct-answer alignment benefit

from unCoT data, I_d^{cot} denotes the signal absorbed from CoT data, C_d^{trace} denotes the cost of noisy or over-long traces, and C_d^{format} denotes the mismatch between CoT-format training and non-thinking decoding. Conceptually, I_d^{cot} contains both a compression component, which teaches the model to form a compact user-interest state, and a reasoning component, which teaches the model to connect this state to target-domain decisions. We use this as a conceptual decomposition only, not as an identifiable additive estimate from the current experiments. The empirical optimum $\hat{\alpha}_d^*$ in Figure 25 can be understood as the point where the marginal benefit of adding more unCoT data balances the marginal loss of reducing CoT supervision:

$$\left. \frac{\partial G_d(\alpha)}{\partial \alpha} \right|_{\alpha=\hat{\alpha}_d^*} \approx 0. \quad (29)$$

This equation is only a conceptual explanation and does not assume that the entire curve follows a quadratic law. It simply states that the best mixture appears when direct-answer alignment and CoT-based supervision reach a domain-specific balance.

One hypothesis for the domain difference is that different tasks rely on different evidence sources. Cross-Video and Cross-Live may benefit from CoT because their user histories often contain diverse signals and multiple plausible intent directions, making evidence organization useful before prediction. Cross-Product may be more sensitive to CoT-heavy training because purchase intent often has to be inferred from content cues and behavior transitions. By contrast, Cross-Ad may depend more on short-term conversion signals and business exposure patterns that are only partially captured by natural-language reasoning traces, which could explain why it favors unCoT-heavy training and shows a flatter curve.

Overall, these experiments reveal a useful but domain-dependent effect of thinking supervision on non-thinking recommendation. High-quality thinking supervision can improve direct non-thinking recommendation in most domains under appropriate mixture ratios, but Cross-Ad shows that CoT is not universally beneficial and should not be used more aggressively by default. This makes CoT data valuable as part of a potential data flywheel for recommendation foundation models: CoT construction can strengthen the backbone, the improved non-thinking model can produce better candidates and filtering signals, and explicit thinking-mode inference can still be reserved for harder cohorts where trace generation is worth the extra latency.

9. Deployment

In the rapidly growing local-services advertising scenario of the Kuaishou App, we have successfully deployed OneReason. Under strict latency, strong baseline, and item-deliverability constraints, the system achieves stable online deployment and delivers the **one of the most significant business uplifts in the Kuaishou local-services advertising scenario**. This section focuses on the online deployment of OneReason, including Industrial Scenario Adaptation and Online Deployment Architecture, as well as a set of reusable methodologies distilled from the OneReason deployment.

9.1. Industrial Scenario Adaptation

Given a strong and high-performing OneReason foundation model, efficiently adapting it to a specific industrial scenario becomes a critical challenge. Notably, the Kuaishou life-service advertising scenario is not included in the OneReason training data. To address this, we propose a standardized and reusable continuous training pipeline that enables rapid domain adaptation and can be easily generalized to other industrial application scenarios. Further details are provided in Appendix A.1.

9.2. Industrial Scenario Online Deployment

Due to the scale of LLMs and the limitations of inference systems, directly deploying OneReason for real-time recommendation still faces significant latency and computational cost challenges. In this section, we propose a new online architecture for OneReason, the “Fast-Slow Thinking” architecture, and further introduce an online incremental training strategy to ensure model performance.

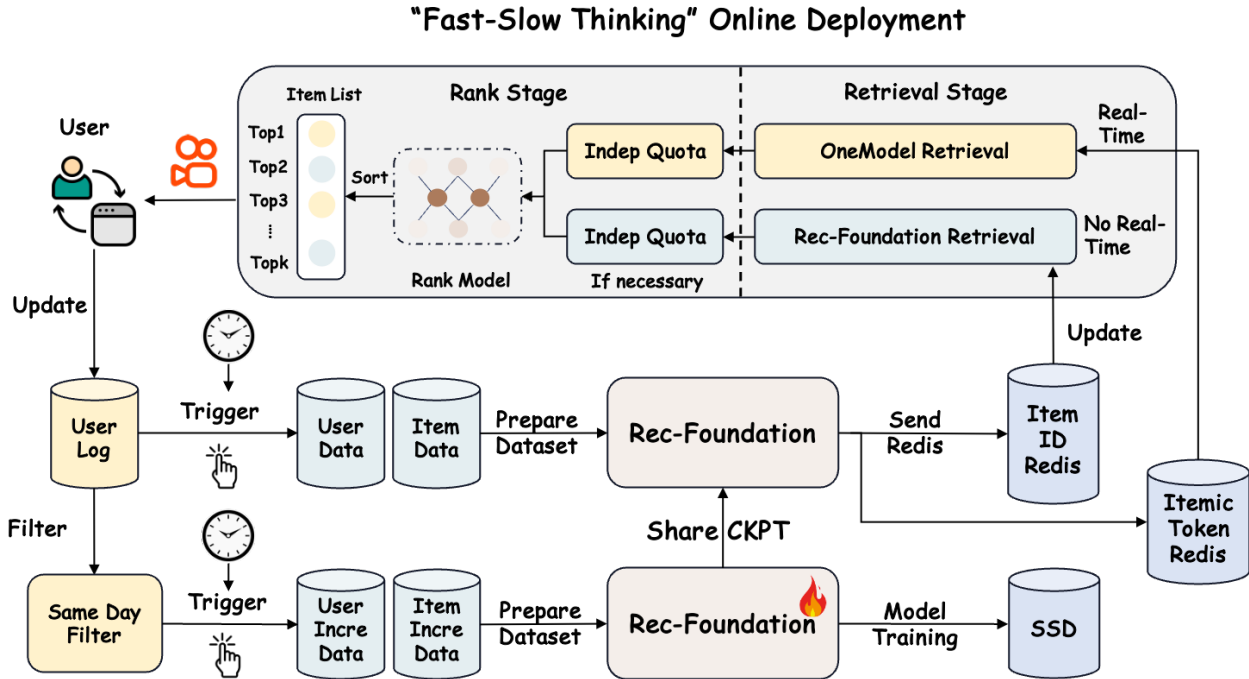


Figure 26 | The online deployment architecture of OneReason.

“Fast-Slow Thinking” Architecture in the Online Deployment. This architecture avoids fully relying on real-time inference and instead introduces a Fast (Online) and Slow (Nearline) collaboration mechanism, continuously injecting OneReason’s knowledge and reasoning capabilities into the system while maintaining low latency and stable gains. The workflow is shown in Figure 26.

- **Step 1: Pipeline Design.** We build a decoupled nearline retrieval pipeline that does not participate in early-stage competition. Instead, it is integrated into the downstream ranking model for joint scoring with the real-time OneRec outputs. When OneReason retrieval results are unavailable, the system automatically falls back to the original OneRec pipeline to ensure stability and coverage.
- **Step 2: Dataset Trigger.** We periodically (daily or hourly) aggregate distributed data jobs to construct user behavior, user profile, and item content data. An online pipeline is then triggered for data fetching, cleaning, and feature processing, forming an end-to-end data flywheel.
- **Step 3: Offline Inference.** After dataset preparation, we load the latest OneReason checkpoint to predict the user’s most likely next itemic tokens offline. These tokens are then decoded into item IDs using our specialized decoding strategy for OneReason (as discussed in Appendix A.4.2).
- **Step 4: Online Serving.** The decoded item IDs are written to Redis via offline inference jobs, forming a candidate pool for online serving. During inference, the nearline OneReason results and real-time OneRec retrieval results are jointly fed into the ranking model for unified fusion, enabling a Fast-Slow Thinking recommendation system.

Online Incremental Training Strategy. The online platform continuously introduces new items,

while user interests and intents also evolve over time. Therefore, an incremental training mechanism is required to adapt OneReason to long-term business evolution while preserving its existing capabilities and ensuring stable online performance. The procedure is as follows:

- **Pre-Training Incremental Training.** We conduct periodic continual pre-training on newly introduced items within fixed time windows to keep pace with the latest content distribution. To prevent catastrophic forgetting and maintain the general reasoning capabilities of OneReason, we jointly train on a mixture of newly collected data and a sampled general-domain corpus.
- **SFT Incremental Training.** We construct daily incremental training data from user logs, using same-day user interactions as supervision signals for OneReason training, which models short-term interest dynamics and improves sensitivity to preference shifts.

Two Application Paradigms of OneReason. We design two application paradigms of OneReason under the “Fast-Slow Thinking” architecture in the online deployment, both yielding significant online gains and demonstrating the strong modeling and generalization capabilities of OneReason.

- **Slow Pipeline: OneReason.** This method directly employs OneReason for retrieval. Specifically, OneReason leverages user profile information and historical interaction sequences to predict the next itemic tokens that the user is most likely to engage with, which are then decoded into a candidate item list serving as the retrieval output of the Slow pipeline.
- **Fast Pipeline: OneReason for OneRec.** We leverage the outputs of OneReason to enhance recommendations, where the most relevant itemic tokens are transformed into embedding representations. In OneRec, we introduce a dedicated *Thinking Token* to incorporate these signals, which serves to distill knowledge from OneReason into the online OneRec within the Fast pipeline. The overall methodology is provided in Appendix A.2.

9.3. Online Experiment

We conducted a 10-day online A/B experiment by deploying OneReason in a subpart of the local-services scenario of the Kuaishou App, using a 5% traffic allocation for both the treatment and control groups. Users in the treatment group were served by the “Fast-Slow Thinking” architecture of OneReason, whereas users in the control group remained on the existing recommendation system.

Table 18 shows that both directly using OneReason for retrieval and using it to enhance the real-time OneRec yield significant improvements in the online A/B experiment. The direct application in the slow pipeline improves Revenue, while enhancing OneRec in the fast pipeline further boosts performance. Combining both paradigms achieves the best results, with substantial gains in Revenue and Impressions, corresponding to hundreds of millions of RMB in annualized commercial revenue for the Kuaishou platform. Moreover, we compute the ratio between computational resource consumption and revenue, demonstrating that OneReason achieves an ROI > 5. More experimental analysis and discussion are provided in Appendix A.3 and A.4.

10. Related Works

With the rapid advancement of artificial intelligence, leveraging AI to reshape industrial systems has become a major direction in both industry and academia. Recommendation systems have increasingly adopted LLMs to improve user experience and business performance. Recent generative recommendation frameworks, such as the OneRec series, demonstrate strong modeling capacity, scalability, and system efficiency. However, most existing methods largely remain within a System-1 single-pass

Model	Impressions	Revenue
OneReason	+0.940%	+4.528%
OneReason for OneRec	+6.831%	+4.636%
Combined	+10.332%	+8.234%

Table 18 | Online performance comparison across different deployment models. **Combined**: joint deployment of OneReason and OneReason for OneRec. **Revenue** refers to the total commercial revenue generated in the online advertising system of the Kuaishou platform.

inference paradigm. Inspired by reasoning-oriented foundation models such as OpenAI O1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025), researchers are beginning to explore a **Think-then-Generate** paradigm. As recommendation environments become increasingly complex and dynamic, incorporating structured reasoning into generative foundation models (GFMs) has emerged as a promising research direction. In this section, we provide an overview of recommendation foundation models and reasoning-enhanced generative frameworks. Building upon these developments, we illustrate how the OneReason **Think-then-Recommendation** paradigm differs from existing GFMs.

Following the rise of LLMs, researchers began exploring their direct application to recommendation tasks, leveraging their rich world knowledge and strong instruction-following capabilities. Typical approaches formulate user profiles and interaction histories as prompts, allowing LLMs to reason about user preferences before generating recommendation results. Representative works include GPT4Rec (Li et al., 2023), Chat-Rec (Gao et al., 2023), LLMRank (Hoffmann et al., 2024), and LLM as Zero-Shot Ranker (Hou et al., 2024b). However, as research on LLM4Rec advances, it has become clear that general-purpose LLMs struggle to fully capture the characteristics of recommendation tasks, leading to performance bottlenecks. This has motivated two main directions for integrating LLMs with recommendation systems. The first follows an LLM-backbone paradigm to enhance existing recommendation models, represented by the OneRec series. The second aims to train recommendation foundation models that can directly perform recommendation.

In the first direction, TIGER (Rajput et al., 2023) is an early industrial work on generative recommendation, introducing RQ-VAE-based itemic tokens and framing recommendation as an end-to-end generation task from user sequence itemic patterns to next-item representations. HSTU (Zhai et al., 2024) further replaces traditional multi-stage pipelines with a Transformer-based backbone and explores scaling laws for generative recommendation. The OneRec series (Deng et al., 2025; Liu et al., 2025; Zhou et al., 2025d; Wang et al., 2026; Zhang et al., 2026a; Zhou et al., 2026) then engineers this paradigm into a unified pre-train and post-train framework, demonstrating strong modeling and efficiency advantages. Subsequent work focuses on issues such as diversity control (Agarwal et al., 2025), end-to-end pipelines (He et al., 2025; Sun et al., 2026a), ultra-long user sequence modeling (Li et al., 2025), and industrial scenario constraints (Lin et al., 2025; Xue et al., 2026; Zhang et al., 2026c).

In the second direction, research focuses on building recommendation foundation models by adapting LLMs to recommendation data, leveraging their world knowledge and reasoning ability. LC-Rec (Zheng et al., 2024a) shows that LLMs can understand recommendation semantics via itemic ID alignment and instruction tuning. RecGPT (Yi et al., 2025b) and RecGPT-V2 (Yi et al., 2025a) further use LLMs for user intent modeling in retrieval and interest mining, with small-scale online validation. However, these methods mainly focus on user intent modeling and do not achieve full end-to-end decision-making. OxygenREC (Hao et al., 2025) follows a similar idea to RecGPT, using fast-slow thinking to generate contextual instructions for downstream models. Nevertheless, these approaches remain LLM-augmented recommendation systems rather than true end-to-end generative foundation

models. In contrast, our goal is to build an industrially deployable GFM paradigm that enables direct reasoning-based recommendation decisions with measurable online impact. OpenOneRec (Zhou et al., 2026) is an important precursor, proposing a novel **Think-then-Recommendation** paradigm. They also propose RecIF-Bench, large-scale datasets, and 1.7B/8B foundation models with multiple training strategies, including Text Alignment, Co-Pretraining, Multi-task SFT, and Rec-RL. However, it remains unclear whether its reasoning consistently translates into real recommendation performance gains, leaving room for further investigation. The most directly comparable diagnostic work is (Zhang et al., 2026b), which formally measures the non-thinking/thinking gap on OpenOneRec and attributes it to textual inertia from the general subspace: itemic token embeddings and general-text embeddings share latent dimensions but are misaligned, so a longer CoT lets the general-text prior dominate and dilute ID-grounded evidence. In OneReason, we address this issue at training time across elaborate training pipelines to guide more robust itemic perception and cognition CoT.

In traditional generative recommendation models, researchers have early on incorporated the reasoning capability of LLMs by introducing explicit or implicit reasoning mechanisms. Since generative recommendation operates within the recommendation space and lacks an explicit CoT paradigm, most existing methods adopt latent reasoning, injecting reasoning signals from LLM outputs or internal system signals. For example, ReaRec (Tang et al., 2026) introduces multi-step reasoning into generative recommendation, iteratively feeding high-dimensional reasoning results back to refine user representations. OnePiece (Dai et al., 2025) further extends this idea by integrating LLM-style context engineering and reasoning into industrial recommendation systems via a reasoning-block design, enabling progressive alignment between intermediate states and recommendation logic. In contrast, OneSearch-V2 (Chen et al., 2026a) leverages LLMs with CoT to deeply understand complex user queries and distills such reasoning capability into recommendation model parameters via self-distillation. Meanwhile, REG4Rec (Xing et al., 2025) addresses reasoning diversity and reliability by constructing multiple item-level semantic representations to enable multi-path reasoning and introducing self-reflection to prune inconsistent reasoning paths.

11. Conclusion & Future Work

In this work, we propose OneReason, a reasoning foundation model that connects the large language model and the recommendation system into one single model, enabling recommendation with wisdom. Specifically, we conduct extensive modifications: (1) in the pre-training stage, we utilize a content-understanding task rather than contrastive supervision to obtain discrete itemic tokens, and use 578B tokens to align the itemic-token and text-token semantic spaces; (2) in SFT, we design a standardized coarse-to-fine, logically coherent CoT structure to ensure the quality of the thinking trace; and (3) in RL, we conduct a specialize-then-unify recipe to improve thinking-mode ability while balancing performance across multiple services.

Equipped with these techniques, to our knowledge, OneReason is the first work in which the thinking mode consistently outperforms the non-thinking mode on downstream recommendation benchmarks, suggesting that reasoning can be translated into real recommendation gains. More interestingly, we also find that replacing ordinary unCoT data with thinking CoT data improves non-thinking inference in several domains. This suggests that some benefits of CoT supervision can transfer to direct decoding, but a causal separation between compression and reasoning requires further ablation. Furthermore, we propose OneReason-Bench, which organizes reasoning-oriented benchmark tasks into a progressive hierarchy to guide our model iteration (e.g., Perception, Derivation, Evolution, and Recommendation). For industrial deployment, we propose a Fast-Slow thinking architecture that continuously injects the knowledge of OneReason into the recommendation pipeline through the collaboration of nearline slow thinking and online fast thinking, contributing significant business

gains.

In the future, we will explore: (1) more stable online incremental-learning mechanisms that continuously adapt to new general-knowledge training corpora and user-interest drift while avoiding potential degradation of instruction-following and general capabilities; (2) more customized decoding and generation strategies tailored to business constraints, achieving a better balance among throughput, latency, quality, and business value; and (3) moving beyond “thinking-then-answer” toward an agentic recommendation foundation model with **planning and tool-calling abilities**. We believe that, with continued progress in model compression, inference infrastructure, and hardware efficiency, recommendation foundation models will become a core component of the recommendation chain, gradually driving the evolution toward an agentic-recommender architecture.

Acknowledgement

We thanks Gaoguo Sun, Peng Zhang, Lixing Zhao, Xinyu Zhang, Xinyue Zhang, Xun Zheng, Zheng Wang for their contributions to the data collection and insightful suggestions.

References

- Tishby, N. and N. Zaslavsky (2015). “Deep Learning and the Information Bottleneck Principle”. In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE, pp. 1–5. DOI: [10.1109/ITW.2015.7133169](https://doi.org/10.1109/ITW.2015.7133169). arXiv: [1503.02406](https://arxiv.org/abs/1503.02406).
- Kang, W.-C. and J. McAuley (2018). “Self-attentive sequential recommendation”. In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 197–206.
- Saxe, A. M., Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox (2018). “On the Information Bottleneck Theory of Deep Learning”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ry_WPG-A-.
- Yelp, Inc. (2019). *Yelp Open Dataset*. <https://www.yelp.com/dataset>.
- Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361*.
- Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, A. Sastry Girish Askell, P. Mishkin, J. Clark, et al. (2021). “Learning transferable visual models from natural language supervision”. In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.
- Tan, J., S. Xu, Y. Ge, Y. Li, X. Chen, and Y. Zhang (2021). “Counterfactual Explainable Recommendation”. In: *arXiv preprint arXiv:2108.10539*.
- Zheng, Y., C. Gao, X. Li, X. He, Y. Li, and D. Jin (2021). “Disentangling User Interest and Conformity for Recommendation with Causal Embedding”. In: *Proceedings of the Web Conference 2021*. Association for Computing Machinery, pp. 2980–2991. DOI: [10.1145/3442381.3449788](https://doi.org/10.1145/3442381.3449788).
- Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre (2022). *Training Compute-Optimal Large Language Models*. arXiv: [2203.15556](https://arxiv.org/abs/2203.15556) [cs.CL]. URL: <https://arxiv.org/abs/2203.15556>.
- Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe (2022). *Training language models to follow instructions with human feedback*. arXiv: [2203.02155](https://arxiv.org/abs/2203.02155) [cs.CL]. URL: <https://arxiv.org/abs/2203.02155>.
- Gao, Y., T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang (2023). “Chat-rec: Towards interactive and explainable llms-augmented recommender system”. In: *arXiv preprint arXiv:2303.14524*.

- Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa (2023). *Large Language Models are Zero-Shot Reasoners*. arXiv: [2205.11916](https://arxiv.org/abs/2205.11916) [cs.CL]. URL: <https://arxiv.org/abs/2205.11916>.
- Li, J., W. Zhang, T. Wang, G. Xiong, A. Lu, and G. Medioni (2023). “GPT4Rec: A generative framework for personalized recommendation and user interests interpretation”. In: *arXiv preprint arXiv:2304.03879*.
- Rajput, S., N. Mehta, A. Singh, R. Erickson, N. Magd, P. Covington, et al. (2023). “Recommender systems with generative retrieval”. In: *Advances in Neural Information Processing Systems* 36.
- Sun, Z., Z. Si, X. Zang, D. Leng, Y. Niu, Y. Song, X. Zhang, and J. Xu (2023). “KuaiSAR: A Unified Search and Recommendation Dataset”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 5407–5411.
- Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: [2201.11903](https://arxiv.org/abs/2201.11903) [cs.CL]. URL: <https://arxiv.org/abs/2201.11903>.
- Yuan, Z., H. Yuan, C. Li, G. Dong, K. Lu, C. Tan, C. Zhou, and J. Zhou (2023). “Scaling relationship on learning mathematical reasoning with large language models”. In: *arXiv preprint arXiv:2308.01825*.
- Cheng, Y., Y. Pan, J. Zhang, Y. Ni, A. Sun, and F. Yuan (2024). “An Image Dataset for Benchmarking Recommender Systems with Raw Pixels”. In: *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*.
- Delétang, G., A. Ruoss, P.-A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau, M. Hutter, and J. Veness (2024). “Language Modeling Is Compression”. In: *International Conference on Learning Representations*. arXiv: [2309.10668](https://arxiv.org/abs/2309.10668).
- Hoffmann, D., K. Budhathoki, and M. Kleindessner (2024). “LLM-Rank: A Graph Theoretical Approach to Pruning Large Language Models”. In: *arXiv preprint arXiv:2410.13299*.
- Hou, Y., J. Li, X. Fu, Z. He, A. Yan, X. Chen, and J. McAuley (2024a). “Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders”. In: *arXiv preprint arXiv:2403.03952*.
- Hou, Y., J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao (2024b). “Large language models are zero-shot rankers for recommender systems”. In: *European conference on information retrieval*. Springer, pp. 364–381.
- Jaech, A., A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. (2024). “Openai o1 system card”. In: *arXiv preprint arXiv:2412.16720*.
- Luo, H., F. Zhuang, R. Xie, H. Zhu, D. Wang, Z. An, and Y. Xu (2024). “A survey on causal inference for recommendation”. In: *The Innovation* 5.2, p. 100590. DOI: [10.1016/j.xinn.2024.100590](https://doi.org/10.1016/j.xinn.2024.100590).
- Parmar, J., S. Satheesh, M. Patwary, M. Shoeybi, and B. Catanzaro (2024). *Reuse, Don’t Retrain: A Recipe for Continued Pretraining of Language Models*. arXiv: [2407.07263](https://arxiv.org/abs/2407.07263) [cs.CL]. URL: <https://arxiv.org/abs/2407.07263>.
- Rusak, E., P. Reizinger, A. Juhos, O. Bringmann, R. S. Zimmermann, and W. Brendel (2024). “Infonce: Identifying the gap between theory and practice”. In: *arXiv preprint arXiv:2407.00143*.
- Shao, Z., P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, et al. (2024). “Deepseek-math: Pushing the limits of mathematical reasoning in open language models”. In: *arXiv preprint arXiv:2402.03300*.
- Wang, Y., X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al. (2024). “Mmlu-pro: A more robust and challenging multi-task language understanding benchmark”. In: *Advances in Neural Information Processing Systems* 37, pp. 95266–95290.
- Zhai, S., T. Likhomanenko, E. Littwin, M. Daniluk, M. Bautista, et al. (2024). “Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations”. In: *arXiv preprint arXiv:2402.17152*.

- Zheng, B., Y. Hou, H. Lu, Y. Chen, W. X. Zhao, M. Chen, and J.-R. Wen (2024a). “Adapting large language models by integrating collaborative semantics for recommendation”. In: *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, pp. 1435–1448.
- Zheng, Y. et al. (2024b). “Large Language Models as Recommender Systems”. In: *arXiv preprint arXiv:2402.xxxx*.
- Agarwal, P., A. Badrinath, L. Bhasin, J. Yang, E. Botta, J. Xu, and C. Rosenberg (2025). “Pinrec: Outcome-conditioned, multi-token generative retrieval for industry-scale recommendation systems”. In: *arXiv preprint arXiv:2504.10507*.
- Dai, S., J. Tang, J. Wu, K. Wang, Y. Zhu, B. Chen, B. Hong, Y. Zhao, C. Fu, K. Wu, et al. (2025). “Onepiece: Bringing context engineering and reasoning to industrial cascade ranking system”. In: *arXiv preprint arXiv:2509.18091*.
- Deng, J., S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou (2025). “Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment”. In: *arXiv preprint arXiv:2502.18965*.
- Gao, C., C. Zheng, X.-H. Chen, K. Dang, S. Liu, B. Yu, A. Yang, S. Bai, J. Zhou, and J. Lin (2025). “Soft adaptive policy optimization”. In: *arXiv preprint arXiv:2511.20347*.
- Guo, D., D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. (2025). “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning”. In: *arXiv preprint arXiv:2501.12948*.
- Hao, X., M. Zhang, A. Li, X. Qian, Z. Ma, Y. Zang, S. Yang, Z. Han, X. Ma, J. Liu, et al. (2025). “OxygenREC: An Instruction-Following Generative Framework for E-commerce Recommendation”. In: *arXiv preprint arXiv:2512.22386*.
- He, T., M. Xie, R. Li, X. Xu, J. Yu, Z. Wang, L. Hu, H. Li, and K. Gai (2025). “An End-to-End Multi-objective Ensemble Ranking Framework for Video Recommendation”. In: *arXiv preprint arXiv:2508.05093*.
- Jiang, D., R. Zhang, Z. Guo, Y. Li, Y. Qi, X. Chen, L. Wang, J. Jin, C. Guo, S. Yan, et al. (2025a). “Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency”. In: *arXiv preprint arXiv:2502.09621*.
- Jiang, J., C. Ma, X. Song, H. Zhang, and J. Luo (2025b). *Corvid: Improving Multimodal Large Language Models Towards Chain-of-Thought Reasoning*. arXiv: 2507.07424 [cs.CV]. URL: <https://arxiv.org/abs/2507.07424>.
- Lee, C., S. Lim, K. Lee, and J. yong Sohn (2025). “On the Similarities of Embeddings in Contrastive Learning”. In: *Forty-second International Conference on Machine Learning*. URL: <https://openreview.net/forum?id=ChmJZ9V2o1>.
- Li, W., S. Wang, K. Cai, J. Deng, X. Wang, Q. Hu, D. Lian, and G. Zhou (2025). “Taming Ultra-Long Behavior Sequence in Session-wise Generative Recommendation”. In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5839–5846.
- Lin, H., Z. Yang, J. Xue, Z. Zhang, L. Wang, Y. Gu, Y. Xu, and X. Li (2025). “Spacetime-gr: A spacetime-aware generative model for large scale online poi recommendation”. In: *arXiv preprint arXiv:2508.16126*.
- Liu, Z., S. Wang, X. Wang, R. Zhang, J. Deng, H. Bao, J. Zhang, W. Li, P. Zheng, X. Wu, Y. Hu, Q. Hu, X. Luo, L. Ren, Z. Zhang, Q. Wang, K. Cai, Y. Wu, H. Cheng, Z. Cheng, L. Ren, H. Wang, Y. Su, R. Tang, K. Gai, and G. Zhou (2025). *OneRec-Think: In-Text Reasoning for Generative Recommendation*. arXiv: 2510.11639 [cs.IR]. URL: <https://arxiv.org/abs/2510.11639>.
- Luo, X., J. Cao, T. Sun, J. Yu, R. Huang, W. Yuan, H. Lin, Y. Zheng, S. Wang, Q. Hu, C. Qiu, J. Zhang, X. Zhang, Z. Yan, J. Zhang, S. Zhang, M. Wen, Z. Liu, and G. Zhou (2025). “QARM: Quantitative Alignment Multi-Modal Recommendation at Kuaishou”. In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. CIKM ’25. Seoul, Republic of Korea: Association for Computing Machinery, 5915–5922.

- Ovadia, O., M. Brief, R. Lemberg, and E. Sheerit (2025). “Knowledge-Instruct: Effective Continual Pre-training from Limited Data using Instructions”. In: *arXiv preprint arXiv:2504.05571*. arXiv: [2504.05571](https://arxiv.org/abs/2504.05571) [cs.CL].
- Qwen3, A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu (2025). “Qwen3 Technical Report”. In: *arXiv preprint arXiv:2505.09388*.
- Ren, Y. and D. J. Sutherland (2025). “Learning Dynamics of LLM Finetuning”. In: *The Thirteenth International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=tPNH0oZF19>.
- Tan, J., Y. Chen, A. Zhang, J. Jiang, B. Liu, Z. Xu, H. Zhu, J. Xu, B. Zheng, and X. Wang (2025). “Reinforced preference optimization for recommendation”. In: *arXiv preprint arXiv:2510.12211*.
- Xing, H., H. Deng, Y. Mao, L. Mu, J. Hu, Y. Xu, H. Zhang, J. Wang, S. Wang, Y. Zhang, et al. (2025). “Reg4rec: Reasoning-enhanced generative model for large-scale recommendation systems”. In: *arXiv preprint arXiv:2508.15308*.
- Yi, C., D. Chen, G. Guo, J. Tang, J. Wu, J. Yu, M. Zhang, W. Chen, W. Yang, Y. Luo, et al. (2025a). “RecGPT-V2 Technical Report”. In: *arXiv preprint arXiv:2512.14503*.
- Yi, C., D. Chen, G. Guo, J. Tang, J. Wu, J. Yu, M. Zhang, S. Dai, W. Chen, W. Yang, et al. (2025b). “Recgpt technical report”. In: *arXiv preprint arXiv:2507.22879*.
- Yu, Q., Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, Y. Song, X. Wei, H. Zhou, J. Liu, W.-Y. Ma, Y.-Q. Zhang, L. Yan, M. Qiao, Y. Wu, and M. Wang (2025). *DAPO: An Open-Source LLM Reinforcement Learning System at Scale*. arXiv: [2503.14476](https://arxiv.org/abs/2503.14476) [cs.LG]. URL: <https://arxiv.org/abs/2503.14476>.
- Zhang, J., Y. Cheng, Y. Ni, Y. Pan, Z. Yuan, J. Fu, Y. Li, J. Wang, and F. Yuan (2025a). “NineRec: A Benchmark Dataset Suite for Evaluating Transferable Recommendation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, pp. 5256–5267.
- Zhang, Y., M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou (2025b). “Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models”. In: *arXiv preprint arXiv:2506.05176*.
- Zheng, C., S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, J. Zhou, and J. Lin (2025). *Group Sequence Policy Optimization*. arXiv: [2507.18071](https://arxiv.org/abs/2507.18071) [cs.LG]. URL: <https://arxiv.org/abs/2507.18071>.
- Zhou, C., M. Wang, Y. Ma, C. Wu, W. Chen, Z. Qian, X. Liu, Y. Zhang, J. Wang, H. Xu, F. Luo, X. Chen, X. Hao, H. Li, A. Zhang, W. Wang, K. Zhang, G. Jia, L. Li, Z. Lu, Y. Lu, and Y. Guo (2025a). *From Perception to Cognition: A Survey of Vision-Language Interactive Reasoning in Multimodal Large Language Models*. arXiv: [2509.25373](https://arxiv.org/abs/2509.25373) [cs.AI]. URL: <https://arxiv.org/abs/2509.25373>.
- Zhou, G., J. Deng, J. Zhang, K. Cai, L. Ren, Q. Luo, Q. Wang, Q. Hu, R. Huang, S. Wang, W. Ding, W. Li, X. Luo, X. Wang, Z. Cheng, Z. Zhang, B. Zhang, B. Wang, C. Ma, C. Song, C. Wang, D. Wang, D. Meng, F. Yang, F. Zhang, F. Jiang, F. Zhang, G. Wang, G. Zhang, H. Li, H. Hu, H. Lin, H. Cheng, H. Cao, H. Wang, J. Huang, J. Chen, J. Liu, J. Jia, K. Gai, L. Hu, L. Zeng, L. Yu, Q. Wang, Q. Zhou, S. Wang, S. He, S. Yang, S. Yang, S. Huang, T. Wu, T. He, T. Gao, W. Yuan, X. Liang, X. Xu, X. Liu, Y. Wang, Y. Wang, Y. Liu, Y. Song, Y. Zhang, Y. Wu, Y. Zhao, and Z. Liu (2025b). *OneRec Technical Report*. arXiv: [2506.13695](https://arxiv.org/abs/2506.13695) [cs.IR]. URL: <https://arxiv.org/abs/2506.13695>.
- Zhou, G., H. Hu, H. Cheng, H. Wang, J. Deng, J. Zhang, K. Cai, L. Ren, L. Ren, L. Yu, P. Zheng, Q. Luo, Q. Wang, Q. Hu, R. Huang, R. Tang, S. Wang, S. Yang, T. Wu, W. Li, X. Luo, X. Wang, Y. Su, Y. Wu,

- Z. Cheng, Z. Liu, Z. Zhang, B. Zhang, B. Wang, C. Ma, C. Song, C. Wang, C. Chu, D. Wang, D. Meng, D. Zang, F. Yang, F. Zhang, F. Jiang, F. Zhang, G. Wang, G. Zhang, H. Li, H. Bao, H. Cao, J. Huang, J. Chen, J. Liu, J. Jia, K. Gai, L. Hu, L. Zeng, Q. Wang, Q. Zhou, R. Zhang, S. Wang, S. He, S. Yang, S. Mao, S. Huang, T. He, T. Gao, W. Yuan, X. Liang, X. Xu, X. Liu, Y. Wang, Y. Zhou, Y. Wang, Y. Liu, Y. Song, Y. Zhang, Y. Zhao, Z. Ling, and Z. Li (2025c). *OneRec-V2 Technical Report*. arXiv: [2508.20900](https://arxiv.org/abs/2508.20900) [cs.IR]. URL: <https://arxiv.org/abs/2508.20900>.
- Zhou, G., H. Hu, H. Cheng, H. Wang, J. Deng, J. Zhang, K. Cai, L. Ren, L. Ren, L. Yu, et al. (2025d). “Onerec-v2 technical report”. In: *arXiv preprint arXiv:2508.20900*.
- Aksenov, V., E. Bodnia, M. H. Freedman, and M. Mulligan (2026). “Compression Is All You Need: Modeling Mathematics”. In: *arXiv preprint arXiv:2603.20396*. arXiv: [2603.20396](https://arxiv.org/abs/2603.20396).
- Chen, B., S. Wang, Y. Ma, Z. Liang, X. Zhang, Y. Lv, Y. Yang, H. Dai, L. Mao, T. Zhao, et al. (2026a). “OneSearch-V2: The Latent Reasoning Enhanced Self-distillation Generative Search Framework”. In: *arXiv preprint arXiv:2603.24422*.
- Chen, T., J. Ou, Z. Liu, R. Tang, J. Liang, and H. Li (2026b). “Counteraction-Aware Multi-Teacher On-Policy Distillation for General Capability Recovery with Domain Preservation”. In: *arXiv preprint arXiv:2605.27115*.
- Chen, Z., R. Lu, A. Zhao, Z. Wang, Y. Yue, S. Song, and G. Huang (2026c). “Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?” In: *Advances in Neural Information Processing Systems* 38, pp. 57654–57689.
- DeepSeek-AI (2026). *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*.
- Ding, Y., Z. Guo, J. Li, L. Peng, S. Shao, W. Shao, X. Luo, L. Simon, J. Shang, J. McAuley, et al. (2026). “How Well Does Generative Recommendation Generalize?” In: *arXiv preprint arXiv:2603.19809*.
- Dou, C., F. Yang, F. Li, J. Jia, Q. Ju, S. Wang, T. Li, X. Zeng, Y. Zhou, H. Zhang, et al. (2026). “Baichuan-M3: Modeling Clinical Inquiry for Reliable Medical Decision-Making”. In: *arXiv preprint arXiv:2602.06570*.
- Feng, J., X. Kong, L. Sheng, B. Wu, C. Yi, F. Yang, X.-R. Sheng, H. Zhu, X. Wang, J. Wu, and X. He (2026). *Fine-grained Semantics Integration for Large Language Model-based Recommendation*. arXiv: [2602.22632](https://arxiv.org/abs/2602.22632) [cs.IR]. URL: <https://arxiv.org/abs/2602.22632>.
- He, Y., Y. Sun, J. Tan, Y. Chen, X. Kong, C. Shen, X. Wang, A. Zhang, and T.-S. Chua (2026). “Reasoning over semantic ids enhances generative recommendation”. In: *arXiv preprint arXiv:2603.23183*.
- Huang, A., A. Li, A. Kong, B. Wang, B. Jiao, B. Dong, B. Wang, B. Chen, B. Li, B. Ma, et al. (2026). “Step 3.5 Flash: Open Frontier-Level Intelligence with 11B Active Parameters”. In: *arXiv preprint arXiv:2602.10604*.
- Li, Y., Y. Zuo, B. He, J. Zhang, C. Xiao, C. Qian, T. Yu, H.-a. Gao, W. Yang, Z. Liu, et al. (2026). “Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe”. In: *arXiv preprint arXiv:2604.13016*.
- Sun, D., Y. Liu, J. Zhou, X. Liu, C. Yu, Y. Li, J. Zhang, H. Yu, and J. Jiang (2026a). “OneRanker: Unified Generation and Ranking with One Model in Industrial Advertising Recommendation”. In: *arXiv preprint arXiv:2603.02999*.
- Sun, K., X. Yuan, H. Liu, C. Zhao, C. Zhang, M. Dredze, and F. Bai (2026b). *Reading, Not Thinking: Understanding and Bridging the Modality Gap When Text Becomes Pixels in Multimodal LLMs*. arXiv: [2603.09095](https://arxiv.org/abs/2603.09095) [cs.CL]. URL: <https://arxiv.org/abs/2603.09095>.
- Tang, J., S. Dai, T. Shi, J. Xu, X. Chen, W. Chen, J. Wu, and Y. Jiang (2026). “Think before recommend: Unleashing the latent reasoning power for sequential recommendation”. In: *IEEE Transactions on Knowledge and Data Engineering*.
- Vyas, A., H.-J. Chang, C.-F. Yang, P.-Y. Huang, L. Gao, J. Richter, S. Chen, M. Le, P. Dollár, C. Feichtenhofer, et al. (2026). “Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30172–30182.

- Wang, S., Y. Huang, R. Yang, S. Wen, P. Xu, J. Cao, Y. Liu, K. Cai, C. Guo, S. Wang, et al. (2026). “OneLive: Dynamically Unified Generative Framework for Live-Streaming Recommendation”. In: *arXiv preprint arXiv:2602.08612*.
- Xiao, B., B. Xia, B. Yang, B. Gao, B. Shen, C. Zhang, C. He, C. Lou, F. Luo, G. Wang, et al. (2026). “Mimo-v2-flash technical report”. In: *arXiv preprint arXiv:2601.02780*.
- Xue, B., D. Liu, L. Wang, M. Sun, P. Wang, P. Zhang, S. Shi, T. Xu, Y. Sha, Z. Liu, et al. (2026). “Generative Recommendation for Large-Scale Advertising”. In: *arXiv preprint arXiv:2602.22732*.
- Yang, Z., Z. Liu, Y. Chen, W. Dai, B. Wang, S.-C. Lin, C. Lee, Y. Chen, D. Jiang, J. He, et al. (2026). “Nemotron-Cascade 2: Post-Training LLMs with Cascade RL and Multi-Domain On-Policy Distillation”. In: *arXiv preprint arXiv:2603.19220*.
- Zhang, K., J. Zhang, W. Cheng, Y. Cheng, J. Zhang, H. Lu, X. Zhang, H. Gan, J. Cao, T. Wang, et al. (2026a). “OneMall: One Model, More Scenarios–End-to-End Generative Recommender Family at Kuaishou E-Commerce”. In: *arXiv preprint arXiv:2601.21770*.
- Zhang, L., Y. Huang, H. Lv, M. Yin, L. Li, Z. Chen, H. Wang, and E. Chen (2026b). “Why Thinking Hurts? Diagnosing and Rectifying the Reasoning Shift in Foundation Recommender Models”. In: *arXiv preprint arXiv:2602.16587*.
- Zhang, X., Y. Xiong, J. Zhou, Z. Guo, Z. Pang, J. Huo, J. Wang, X. Sun, E. Zhang, J. Jin, et al. (2026c). “Unified Value Alignment for Generative Recommendation in Industrial Advertising”. In: *arXiv preprint arXiv:2605.05803*.
- Zhou, G., H. Bao, J. Huang, J. Deng, J. Zhang, J. She, K. Cai, L. Ren, L. Ren, Q. Luo, Q. Wang, Q. Hu, R. Zhang, R. Tang, S. Wang, W. Li, X. Wu, X. Luo, X. Wang, Y. Hu, Y. Wu, Z. Liu, Z. Zhang, Z. Zhang, B. Chen, B. Wen, C. Ma, C. Song, C. Chu, D. Lian, F. Yang, F. Jiang, H. Cheng, H. Wang, K. Gai, P. Zheng, Q. Wang, R. Huang, S. Mao, T. Gao, W. Yuan, Y. Wang, Y. Zhou, Y. Su, Z. Cheng, Z. Ling, and Z. Li (2026). *OpenOneRec Technical Report*. arXiv: [2512.24762](https://arxiv.org/abs/2512.24762) [cs.IR]. URL: <https://arxiv.org/abs/2512.24762>.

12. Author List

Core Contributors Biao Yang, Boyang Ding^{*}, Chenglong Chu, Dunju Zang, Fei Pan, Han Li, Hao Jiang, Honghui Bao, Huanjie Wang, Jian Liang, Jiangxia Cao, Jiao Ou, Jiabin Deng, Jinghao Zhang, Kun Gai, Lu Ren, Peiru Du, Pengfei Zheng^{*}, Rongzhou Zhang, Ruiming Tang, Shiyao Wang^{*}, Siyang Mao, Siyuan Lou, Teng Shi^{*}, Wei Yuan, Wenlong Xu, Xingchen Liu, Xingmei Wang, Xinqi Jin, Yan Sun, Yan Wang^{*}, Yifei Hu, Yingzhi He, Yufei Ye, Yuhao Wang, Yunhao Zhou, Yuqin Dai, Zhao Liu, Zhipeng Wei, Zhixin Ling, Ziming Li, Zixing Zhang^{*}, Ziyuan Liu.

Contributors An Zhang, Changxin Lao, Chaoyi Ma, Chengru Song, Defu Lian, Fan Yang, Guowang Zhang, Hao Peng, Jiayao Shen, Jie Chen, Jun Xu, Junmin Chen, Kun Zhang, Kuo Cai, Mingxing Wen, Minmao Wang, Minxuan Lv, Qi Zhang, Qiang Luo, Sheng Yu, Shijie Li, Shijie Yi, Shuang Yang, Shugui Liu, Shuni Chen, Tinghai Zhang, Tingting Gao, Xiang Wang, Xiangyu Wu, Xiangyu Zhao, Xiao Lv, Xiaoyou Zhou, Xuming Wang, Yong Du, Zejian Zhang, Zhaojie Liu, Zhiyang Zhang, Zhuang Zhuang, Ziqi Wang, Ziyi Zhao.

All the authors listed alphabetically by first name. ^{*} individuals who have departed from our team.

A. Deployment Details

A.1. Industrial Scenario Adaptation

Facing a new industrial scenario, how can OneReason be rapidly adapted? This setting presents two key challenges: understanding new item corpora and learning emerging user interest distributions, both of which are absent during OneReason training. To address these challenges, we propose a reusable deployment recipe called Industrial Scenario Adaptation.

Scenario-aware Continual Pre-Train. We start from the checkpoint of the final OneReason foundation model. For item understanding in new industrial scenarios, we adapt the RO: Perception 5.1 method for bidirectional alignment between itemic tokens and captions. For user understanding in new industrial scenarios, we adopt a User-Profile-Based Continual Pre-Training strategy. A large-scale LLM is first used to summarize user behaviors and extract preference signals, which are then combined with user profiles to form user portrait texts. These user portraits texts are used for continual pre-training, enabling the model to capture scenario-specific user characteristics.

Scenario-aware Continual Supervised Fine-Tuning. We further train OneReason for recommendation by conditioning on user portrait texts and historical interactions to predict the next itemic token. To improve robustness across users with different activity levels, we introduce a curriculum learning strategy that progresses from high-activity to low-activity users. This allows the model to first learn from dense behavioral signals and gradually adapt to sparse scenarios, while user portraits provide consistent semantic guidance for generalization.

A.2. The Overall Methodology of OneReason for OneRec

As reflected in the core design philosophy of the OneReason, **Thinking-then-Generation** have been demonstrated to be an important pathway in the evolution of LLMs. Prior work has also shown that introducing latent reasoning mechanisms into generative models can effectively enhance reasoning capabilities and further improve recommendation performance. Therefore, we attempt to leverage the outputs of OneReason to provide reasoning supervision, and we have obtained business uplift in the online fast pipeline, validating that OneReason’s reasoning capabilities and knowledge can be transferred to generative recommendation model via distillation.

The overall framework is illustrated in the Figure 27. Specifically, we first use OneReason to predict the next itemic token with the highest probability of user interaction, and then decode it into the embedding representation corresponding to the itemic token via a quantized model, serving as a supervision signal for the reasoning process. Subsequently, we introduce a **Thinking Token** into the generative recommendation model and use the hidden state at the decoder-side <BOS> position as the supervision target. We then apply an **Alignment Network** to impose representation-level constraints, thereby distilling OneReason’s knowledge and reasoning capabilities into the generative recommendation model within a high-dimensional semantic space.

On this basis, we further explore several variant designs. For example, we introduce a User Representation Decoder to compress the encoder outputs, align them with OneReason’s outputs, and further distill them into the Thinking Token. In addition, to address the uncontrollability of the reasoning process, we design a dual-branch architecture consisting of a Decoder with Reason and a Decoder without Reason, and apply mutual supervision to improve the stability of the reasoning process. Finally, considering the inherent misalignment between the recommendation space and the semantic space, we further propose a contrastive distillation method based on spatial residuals to enhance cross-space alignment.

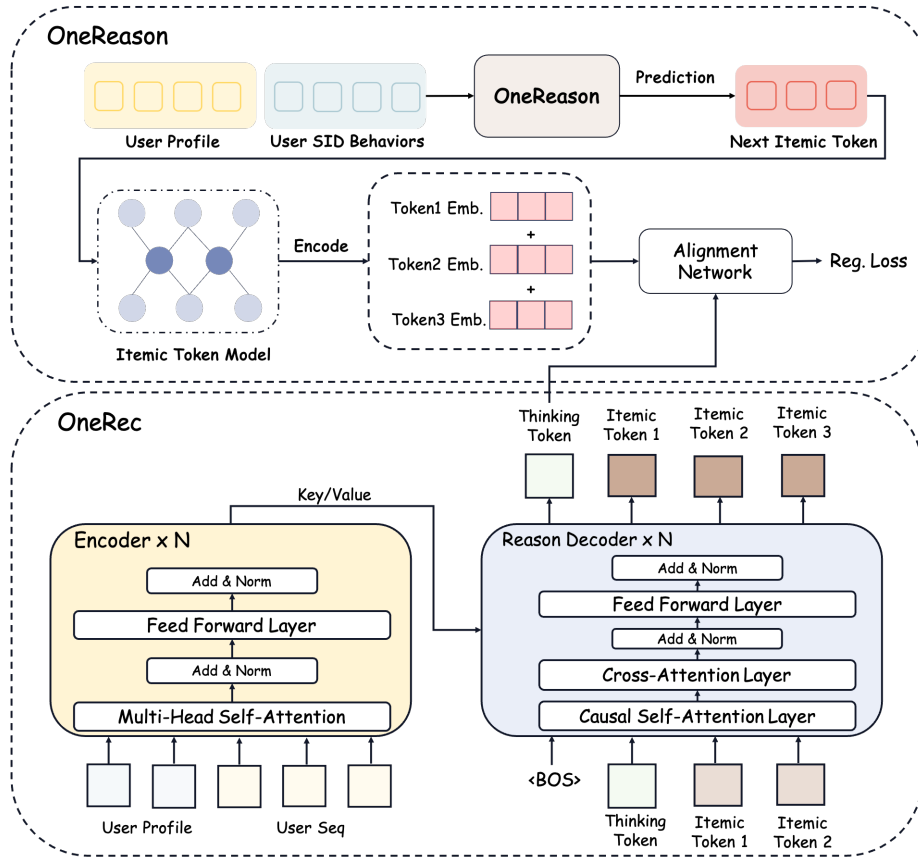


Figure 27 | The overall structure of OneReason for OneRec.

A.3. Experimental Analysis

We conduct a detailed analysis of the above A/B test results. We decompose the overall effects from two perspectives: (1) the changes in efficiency brought by OneReason, and (2) the advantages of OneReason across different user groups.

Distribution and Efficiency Analysis. From Table 19, we observe that OneReason brings consistent system-level improvements under different deployment paradigms. In the slow pipeline, direct application of OneReason improves conversion and click performance by enhancing traffic quality through better semantic matching. In contrast, enhancing OneRec in the fast pipeline mainly improves system scale via expanded retrieval coverage, while per-impression metrics remain relatively stable. The combined strategy achieves the best overall performance with further increased coverage, demonstrating the complementarity between the two paradigms: the slow pipeline improves traffic efficiency, while the fast pipeline expands system scale.

Model	Conv.	Click	CVR	CTR	Retr. Share
OneReason	+5.870%	+3.986%	+4.516%	+2.497%	3.8%
OneReason for OneRec	+7.935%	+0.207%	+0.459%	+0.195%	22.1%
Combined	+12.643%	+1.709%	+1.865%	+0.519%	27.2%

Table 19 | Online traffic efficiency and retrieval distribution analysis. **Conv.** denotes conversion numbers, **Click** denotes click numbers, **CVR** denotes conversion rate, **CTR** denotes click rate.

User Level Analysis. To evaluate the user-level impact of OneReason, we conduct a stratified analysis across high-activity, medium-activity, and low-activity users. Table 20 shows that business revenue gains increase as user activity decreases, with the largest improvements observed in low-activity users. Compared with traditional recommendation models that rely heavily on users’ historical behaviors, OneReason enables reasoning-based interest completion by integrating user profiles, limited behaviors, and content semantics, enabling effective recommendations under sparse interactions. Overall, these results validate the value of its reasoning capability for low-activity users and demonstrate substantial information gain for the overall system.

User Level	Impressions	Revenue
Active Users	+0.763%	+2.419%
Mid-Active Users	+0.902%	+4.815%
Low-Active Users	+1.125%	+13.323%

Table 20 | User-level performance breakdown under OneReason direct retrieval.

A.4. Discussions in Industrial Deployment

We analyze the key challenges encountered during the online deployment of OneReason and discuss several important issues, along with future directions.

A.4.1. D1: Online Incremental Training and General Model Performance.

OneReason requires incremental updates to adapt to evolving item corpora and user interests. Without such updates, we observe significant performance degradation. However, the current approach primarily focuses on continual learning for recommendation objectives, while partially neglecting instruction-following capabilities and general capability retention. Future work should explore more stable incremental learning strategies that preserve foundational capabilities while adapting to evolving item corpora and shifting user-interest distributions.

A.4.2. D2: Generation Strategies of OneReason.

OneReason has higher inference cost, making large-scale beam search infeasible. We therefore design a hierarchical generation strategy for item ID decoding under small-scale beam search. The model first generates candidates at the top itemic token-level to ensure coverage, and then completes lower-level structures via greedy decoding, balancing diversity, accuracy, and efficiency under latency constraints. However, the current strategy still lacks consideration of business value and resource constraints. Future work may explore more customized decoding strategies to improve recommendation quality.

A.4.3. D3: Model Parameter Size and Real-time Applications.

In fact, the “Fast-Slow Thinking” architecture is an engineering trade-off. It preserves real-time system stability while incorporating information gains from OneReason via a nearline mechanism. However, it cannot fully replace real-time retrieval nor fully exploit the model’s capacity. A key future direction is model parameter size, i.e., exploring smaller models (e.g., 0.8B level) that maintain comparable performance while enabling true real-time retrieval, further improving system efficiency and integration.

A.4.4. D4: ROI of OneReason for Online Deployment.

We first note that ROI is not the primary objective of OneReason. Our main goal is to evaluate whether the model provides meaningful information gain in recommendation systems. However, from an engineering perspective, the system consumes approximately 600 flagship GPUs per day to serve full traffic across 400 million users in the Kuaishou App. We analyze A/B test experimental results and estimate an ROI > 5 , indicating that system-level gains outweigh inference costs and yield a positive return. With advances in model compression, inference infrastructure, and hardware efficiency, OneReason is expected to become a core component of future recommendation systems, driving a shift toward a foundation-centric architecture.

B. Benchmark Details

B.1. Task Construction

We provide the construction details of the four reasoning layers in OneReason-Bench here.

R0: Perception. We use multimodal LLMs to generate reference item captions from native multi-modal signals across video, live-streaming, product, and advertising domains, followed by LLM-based and human quality checks. We next construct content-grounded QA instances from the verified captions and item metadata, and further filter them by answerability, difficulty, and human spot-checking.

R1: Derivation. We extract item-to-item association pairs from content knowledge graphs. We further apply LLM-based difficulty filtering and human quality checks to retain samples that require semantic derivation rather than surface-level matching.

R2: Evolution. We mine interest evolution chains with temporal progression and cognitive increments from full-domain user behavior timelines spanning short video, e-commerce, advertising, live-streaming, and additional search. Rule-based, LLM-based, and human quality checks are used to filter pseudo-logical chains, and the resulting chains are converted into the three task formats above.

R3: Recommendation. We build all-domain history sequences from real user behavior logs and use users' subsequent high-value interacted items as prediction targets, where high-value interaction is defined domain-specifically: clicks for Product, watch-time above the 75th percentile within the same duration bucket for Video, conversion actions (e.g., activation or payment) for Ad, and first gifting events for Live. Target items are further processed with abnormal-sample filtering, popularity downsampling, and category balancing to reduce head-item dominance.

B.2. Recommendation Data Statistics and Analysis

Table 21 presents the scale statistics of R3 across the four target domains.

Beyond the scale statistics, we further characterize train-evaluation overlap from two complementary perspectives: target-level visibility and transition-level memorization/generalization. For target-level visibility, 33.69% of target item origin IDs are unseen during training, whereas only 11.55% of target itemic patterns are unseen. This suggests that many item origin ID-level cold-start targets still support content-based generalization through seen itemic patterns.

Table 21 | Statistics of R3 recommendation data in OneReason-Bench across four target domains. Each user corresponds to one instance.

Domain	# Users	# In-Domain Items	# Out-of-Domain Items	# In-Domain Inter.	# Out-of-Domain Inter.	Avg. In-Domain Hist.	Avg. Out-of-Domain Hist.	Avg. Tgt.
Video	10,285	694,507	419,428	770,228	781,533	74.89	75.99	13.92
Product	9,859	502,668	599,238	760,277	986,817	77.12	100.09	1.62
Ad	9,081	249,309	718,771	840,760	1,089,317	92.58	119.96	4.24
Live	8,083	66,831	810,269	240,577	1,229,617	29.76	152.12	1.58

Following Ding et al. (2026), we then examine transition-level memorization and generalization at the itemic-pattern granularity. For an evaluation instance with history ($\text{item}_1, \text{item}_2, \dots, \text{item}_n$) and target set \mathcal{T} , each pair (item_n, t) for $t \in \mathcal{T}$ is counted as a memorized pair if the same transition appears in training, and as a generalized pair otherwise. As shown in Table 22, Video and Product are dominated by generalized transitions, whereas Ad and Live are more memorization-heavy, allowing OneReason-Bench to cover both settings.

Table 22 | Memorization and generalization statistics of R3 target transitions at the itemic-pattern granularity in OneReason-Bench.

Domain	Memorized Pairs	Generalized Pairs
Video	4.7%	95.3%
Product	27.8%	72.2%
Ad	75.9%	24.1%
Live	73.0%	27.0%

B.3. Comparison with Existing Benchmarks

Existing benchmarks have progressively expanded the evaluation scope of recommender systems. Traditional recommendation benchmarks, including Amazon (Hou et al., 2024a), Yelp (Yelp, Inc., 2019), PixelRec (Cheng et al., 2024), NineRec (Zhang et al., 2025a), and KuaiSAR (Sun et al., 2023), mainly focus on outcome-level evaluation such as ranking accuracy, next-item prediction, and behavior prediction. Within recommendation foundation-model evaluation, ReCF-Bench provides the immediate basis for this work by establishing a holistic instruction-following benchmark.

OneReason-Bench continues this line of work by focusing on recommendation-specific reasoning diagnosis. Instead of only evaluating the final recommendation outcome or response-level quality, it introduces explicit diagnostic tasks for item-level relation derivation and temporal user-interest evolution as complementary probes of recommendation-specific reasoning, while retaining next-items prediction as a downstream recommendation task. Table 23 summarizes this positioning.

B.4. Evaluation Metrics

B.4.1. Item Understanding Evaluation

For the Item Understanding task, we reuse the LLM-as-a-Judge protocol introduced in OpenOneRec. The metric definition is as follows: generated descriptions are decomposed into weighted information points and matched against the corresponding ground-truth description. Each information point carries a predefined weight reflecting its semantic importance; the final score is the OpenOneRec double-weighted F1, which penalizes both hallucinated points (false positives) and omitted points

Category	Benchmark	Multi-domain Knowledge	Instruction-Following	Reasoning-Centric	SeqRec	Cross-Domain Rec
General LLM	GSM8K / MATH	X	✓	✓	X	X
	MMLU-Pro / GPQA	✓	✓	✓	X	X
	CorrectBench	X	✓	✓	X	X
	CriticBench	✓	✓	✓	X	X
Classic Rec	Yelp / MovieLens / PixelRec	X	X	X	✓	X
	Amazon / NineRec	X	X	X	✓	✓
	KuaiSAR	X	X	X	✓	X
Rec (LLM era)	RecBench+	X	✓	X	✓	X
	AgentRecBench	X	✓	X	✓	X
	HORIZON	X	X	X	✓	✓
RecFM	RecIF-Bench	✓	✓	X	✓	✓
	★ OneReason-Bench (Ours)	✓	✓	✓	✓	✓

✓ supported X none

Table 23 | Comparison of benchmarks across categories. OneReason-Bench is the only benchmark covering all five capabilities.

(false negatives) in proportion to their weights. Meanwhile, same with OpenOneRec, the evaluation is conducted over all examples in the Item Understanding split and the average double-weighted F1 is used to reported as the task score.

B.4.2. Action–Logic Score

The Action–Logic Score evaluates two generative tasks: Evolution Topic Generation and Evolution Direct Generation. Both require the model to produce *evolution chains*, each a sequence of events $e_k = (A_k, l_k)$, where A_k is an *action group*, i.e., a set of user actions characterizing a phase in the user’s interest-evolution process, and l_k is the corresponding logic statement. Let the ground-truth chain be $\mathcal{E}^* = (e_1^*, \dots, e_n^*)$ with $e_k^* = (A_k^*, l_k^*)$, and the generated chain be $\hat{\mathcal{E}} = (\hat{e}_1, \dots, \hat{e}_m)$ with $\hat{e}_j = (\hat{A}_j, \hat{l}_j)$. The metric combines action alignment and logic alignment quality into a single composite score.

Action Alignment. For a ground-truth event e_k^* and a generated event \hat{e}_j , the pairwise action similarity is defined as:

$$m_{F1}(A_k^*, \hat{A}_j) = \frac{2 |A_k^* \cap \hat{A}_j|}{|A_k^*| + |\hat{A}_j|}. \quad (30)$$

The resulting similarity matrix is aligned via maximum-weight monotone matching, equivalent to a weighted longest-common-subsequence alignment: it selects a set of one-to-one pairs \mathcal{M}_a with

strictly increasing ground-truth and generated indices that maximizes total action similarity.

The chain-level **action alignment score** aggregates pairwise similarities with a precision–recall F1 over the generated and ground-truth chains:

$$P_a = \frac{\sum_{(k,j) \in \mathcal{M}_a} m_{F1}(A_k^*, \hat{A}_j)}{|\hat{\mathcal{E}}|}, \quad (31)$$

$$R_a = \frac{\sum_{(k,j) \in \mathcal{M}_a} m_{F1}(A_k^*, \hat{A}_j)}{|\mathcal{E}^*|}, \quad (32)$$

$$\text{Action Alignment} = \frac{2P_a R_a}{P_a + R_a}. \quad (33)$$

Logic Evaluation. The **logic alignment score** reuses the matched pairs \mathcal{M}_a from Action Alignment. For each $(k, j) \in \mathcal{M}_a$, the mixed logic similarity is:

$$s_{\text{logic}}(l_k^*, \hat{l}_j) = 0.5 \text{Token}_{F1}(l_k^*, \hat{l}_j) + 0.5 \text{ROUGE-L}_{F1}(l_k^*, \hat{l}_j). \quad (34)$$

Here, Token_{F1} is the token-overlap F1 between the ground-truth and generated logic statements, and ROUGE-L_{F1} is the F1 variant of ROUGE-L, derived from the longest common subsequence of the two statements. Defining the total logic similarity over matched pairs as:

$$T = \sum_{(k,j) \in \mathcal{M}_a} s_{\text{logic}}(l_k^*, \hat{l}_j), \quad (35)$$

precision, recall, and the chain-level score are then:

$$P_\ell = \frac{T}{|\hat{\mathcal{E}}|}, \quad (36)$$

$$R_\ell = \frac{T}{|\mathcal{E}^*|}, \quad (37)$$

$$\text{Logic Alignment} = \frac{2P_\ell R_\ell}{P_\ell + R_\ell}. \quad (38)$$

Action–Logic Score Definition. The **Action–Logic Score** is the primary metric for generated evolution chains, defined as the equal-weight average of the two alignment sub-scores:

$$\text{Action–Logic Score} = 0.5 \times \text{Action Alignment} + 0.5 \times \text{Logic Alignment}. \quad (39)$$

For Evolution Topic Generation, scores are averaged across generations for each sample and then across all samples; parse failures receive zero score and remain in the denominator.

Multi-Chain Action–Logic Score. Evolution Direct Generation requires the model to produce a set of evolution chains $\{\hat{\mathcal{E}}\}$ without topic guidance, compared against a ground-truth set $\{\mathcal{E}^*\}$. The evaluator first computes a chain-level action score for every ground-truth–generated chain pair and applies the Hungarian algorithm to find the maximum-weight one-to-one assignment.

Each matched chain pair is then evaluated using the single-chain procedure above. Unmatched ground-truth chains receive zero score, and the **Multi-Chain Action–Logic Score** is averaged over the number of ground-truth chains. Extra generated chains do not directly enter this average, but they are penalized in the auxiliary chain-level F1 metrics through the precision denominator.

B.5. Task Formats

In this appendix, we provide representative input–output templates for the benchmark task formats.

Prompt Bench.R0.1: Item Understanding (R0 — Perception)

System: 你是一个直播内容分析专家。

User:

请将以下主播token还原为主播画像描述，涵盖形象气质与表达方式、内容方向与受众定位、互动氛围与风格特点: <|living_begin|><a_1658><b_451><c_7635>

Assistant:

该主播“XXXXXX”是一位极具生活气息的聊天娱乐类主播。她外在形象朴素亲切，常戴眼镜身穿棉服，以边吃边聊的真实状态示人。其直播内容聚焦于情感调解、家长里短及婚恋话题（如彩礼、相亲），通过多人连麦的形式搭建了一个“线上茶话会”。她性格豪爽、接地气，扮演着倾听者与调解员的角色...

English Example

System: *You are an expert in live-streaming content analysis.*

User:

Decode the following streamer token into a portrait description covering appearance and expression, content direction and audience positioning, and interaction style:

<|living_begin|><a_1658><b_451><c_7635>

Assistant:

The streamer “XXXXXX” is a lively chat-entertainment host whose style is steeped in everyday life. She appears plain and approachable, often wearing glasses and a padded jacket, presenting herself in the authentic state of chatting while eating. Her content focuses on emotional mediation, family gossip, and marriage-related topics (such as bride prices and matchmaking), building an “online tea party” through multi-person live-chat sessions. Outgoing and down-to-earth, she serves as both listener and mediator...

Prompt Bench.R0.2: Itemic Pattern Grounding (R0 — Perception)

System: 你是一个专业的直播主播分析助手。

User:

请分析以下主播的形象、内容和风格描述，生成匹配的主播token: 该主播“XXXXXX”是一位极具生活气息的聊天娱乐类主播。她外在形象朴素亲切，常戴眼镜身穿棉服，以边吃边聊的真实状态示人。其直播内容聚焦于情感调解、家长里短及婚恋话题（如彩礼、相亲），通过多人连麦的形式搭建了一个“线上茶话会”。她性格豪爽、接地气，扮演着倾听者与调解员的角色...

Assistant:

<|living_begin|><a_1658><b_451><c_7635>

English Example

System: *You are a professional live-streamer analysis assistant.*

User:

Analyze the following streamer’s appearance, content, and style description, and generate the matching streamer token: *The streamer “XXXXXX” is a lively chat-entertainment host whose style is steeped in everyday life. She appears plain and approachable, often wearing glasses and a padded jacket, presenting herself in the authentic state of chatting while eating. Her content focuses on emotional mediation, family gossip, and marriage-related topics (such as bride prices and matchmaking), building an “online tea party” through multi-person live-chat sessions. Outgoing and down-to-earth, she serves as both listener and mediator...*

Assistant:

<|living_begin|><a_1658><b_451><c_7635>

Prompt Bench.R0.3: Item QA (R0 — Perception)

System: 你是一位专业的短视频理解助手，能够精准掌握视频描述，并清晰准确地回答相关问题。

User:

视频<|video_begin|><a_5946><b_1435><c_6589>

视频如何体现地域文化？

- A. 呈现郑州现代城市景观与科技馆
- B. 使用河南方言和豫剧唱段
- C. 展示陕西兵马俑和秦腔艺术
- D. 展示河南古建筑和雪景

请直接回答A/B/C/D中的一个选项，不要输出任何其他内容。

Assistant:

D

English Example

System: You are a professional short-video understanding assistant, capable of accurately grasping video descriptions and clearly answering related questions.

User:

Video <|video_begin|><a_5946><b_1435><c_6589>

How does the video reflect regional culture?

- A. Presents Zhengzhou's modern cityscape and science museum
- B. Uses Henan dialect and Yu opera excerpts
- C. Showcases Shaanxi's terracotta warriors and Qinqiang art
- D. Shows ancient architecture of Henan and snow scenes

Please answer with one of A/B/C/D and nothing else.

Assistant:

D

Prompt Bench.R1.1: Item2Item (R1 — Derivation)

System: 你是一位视频推荐系统专家，擅长发现视频之间的深层关系。

User:

给定源视频: <|video_begin|><a_1584><b_6006><c_2381>

该视频与某个候选视频之间存在「潜在需求共鸣」关系。（潜在需求共鸣的定义：潜在需求共鸣：两个表面完全不同的视频，满足用户同一个深层心理需求。例：「30天极简断舍离挑战」「手机APP大清理只留5个」，深层需求都是简化生活获得掌控感。）

请从以下候选视频中选出与源视频存在该关系的视频：

- A. <|video_begin|><a_8152><b_3417><c_4595>
- B. <|video_begin|><a_1584><b_254><c_1317>
- C. <|video_begin|><a_6219><b_6006><c_4595>
- D. <|video_begin|><a_6219><b_3504><c_5673>

请以"答案: [选项字母]"的格式返回最终答案。

Assistant:

[答案] C

English example

System: You are a video recommendation system expert, skilled at discovering deep relationships between videos.

User:

Given the source video: <|video_begin|><a_1584><b_6006><c_2381>

This video shares a "Latent Need Resonance" relationship with one of the candidate videos below.

(Definition of Latent Need Resonance: two superficially different videos that satisfy the same deep psychological need. Example: "30-Day Minimalist Decluttering Challenge" "Phone App Purge: Keep Only

5,” both serving the deep need of simplifying life and gaining a sense of control.)

Please select the candidate video that has this relationship with the source video:

A. <|video_begin|><a_8152><b_3417><c_4595>

B. <|video_begin|><a_1584><b_254><c_1317>

C. <|video_begin|><a_6219><b_6006><c_4595>

D. <|video_begin|><a_6219><b_3504><c_5673>

Please return your final answer in the format “Answer: [Option Letter].”

Assistant:

[Answer] C

Prompt Bench.R2.1: Evolution Action Selection (R2 — Evolution)

User:

你是一名专业的数据标签专家。请根据给定的用户行为逻辑链，从行为时间线中提取所有相关的行为记录。

逻辑链：从赛那泛化改装兴趣到具体部件方案筛选

用户行为时间线：

【2023-10-25】[商品-购买] <|prod_begin|><a_5967><b_2249><c_743>

【2024-06-04】[直播-关注] <|living_begin|><a_4318><b_7342><c_5563>

...

Assistant:

<|video_begin|><a_4750><b_2230><c_4729> <|video_begin|><a_5844><b_5861><c_361>

<|video_begin|><a_3475><b_2493><c_1218>

English Example

User:

You are a professional data-labeling expert. Given a logic chain of user behavior, extract all related action records from the behavioral timeline.

Logic Chain: From generic Sienna modification interest to specific component plan selection.

User Behavioral Timeline:

[2023-10-25] [Product-Purchase] <|prod_begin|><a_5967><b_2249><c_743>

[2024-06-04] [Live-Follow] <|living_begin|><a_4318><b_7342><c_5563>

...

Assistant:

<|video_begin|><a_4750><b_2230><c_4729> <|video_begin|><a_5844><b_5861><c_361>

<|video_begin|><a_3475><b_2493><c_1218>

Prompt Bench.R2.2: Evolution Topic Generation (R2 — Evolution)

User:

【用户交互历史】：

【2023-10-25】[商品-购买] <|prod_begin|><a_5967><b_2249><c_743>

... (多域时间线) ...

【2025-12-25】[搜索] carplay转安卓手机盒子

...

目标兴趣主题：从赛那泛化改装兴趣到具体部件方案筛选

请基于历史生成完整的逻辑链。

Assistant:

{

 "logic_chain": {

 "name": "从赛那泛化改装兴趣到具体部件方案筛选",

 "events": [

 { "date": "2025-12-21", "action": "[搜索] 塞纳改装大屏教程,

```

    "logic": 触发阶段：以教程形式主动获取赛那改装路径，显示明确改装方向需求。 },
    { "date": "2025-12-23", "action": [搜索] 玩塞纳的十大忠告,
      "logic": 认知补充：从做法走向风险与经验框架。 },
    { "date": "2025-12-25", "action": [商品-点击] CarPlay转安卓手机盒子,
      "logic": 决策收敛：从教程/忠告调研落到可购买的改装硬件点击。 },
    ...
  ]
}
}

```

English Example

User:

[User Interaction History]:

[2023-10-25] [Product-Purchase] <|prod_begin|><a_5967><b_2249><c_743>

... (multi-domain timeline) ...

[2025-12-25] [Search] CarPlay-to-Android phone adapter box...

Target Interest Topic: From generic Sienna modification interest to specific component plan selection.

Generate the complete logic chain based on the history.

Assistant:

```

{
  "logic_chain": {
    "name": "From generic Sienna modification interest to specific component plan selection",
    "events": [
      { "date": "2025-12-21", "action": [Search] Sienna modification large-screen tutorial,
        "logic": Trigger stage: actively acquiring the Sienna modification path through tutorial searches,
        indicating a clear modification-direction need. },
      { "date": "2025-12-23", "action": [Search] Top-10 tips for Sienna owners,
        "logic": Cognitive complement: shifting from how-to toward risk and experience frameworks. },
      { "date": "2025-12-25", "action": [Product-Click] CarPlay-to-Android adapter box,
        "logic": Decision convergence: from tutorial/advice research down to a purchasable modification
        hardware click. },
      ...
    ]
  }
}

```

Prompt Bench.R2.3: Evolution Direct Generation (R2 — Evolution)

User:

【用户交互历史】：

【2024-01-22】 [商品-购买] <|prod_begin|><a_5175><b_1583><c_3381>

【2025-12-01】 [广告-点击] <|ad_begin|><a_6526><b_797><c_4992>

【2026-01-04】 [搜索] 美甲构建和不构建什么区别

... (无主题提示——模型须自主发现所有兴趣线索) ...

Assistant:

```

[
  { "logic_chain": {
    "name": "从美甲知识补课到工具/材料闭环采购的DIY落地链",
    "events": [
      { "date": "2026-01-04", "action": [搜索] 美甲构建和不构建什么区别,
        "logic": 问题触发：从工艺分叉点切入，属于概念级求解。 },
      { "date": "2026-01-15", "action": [视频-长播] <|video_begin|><a_7233><b_649><c_7927>,
        "logic": 知识补充：转向零基础步骤补齐执行路径。 },
    ]
  }
}

```

```

    { "date": "2026-01-16", "action": [商品-购买] <|prod_begin|><a_1356><b_5951><c_167>,
      "logic": 落地收敛：从认知到行动的闭环。 }
  ]
} },
{ ... 其他演化链... }
]

```

English Example

User:

[User Interaction History]:

```

[2024-01-22] [Product-Purchase] <|prod_begin|><a_5175><b_1583><c_3381>
[2025-12-01] [Ad-Click] <|ad_begin|><a_6526><b_797><c_4992>
[2026-01-04] [Search] difference between built-up vs. non-built-up manicure
... (no topic provided—the model must autonomously discover all interest threads) ...

```

Assistant:

```

[
  { "logic_chain": {
    "name": "From manicure-knowledge catch-up to closed-loop tool/material procurement, a DIY-execution chain",
    "events": [
      { "date": "2026-01-04", "action": [Search] difference between built-up vs. non-built-up manicure,
        "logic": Problem trigger: entry from the technique-bifurcation point, a concept-level inquiry. },
      { "date": "2026-01-15", "action": [Video-Long-View]
<|video_begin|><a_7233><b_649><c_7927>,
        "logic": Knowledge complement: shifting to zero-basis step-by-step execution paths. },
      { "date": "2026-01-16", "action": [Product-Purchase]
<|prod_begin|><a_1356><b_5951><c_167>,
        "logic": Execution convergence: closing the loop from cognition to action. }
    ]
  } },
  { ... additional evolution chains ... }
]

```

Note: The following examples illustrate the video target domain. The same task format applies equally to product, ad, and live target domains.

Prompt Bench.R3.1: Single-Domain Recommendation (R3 — Recommendation)

System: 你是一个推荐系统助手，擅长根据用户属性与视频历史行为预测用户的视频偏好。

User:

用户属性信息：这是一个女性用户，她关注的博主类型分布为：【亲子】占18.2%，【二次元】占9.54%，【游戏】占9.54%。用户点赞最多的类型是【美食、明星娱乐】，收藏最多的视频类型是【时尚、美妆】。近30日短剧消费情况：[逆袭_女性成长]看了17.1分钟...

用户视频域的历史行为：深度观看过的视频有<|video_begin|><a_4027><b_6474><c_6011>, <|video_begin|><a_631><b_5499><c_8167>, <|video_begin|><a_3433><b_1769><c_1441>... 请预测用户接下来可能感兴趣的视频。

Assistant:

```

<|video_begin|><a_4750><b_2230><c_4729>, <|video_begin|><a_5844><b_5861><c_361>,
<|video_begin|><a_3475><b_2493><c_1218>,
<|video_begin|><a_7092><b_5735><c_3928>...

```

English Example

System: You are a recommendation-system assistant, skilled at predicting the user's video preferences based on user attributes and historical video behavior.

User:

User Attributes: The user is female, and her followed-blogger type distribution is: [Parenting] 18.2%, [Anime] 9.54%, [Gaming] 9.54%. The user's most-liked categories are [Food, Celebrity Entertainment]; the most-collected video categories are [Fashion, Beauty]. Last-30-day short-drama consumption: [Counterattack_Female-Growth] 17.1 minutes...

User Video-Domain Behavioral History: deeply-watched videos include

<|video_begin|><a_4027><b_6474><c_6011>, <|video_begin|><a_631><b_5499><c_8167>, <|video_begin|><a_3433><b_1769><c_1441>...

Please predict the videos the user is likely to be interested in next.

Assistant:

<|video_begin|><a_4750><b_2230><c_4729>, <|video_begin|><a_5844><b_5861><c_361>, <|video_begin|><a_3475><b_2493><c_1218>, <|video_begin|><a_7092><b_5735><c_3928>...

Prompt Bench.R3.2: Cross-Domain Recommendation (R3 — Recommendation)

System: 你是一个推荐系统助手，擅长根据用户属性与多域历史行为预测用户的视频偏好。

User:

用户属性信息：这是一个女性用户。点赞最多的类型是【美食、明星娱乐】...

用户多域历史行为：

用户在直播域：关注了主播<|living_begin|><a_7490><b_3><c_2159>，

<|living_begin|><a_959><b_263><c_3562>。

用户在电商域：浏览过的商品有<|prod_begin|><a_4698><b_6949><c_7779>，

<|prod_begin|><a_7463><b_2719><c_4393>...

用户在广告域：点击过的广告有<|ad_begin|><a_3120><b_5580><c_1890>...

请基于多域行为预测用户在视频域接下来可能感兴趣的内容。

Assistant:

<|video_begin|><a_4750><b_2230><c_4729>, <|video_begin|><a_5844><b_5861><c_361>, <|video_begin|><a_3475><b_2493><c_1218>...

English Example

System: You are a recommendation-system assistant, skilled at predicting the user's video preferences based on user attributes and multi-domain historical behavior.

User:

User Attributes: The user is female. The most-liked categories are [Food, Celebrity Entertainment]...

User Multi-Domain Behavioral History:

Live-streaming domain: the user has followed streamers <|living_begin|><a_7490><b_3><c_2159>, <|living_begin|><a_959><b_263><c_3562>.

E-commerce domain: the user has browsed products <|prod_begin|><a_4698><b_6949><c_7779>, <|prod_begin|><a_7463><b_2719><c_4393>...

Advertising domain: the user has clicked ads <|ad_begin|><a_3120><b_5580><c_1890>...

Based on the multi-domain behavior, predict the content the user is likely to be interested in next in the video domain.

Assistant:

<|video_begin|><a_4750><b_2230><c_4729>, <|video_begin|><a_5844><b_5861><c_361>, <|video_begin|><a_3475><b_2493><c_1218>...

C. Pre-Training Details

C.1. Positive–Negative Similarity Margin Study

This appendix details the positive-negative similarity margin study summarized in Section 4.2. We examine whether the proposed pre-training corpora yield item representations that better separate semantically matched (positive) from unrelated (negative) captions.

Experimental Setup. We compare two checkpoints under a controlled protocol: the reproduced OpenOneRec serves as the *baseline*, while *ours* replaces only the pre-training corpora, holding the token budget, optimization schedule, and architecture fixed, thus isolating the effect of corpus composition. For each item, we probe two complementary representations: (i) the mean-pooled output of the *first embedding layer*, reflecting lexical-level priors in the learned token embeddings, and (ii) the last output of the *final hidden layer*, capturing the contextualized representation consumed by downstream heads. This lets us examine whether the effect of our corpora appears at the input-side, the task-relevant representation, or both. All similarities use cosine similarity in a shared representation space.

Metric Definition. Following common practice in metric learning, we summarize representational separation through the difference between matched and mismatched similarities rather than reporting them in isolation (Rusak et al., 2024; Lee et al., 2025). We define the *positive–negative similarity margin* for each anchor item i as

$$\text{margin}(i) = \cos(e_i^{\text{item}}, e_i^{\text{cap}}) - \cos(e_i^{\text{item}}, e_j^{\text{cap}}), \quad j \sim \text{Uniform}(\mathcal{D}_d \setminus \{i\}), \quad (40)$$

where e_i^{item} is the mean-pooled embedding of item i at the probed layer, e_i^{cap} is its ground-truth caption embedding, and e_j^{cap} is a caption sampled uniformly from the same domain \mathcal{D}_d . The definition is applied independently to both probes, with the negative term taken in expectation over j , so the reported quantity is $\mathbb{E}_j[\text{margin}(i)]$. The first term measures how tightly an item couples to its true caption, the second the residual similarity to an unrelated one. A larger $\text{margin}(i)$ indicates the model pushes the matched caption closer while keeping unrelated captions sufficiently far.

Table 24 | Mean cosine similarity of positive and negative item–caption pairs, and the resulting average margin Δ , computed under the baseline (the reproduced OpenOneRec in Section 4.3) and our model across four domains, at two probe depths: the first embedding layer and the final hidden layer. Larger Δ indicates stronger representational discriminability.

Probe	Domain	Baseline			Ours		
		Pos	Neg	Δ	Pos	Neg	Δ
First layer	Video	+0.043	+0.001	+0.042	+0.072	+0.001	+0.071
	Live	+0.021	0.000	+0.021	+0.043	−0.002	+0.045
	Product	+0.059	0.000	+0.059	+0.089	−0.001	+0.090
	Ad	+0.034	0.000	+0.034	+0.052	−0.001	+0.053
Final layer	Video	+0.012	0.000	+0.012	+0.009	−0.002	+0.011
	Live	+0.002	−0.002	+0.004	+0.007	−0.006	+0.013
	Product	+0.016	0.000	+0.016	+0.039	+0.001	+0.038
	Ad	+0.004	−0.001	+0.005	+0.017	0.000	+0.017

Quantitative Summary. Table 24 reports, for each domain and both probed layers, the mean positive and negative cosine similarities and the resulting margin $\Delta = \mathbb{E}_i[\text{margin}(i)]$ under the baseline and

our model. At the *first-layer* probe, our model widens the margin across all four domains (roughly 1.5–2×), showing that the proposed corpora inject stronger item–caption alignment at the input side. At the *final-layer* probe, the margins shrink for both models—as deeper representations become more contextualized—yet our model still improves clearly on Live, Product, and Ad (~ 2.4–3.4×). The only exception is Video (0.012 vs. 0.011), where the baseline is already well aligned at depth; this is consistent with the near-identical Video results in the Cross-Domain Reco block of Table 2.

C.2. Data Mixture for Pre-Training

Section 4.2 described the construction of our pre-training corpora. In this section, we summarize the overall data mixture used throughout pre-training, where Table 26 and Table 25 report the token ratio of each dataset. Echoing the progressive multi-granularity modality-alignment principle of Section 4.2, the mixture comprises two complementary parts: **recommendation corpora**, which carry the dominant modality-alignment signal, and **general-domain corpora**, which preserve the model’s general competencies.

Recommendation Corpora. This part comprises the four-granularity corpora—*token*, *item*, *relational*, and *user* granularity—whose construction is detailed in Section 4.2. Their datasets and token ratios are listed in Table 25.

General-Domain Corpora. As described in Section 4.2, this part consists of two subsets: *general-domain text data*, spanning mathematical reasoning, code, general reasoning, and medical corpora together with in-house text data, and *multimodal data*, which further comprises image generation and image editing data, in-house image captions, and in-house interleaved image-text corpus. All the publicly available datasets are downloadable from the HuggingFace repository (see footnotes). The corresponding data mixtures are listed in Table 26.

Table 25 | Data mixtures for pre-training (Recommendation Corpora).

Category	Modality	Dataset	Token Ratio (%)
Token Granularity	Itemic +Text	Single-Token-Semantic-Prediction	0.35
		Compositional Prefix Semantic Prediction	1.29
		Prefix Itemic Token Grounding	0.55
		Part-to-Whole Semantic Prediction	0.31
		Subtotal	2.50
Item Granularity	Itemic +Text	Item Captioning	17.28
		Item Grounding	12.12
		Item QA	2.44
		Subtotal	31.84
Relational Granularity	Itemic +Text	Direct-Item-to-Item Associations	0.24
		Multi-Item Interest Flows	2.94
		Subtotal	3.18
User Granularity	Itemic +Text	Domain-Grouped Behavior Sequence	21.41
		Chronologically-Interleaved Behavior Sequence	12.64
		Subtotal	34.05

Table 26 | Data mixtures for pre-training (General-Domain Corpora).

Category	Modality	Dataset	Token Ratio (%)
General Domain	Text	In-House-Text-Data	8.91
		Nemotron-CC-Math-v1 ¹	4.25
		reasoning-v1-20m ²	3.89
		Nemotron-CC-v2 ³	3.26
		Nemotron-Pretraining-Code-v1 ⁴	2.35
		OpenMathReasoning ⁵	1.04
		OpenCodeReasoning ⁶	0.39
		KodCode-V1-SFT-R1 ⁷	0.15
		Chinese-Reasoning-Distil-Data ⁸	0.10
		Bespoke-Stratos-17k ⁹	0.04
		medical-o1-reasoning-SFT ¹⁰	0.02
		MegaMath ¹¹	0.78
		OpenMathInstruct-2, NuminaMath-CoT & MMIQC ¹²¹³¹⁴	0.49
		Infinity-Instruct ¹⁵	0.44
		FineVision ¹⁶	0.18
		GeneralThought-430K ¹⁷	0.07
		Multi-subject-RLVR ¹⁸	0.04
		Big-Math-RL-Verified ¹⁹	0.08
		rStar-Coder ²⁰	0.05
		CuratedThoughts ²¹	0.26
		DeepMath-103K ²²	0.05
		DAPO-Math-17k ²³	0.01
		Subtotal	
Itemic +Text		Image Generation and Image Edit ²⁴²⁵²⁶²⁷	0.29
		Image Captions(In-house and open-source ²⁸)	1.10
		In-house Interleave Corpora	0.20
		Subtotal	

C.3. Four-Granularity Pre-Training Samples

The following are representative training samples from each of the four granularities, illustrating the data formats described in Section 4.2.

Token Granularity: Compositional Prefix Semantic Prediction

中文示例

User: 短视频域中，<a_3664><b_3076> 表示什么？

Assistant:

这些短视频在内容主题上高度聚焦于“都市夜景中的个体孤独与情感内省”。画面通常选取夜晚的城市街道、霓虹灯下或夜市背景，核心视觉元素为背对镜头的单人（多为背影），通过人物与繁华或静谧环境的对比，营造一种疏离、沉思或略带忧郁的氛围。视频往往借助文字叠加来外化人物内心的挣扎、对过去的追忆、对情感的执着或对生活的感悟，将物理空间的“夜”转化为心理空间的“独白”。

从受众角度来看，这类内容主要吸引那些在快节奏都市生活中感到孤独、压力或情感共鸣的观众。目标人群倾向于通过观看此类视频寻找情感慰藉、确认自身情绪，或在静谧的视觉与文字中寻求片刻的内心平静与精神寄托。观众画像多为年轻至中年群体，对情感类、治愈系或略带伤感氛围的都市生活记录有较高偏好。

English Translation

User: In the short-video domain, what does <a_3664><b_3076> represent?

Assistant:

These short videos are highly focused on the theme of “individual loneliness and emotional introspection in urban night scenes.” The frames typically feature nighttime city streets, under neon lights, or night market backgrounds, with the core visual element being a single person (mostly a back view) facing away from the camera. The contrast between the character and the bustling or tranquil environment creates a sense of alienation, contemplation, or a slightly melancholic atmosphere. The videos often externalize the character’s inner struggles, reminiscence of the past, persistence in emotions, or reflections on life through overlaid text, transforming the physical space of “night” into a psychological space of “inner monologue.”

¹<https://huggingface.co/datasets/nvidia/Nemotron-CC-Math-v1>

²<https://huggingface.co/datasets/glaiveai/reasoning-v1-20m>

³<https://huggingface.co/datasets/nvidia/Nemotron-CC-v2>

⁴<https://huggingface.co/datasets/nvidia/Nemotron-Pretraining-Code-v1>

⁵<https://huggingface.co/datasets/nvidia/OpenMathReasoning>

⁶<https://huggingface.co/datasets/nvidia/OpenCodeReasoning>

⁷<https://huggingface.co/datasets/KodCode/KodCode-V1-SFT-R1>

⁸<https://huggingface.co/datasets/Mxode/Chinese-Reasoning-Distil-Data>

⁹<https://huggingface.co/datasets/bespokelabs/Bespoke-Stratos-17k>

¹⁰<https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT>

¹¹<https://huggingface.co/datasets/LLM360/MegaMath>

¹²<https://huggingface.co/datasets/nvidia/OpenMathInstruct-2>

¹³<https://huggingface.co/datasets/AI-MO/NuminaMath-CoT>

¹⁴<https://huggingface.co/datasets/Vivacem/MMIQC>

¹⁵<https://huggingface.co/datasets/BAAI/Infinity-Instruct>

¹⁶<https://huggingface.co/datasets/HuggingFaceM4/FineVision>

¹⁷<https://huggingface.co/datasets/GeneralReasoning/GeneralThought-430K>

¹⁸<https://huggingface.co/datasets/virtuoussy/Multi-subject-RLVR>

¹⁹<https://huggingface.co/datasets/SynthLabsAI/Big-Math-RL-Verified>

²⁰<https://huggingface.co/datasets/microsoft/rStar-Coder>

²¹<https://huggingface.co/datasets/bethgelab/CuratedThoughts>

²²<https://huggingface.co/datasets/zwhe99/DeepMath-103K>

²³<https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math-17k>

²⁴<https://huggingface.co/datasets/bytedance-research/UNO-1M>

²⁵<https://huggingface.co/datasets/BLIP3o/BLIP3o-60k>

²⁶<https://huggingface.co/datasets/Bin1117/AnyEdit>

²⁷<https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions>

²⁸<https://huggingface.co/datasets/BAAI/Infinity-MM>

From an audience perspective, such content mainly attracts viewers who feel loneliness, stress, or emotional resonance in fast-paced urban life. The target group tends to seek emotional solace, validate their own feelings through watching such videos, or find a moment of inner peace and spiritual sustenance in the quiet visuals and text. The viewer profile is largely young to middle-aged groups with a high preference for emotional, healing, or slightly sentimental urban life records.

Item Granularity: Caption-Aware Caption Coarse-Graining)

中文示例

User: 请描述一下这个电商商品<|prod_begin|><a_2147><b_7978><c_5031>。

Assistant:

该商品是一款户外软壳冲锋衣，一级类目是运动户外鞋服，二级类目是户外服装，三级类目是软壳衣裤。采用连帽拉链设计，软壳材质兼具防风防水和加绒保暖功能，标准版型，加绒内里，多色可选，简约时尚，适合户外活动及日常休闲。价格区间为100-120元。无品牌。

English Translation

User: Please describe this e-commerce product <|prod_begin|><a_2147><b_7978><c_5031>.

Assistant:

This product is an outdoor softshell jacket. Its primary category is Sports & Outdoor Footwear and Apparel, secondary category is Outdoor Clothing, and tertiary category is Softshell Jackets & Pants. It features a hooded zipper design, with the softshell material providing windproof, waterproof, and fleece-lined thermal functions. It has a standard fit, fleece-lined interior, multiple color options, and a minimalist, stylish look, suitable for outdoor activities and daily casual wear. The price range is 100-120 yuan. Unbranded.

Relational Granularity: Multi-Item Interest Flow

中文示例

<|living_begin|><a_4515><b_6234><c_6278> → 关注长发美女和马甲线身材的用户，往往对形体管理与健康生活方式感兴趣。直播中展现的性感形体和生活化互动，容易引发对专业健身指导的需求，而内容中刘教练的腿部训练视频，正好满足了他们对科学塑形和健康提升的追求。→

<|video_begin|><a_458><b_3589><c_5563> → 看过腿部锻炼指导视频的人，往往对提升肌肉力量和锻炼方式感兴趣，而普拉提圈作为辅助训练工具，正好能满足他们对多样化训练器材的需求，因此容易被吸引。→ <|prod_begin|><a_1162><b_1966><c_7428> → 关注普拉提圈训练的用户，往往也在寻找适合中老年人的肌肉锻炼工具。他们在观看腿部锻炼教学视频时，接触到类似瑜伽圈、抗阻圈的产品，进而对专为大腿设计的抗阻训练器产生兴趣。→ <|prod_begin|><a_7433><b_6972><c_1162>

English Translation

<|living_begin|><a_4515><b_6234><c_6278> → Users following long-haired beauties and toned abs often care about body shaping and healthy lifestyles. The attractive physique and casual interactions in the livestream can trigger a need for professional fitness guidance, and the leg training video by Coach Liu exactly fulfills their pursuit of scientific body sculpting and health improvement. → <|video_begin|><a_458><b_3589><c_5563> → Those who watched leg exercise tutorials are often interested in improving muscle strength and workout variety, and the Pilates ring, as an auxiliary training tool, precisely meets their demand for diverse training equipment, thus easily capturing their attention. → <|prod_begin|><a_1162><b_1966><c_7428> → Users interested in Pilates ring exercises are also often looking for muscle-building tools suitable for middle-aged and elderly people. When watching leg training videos, they encounter products similar to yoga rings and resistance bands, and further develop interest in resistance trainers designed specifically for thighs. → <|prod_begin|><a_7433><b_6972><c_1162>

User Granularity: Domain-Grouped Behavior Sequences

中文示例

User: 已知用户信息如下：用户专注于6个时尚领域的内容创作。用户安装了小红书-社区、支付宝-支付...该用户可能购买了什么商品？输出JSON，使用字段名"purchase"。

Assistant: {"purchase": [{"<|prod_begin|><a_7552><b_5608><c_912>", "<|prod_begin|><a_7651><b_4552><c_4508>", "<|prod_begin|><a_7618><b_4552><c_4508>"]}]

User: <|ad_begin|><a_7112><b_7554><c_6156>, <|ad_begin|><a_4597><b_4106><c_3984>, <|ad_begin|><a_1169><b_3795><c_3984> 是该用户的部分记录，请继续给出其他点击广告。

Assistant: ['<|ad_begin|><a_7586><b_7762><c_7321>', '<|ad_begin|><a_1512><b_2007><c_5218>']

User: 以下哪些商品是该用户浏览过的？（用分号分隔答案）

- A. <|prod_begin|><a_7278><b_487><c_3329>
- B. <|prod_begin|><a_7698><b_4069><c_6585>
- C. <|prod_begin|><a_5891><b_4945><c_335>

Assistant: A; C

User: 根据广告和购买记录，该用户最近可能感兴趣的视频有哪些？请返回一个list。

Assistant: ['<|video_begin|><a_5991><b_2799><c_6278>', '<|video_begin|><a_4751><b_765><c_2584>']

English Translation

User: Given the following user info: The user focuses on creating content in 6 fashion-related domains. The user has installed Xiaohongshu-Community, Alipay-Payment... What products might this user have purchased? Output JSON with the field name "purchase".

Assistant: {"purchase": [{"<|prod_begin|><a_7552><b_5608><c_912>", "<|prod_begin|><a_7651><b_4552><c_4508>", "<|prod_begin|><a_7618><b_4552><c_4508>"]}]

User: <|ad_begin|><a_7112><b_7554><c_6156>, <|ad_begin|><a_4597><b_4106><c_3984>, <|ad_begin|><a_1169><b_3795><c_3984> are part of this user's records. Please continue and provide other clicked ads.

Assistant: ['<|ad_begin|><a_7586><b_7762><c_7321>', '<|ad_begin|><a_1512><b_2007><c_5218>']

User: Which of the following products has this user browsed? (Separate answers with semicolons)

- A. <|prod_begin|><a_7278><b_487><c_3329>
- B. <|prod_begin|><a_7698><b_4069><c_6585>
- C. <|prod_begin|><a_5891><b_4945><c_335>

Assistant: A; C

User: Based on the ad and purchase records, what videos might this user be recently interested in? Return a list.

Assistant: ['<|video_begin|><a_5991><b_2799><c_6278>', '<|video_begin|><a_4751><b_765><c_2584>']

D. SFT Details**D.1. SFT Data and Itemic-Token Perception****D.1.1. R0 Item-Perception Examples**

R0 trains the base model to map between itemic tokens and natural-language captions, answer content-grounded questions, and recover itemic tokens from descriptions. The examples below are sanitized bilingual cases that preserve the observed task formats while replacing raw identifiers with synthetic three-token examples. They cover short-video, product, live-streaming, and advertising domains, and include both CoT and unCoT formats.

Prompt SFT.R0.1: CoT Itemic-Token-to-Caption Generation

中文示例

System: 你是一名专业的视频内容理解与描述助手，请根据短视频itemic tokens 生成准确的视频描述。

User: 请描述短视频<|video_begin|><a_2344><b_3729><c_6164> 的视频内容。

Assistant:

<think>从<a_2344> 看，该视频属于线下休闲娱乐和商场体验类内容；<b_3729> 进一步指向新奇互动玩法，而不是普通购物或餐饮记录；<c_6164> 支持“悬挂式抓零食”这一具体场景，核心是参与者被安全绳吊起，在零食池中抓取零食并制造轻松搞笑效果。三层itemic tokens 的语义交集指向一种面向年轻消费者的商场趣味体验活动。</think>

视频记录了一名女性在商场内体验悬挂式抓零食游戏的过程。参与者被安全绳吊在装满零食的区域中，通过身体动作抓取零食，旁边有工作人员协助。画面字幕强调单次体验价格和“抓零食”的趣味性，整体呈现为轻松、搞笑、适合线下娱乐分享的短视频。

English Example

System: You are a professional video-understanding assistant. Generate an accurate caption from short-video itemic tokens.

User: Describe the content represented by <|video_begin|><a_2344><b_3729><c_6164>.

Assistant:

<think>The coarse token <a_2344> points to offline entertainment and mall experiences; <b_3729> narrows it to a playful interactive activity rather than ordinary shopping or dining; and <c_6164> supports the specific scene of a suspended snack-grabbing game. Together, the itemic tokens describe a novelty mall activity for young consumers.</think>

The video records a woman trying a suspended snack-grabbing game inside a shopping mall. The participant is lifted by a safety harness above a snack-filled area and tries to grab snacks with body movements while staff assist nearby. On-screen captions emphasize the price and the fun of the game, making the clip a light, humorous record of an offline entertainment experience.

Prompt SFT.R0.2: unCoT Itemic-Token-to-Caption Generation

中文示例

System: 你是一位直播理解专家，擅长将主播itemic tokens 转换为自然的直播描述。

User: 已知主播<|living_begin|><a_522><b_6604><c_512>，请说明这位主播通常会直播什么内容。

Assistant:

该主播定位为音乐轮播类无人直播博主，以虚拟直播间为载体，主打流行歌曲、古风歌曲和轻量DJ音效的持续播放。直播中通常没有真人强互动，更多承担背景音和陪伴功能，适合观众在休息、睡前或碎片化时间收听。整体风格简约、节奏感强，主要吸引喜欢听歌放松、需要背景音陪伴的年轻用户。

English Example

System: You are a live-streaming understanding expert. Convert streamer itemic tokens into a natural live-room description.

User: Given the streamer itemic tokens <|living_begin|><a_522><b_6604><c_512>, describe what this streamer usually broadcasts.

Assistant:

This streamer is best described as an automated music-broadcast host. The live room mainly plays pop songs, classical-style tracks, and light DJ effects through a virtual setup. It has little direct human interaction and works more as background audio or emotional companionship, suitable for rest time, bedtime, or casual listening. The style is simple and rhythmic, attracting young users who want music for relaxation or ambient company.

Prompt SFT.R0.3: Product QA

中文示例

System: 你是一名商品理解助手，能够根据商品itemic tokens 和描述回答问题。

User: 商品<|prod_begin|><a_7516><b_5044><c_7205> 是一款700W 电动抛光机，采用全铜电机和无级变速设计，可用于汽车打蜡、封釉、划痕修复，也可用于皮具保养和木材打磨。这个商品最适合下面哪类使用场景？

A. 室内空气净化 B. 汽车美容与表面抛光 C. 儿童绘本阅读 D. 生鲜冷链配送

请只回答A/B/C/D。

Assistant: B

English Example

System: You are a product-understanding assistant. Answer questions based on product itemic tokens and descriptions.

User: Product <|prod_begin|><a_7516><b_5044><c_7205> is a 700W electric polisher with a copper motor and stepless speed control. It can be used for car waxing, glazing, scratch repair, leather care, and wood polishing. Which usage scenario best matches this product?

A. Indoor air purification B. Car detailing and surface polishing C. Children's picture-book reading
D. Fresh-food cold-chain delivery

Answer only A/B/C/D.

Assistant: B

Prompt SFT.R0.4: Caption-to-Itemic-Token Grounding

中文示例

System: 请根据输入的广告描述，输出能与其语义最匹配的广告itemic tokens。

User: 请根据以下广告内容描述生成匹配的广告itemic tokens：视频为生活技巧类广告，通过真人演示卫裤绳子的多种系法，展示打结、蝴蝶结、松紧节等做法，重点解决裤绳容易缩回裤腰的问题，并用限时补贴弹窗引导用户点击领取福利。

Assistant:

<|ad_begin|><a_2350><b_6885><c_2112>

English Example

System: Given an advertising description, output the semantically matching advertising itemic tokens.

User: Generate the matching advertising itemic tokens for this ad description: the video is a life-hack advertisement that demonstrates several ways to tie sweatpants drawstrings, including knots, bows, and adjustable loops. It focuses on preventing the drawstring from slipping back into the waistband and uses a limited-time subsidy pop-up to guide users to claim a benefit.

Assistant:

<|ad_begin|><a_2350><b_6885><c_2112>

D.1.2. SFT Data Composition

Table 27 | Data mixture for the SFT stage. Sample counts in units of 10^4 ; ratio in %.

Category	Dataset	Samples ($\times 10^4$)	Ratio (%)
R0: Perception	Short-Video-Caption	20.00	5.06
	Commercial-Ad-Caption	16.98	4.30
	Live-Stream-Caption	13.70	3.47
	E-Commerce-Caption	17.52	4.43
	Short-Video-QA	24.09	6.10
	E-Commerce-QA	1.84	0.47
	Subtotal	94.13	23.82
R1: Derivation	i2i-TagNext-CF	20.00	5.06
	i2i-Watch-then-Search	20.00	5.06
	Subtotal	40.00	10.12
R2: Evolution	Multi-Hop-Reasoning	12.38	3.13
R3: Recommendation	Cross-Domain (CoT)	29.56	7.48
	Cross-Domain (UnCoT)	58.80	14.88
	Subtotal	88.36	22.36
Itemic Instruction Data	Item-Instruction-Following	10.30	2.61
General-Domain Data	StepFun-General ²⁹	150.00	37.96
	Total	395.17	100.00

D.2. Cognitive-Reasoning Data Construction

D.2.1. R1 Explicit-Relevance Judgement Prompt

The first relation-quality stage in R1 uses a locally deployed flagship model to define the boundary between explicit relevance, explicit irrelevance, and uncertainty. The prompt is applied to both TagNex-derived and after-play-search item-to-item candidates, using the available item metadata and dense captions as evidence.

Prompt SFT.R1.1: R1 Explicit-Relevance Judgement Prompt

System Prompt

你是一名推荐关系判断专家。你的任务是根据两个候选内容的可见信息，判断它们是否存在直接、稳定、可解释的一跳关系。请只依据给定的内容信息，不要使用“同一人群可能喜欢”“同平台常见共现”等宽泛假设。最终只输出三类之一：显性相关、显性不相关、不确定。

User Prompt Template

请判断下面两个内容之间是否存在直接、稳定、可解释的一跳关系。

【源内容信息】：{src_metadata}

【目标内容信息】：{dest_metadata}

²⁹<https://huggingface.co/datasets/stepfun-ai/Step-3.5-Flash-SFT>

【源内容描述】: {src_dense_caption}
 【目标内容描述】: {dest_dense_caption}

判断标准:

- 1) 如果两者在主题、场景、对象、任务、风格、内容形态或近邻互补关系上存在清晰的一跳连接，输出“显性相关”。
- 2) 如果两者主题、场景或对象明显不同，或只能依赖宽泛人群偏好、平台共现、热度相近等假设建立联系，输出“显性不相关”。
- 3) 如果给定信息不足、描述噪声较高，或关系无法从可见内容中稳定确认，输出“不确定”。

请注意:

- 1) 不要把多跳推理或抽象联想当作直接关系。
- 2) 不要把“可能同一批用户会看”当作相关依据。
- 3) 只输出一个标签: 显性相关、显性不相关或不确定。
- 4) 不要输出解释、示例、标点或其他额外内容。

English Translation

System Prompt

You are a recommendation-relation judgement expert. Your task is to determine, from the visible information of two candidate items, whether they have a direct, stable, and explainable one-hop relation. Use only the provided content evidence, and do not rely on broad assumptions such as “the same audience may like both” or “they often co-occur on the platform”. Output exactly one of three labels: explicitly related, explicitly unrelated, or uncertain.

User Prompt Template

Please judge whether the following two pieces of content have a direct, stable, and explainable one-hop relation.

Source content information: {src_metadata}

Target content information: {dest_metadata}

Source content description: {src_dense_caption}

Target content description: {dest_dense_caption}

Judgement criteria:

- 1) If the two items have a clear one-hop connection in topic, scenario, object, task, style, content form, or neighboring complementarity, output “explicitly related”.
- 2) If their topic, scenario, or object is clearly different, or the connection only relies on broad assumptions such as audience overlap, platform co-occurrence, or similar popularity, output “explicitly unrelated”.
- 3) If the provided information is insufficient, the descriptions are noisy, or the relation cannot be stably confirmed from the visible content, output “uncertain”.

Notes:

- 1) Do not treat multi-hop reasoning or abstract association as a direct relation.
- 2) Do not use “the same users may watch both” as relevance evidence.
- 3) Output exactly one label: explicitly related, explicitly unrelated, or uncertain.
- 4) Do not output explanations, examples, punctuation, or any extra text.

D.2.2. R1 Relation-Explanation Generation Prompt

After explicit relation judgement and bridge-variable extraction, R1 generates the final relation explanation used in recommendation SFT samples. The prompt below is applied with source-item evidence and abstract bridge variables, while the destination item is used only for consistency checking.

Prompt SFT.R1.2: Relation-Explanation Generation Prompt**System Prompt**

你是一名推荐分析师。你需要根据源视频线索和抽象桥接变量，以第一人称输出一段自然、可信、可解释的中文推理。候选目标信息仅用于内部校验，不能在结果里显式提及。

User Prompt Template

给定源视频信息和一组抽象桥接变量，请你以推荐分析师视角，输出一段自然流畅的第一人称中文推理，说明用户为什么会从当前内容继续延伸到某个相邻内容方向。候选目标信息仅用于内部校验，你不能在结果里显式提到目标，也不能把目标表层内容逐条复述成答案。

源视频信息:

```
- src_pid: {src_pid}
- src_tag: {src_tag}
- src_dense_caption: {src_dense_caption}
```

抽象桥接变量:

```
- source_need: {source_need}
- bridge_type: {bridge_type}
- bridge: {bridge}
- direction: {direction}
- reason_seed: {reason_seed}
```

参考候选目标（仅内部校验，不可显式提及）:

```
- dest_pid: {dest_pid}
```

输出要求:

1. 只输出最终推理正文，不要输出标题、编号、Markdown、引号或额外说明。
2. 必须使用第一人称分析口吻，且正文必须以“我注意到”开头。
3. 核心依据必须来自源视频信息和抽象桥接变量；候选目标只用于内部一致性校验。
4. 重点说明源内容激活了什么需求、这种需求为什么会延伸、后续可能会滑向什么方向，不要简单复述源描述。
5. 不要显式提及目标、候选、dest_pid，也不要目标表层关键词直接写进正文。
6. 结尾保持非确定性，例如“可能会继续关注”“更可能延伸到”“有较大概率被这类内容接住”。
7. 正文长度控制在220到520字之间，信息密度高但自然。

你的推理:

English Translation**System Prompt**

You are a recommendation analyst. Based on source-video signals and abstract bridge variables, output a natural, credible, and explainable Chinese reasoning paragraph in the first person. The candidate target is used only for internal consistency checking and must not be explicitly mentioned in the result.

User Prompt Template

Given source-video information and a set of abstract bridge variables, output a fluent first-person Chinese reasoning paragraph from the perspective of a recommendation analyst, explaining why the user may extend from the current content to a neighboring content direction. The candidate target is used only for internal consistency checking; do not explicitly mention the target or restate its surface content as the answer.

Source-video information:

```
- src_pid: {src_pid}
- src_tag: {src_tag}
- src_dense_caption: {src_dense_caption}
```

Abstract bridge variables:

```
- source_need: {source_need}
- bridge_type: {bridge_type}
- bridge: {bridge}
- direction: {direction}
```

- reason_seed: {reason_seed}

Reference candidate target, used only for internal consistency checking and not to be explicitly mentioned:

- dest_pid: {dest_pid}

Output requirements:

1. Output only the final reasoning paragraph, without titles, numbering, Markdown, quotation marks, or extra explanation.
2. Use a first-person analytical tone, and the paragraph must start with “我注意到”.
3. The core evidence must come from the source-video information and abstract bridge variables; the candidate target is only for internal consistency checking.
4. Focus on what need is activated by the source content, why the need extends, and what direction the user may move toward next; do not simply restate the source description.
5. Do not explicitly mention the target, candidate, dest_pid, or directly copy surface keywords from the target content.
6. Keep the ending non-deterministic, such as “may continue to follow”, “is more likely to extend to”, or “has a high probability of being captured by this type of content”.
7. Keep the paragraph between 220 and 520 Chinese characters, dense but natural.

Your reasoning:

D.2.3. R1 Training Examples

Prompt SFT.R1.3: Additional Sanitized R1 Training Examples

Example A: Training and Recovery

Source topic: 短期增肌训练 Target topic: 避免肌肉劳损

User

和输入视频<|video_begin|><a_1842><b_7365><c_2091>常识上相似的视频有哪些?

Assistant

<think>当用户对短期增肌训练产生浓厚兴趣时，由于抗阻训练通常伴随较高强度和较高频次，肌肉疲劳与劳损风险也会随之增加。从常识角度看，这类用户往往不仅关注如何提升训练效率，也会主动寻找科学的恢复策略，以避免训练损伤并保持后续训练表现。睡眠是肌肉修复和炎症缓解的重要阶段，而合理的睡姿能够减少局部压力和夜间酸痛，帮助身体更好地恢复。因此，用户的兴趣很容易从“如何增肌”延伸到“如何减少训练后的身体负担”。这种关联来自训练与恢复之间的协同关系，符合健身人群对健康管理 and 持续训练能力的共同需求。</think>

<|video_begin|><a_6219><b_0487><c_7742>

English Translation

<think>When a user shows strong interest in short-term muscle-gain training, resistance training usually implies high intensity and repeated load, which also increases the risk of muscle fatigue and strain. From common sense, such users care not only about improving training efficiency, but also about scientific recovery strategies that help avoid injury and maintain later performance. Sleep is an important stage for muscle repair and inflammation relief, and an appropriate sleeping posture can reduce local pressure and nighttime soreness. Therefore, the user's interest can naturally extend from “how to gain muscle” to “how to reduce physical burden after training”. This relation comes from the coordination between training and recovery, matching the fitness audience's need for health management and sustainable training.</think>

<|video_begin|><a_6219><b_0487><c_7742>

Example B: Off-Road Driving and Vehicle Suspension

Source topic: 泥泞地形越野驾驶 Target topic: 汽车减震改装

User

请推荐一条与短视频<|video_begin|><a_4027><b_1186><c_9304>主题接近、语义相关的视频。

Assistant

<think>我注意到源内容聚焦于车辆在泥泞等复杂地形中的越野表现，核心吸引点并不只是驾驶画面本身，而是车辆如何在恶劣路况下保持通过性、稳定性和操控感。对这类内容感兴趣的用戶，通常会进一步关注影响越野表现的关键部件，例如悬挂系统、减震能力和底盘调校。因为在复杂路面中，车辆能否稳定通过坑洼、泥地或颠簸路段，很大程度上取决于减震结构对冲击的吸收和车身姿态的控制。因此，用户的兴趣很自然会从“看车辆越野”延伸到“理解车辆为什么能越野”，进而关注改装方案和性能优化。这种关系属于从真实驾驶场景到车辆性能机制的自然延伸，有较大概率被相关的减震改装内容接住。</think>

</video_begin|><a_7750><b_3641><c_5088>

English Translation

<think>I notice that the source content focuses on how a vehicle performs off-road in muddy and complex terrain. The core attraction is not only the driving scene itself, but also how the vehicle maintains passability, stability, and controllability under difficult road conditions. Users interested in such content are likely to further care about key components that affect off-road performance, such as the suspension system, damping capability, and chassis tuning. On complex roads, whether a vehicle can pass pits, mud, and bumps smoothly depends heavily on how the damping structure absorbs impact and controls body posture. Therefore, the user's interest can naturally move from "watching off-road driving" to "understanding why the vehicle can handle off-road conditions", and then to modification plans and performance optimization. This is a natural extension from a driving scenario to the vehicle-performance mechanism.</think>

</video_begin|><a_7750><b_3641><c_5088>

Example C: Athlete and Event Continuation

Source topic: 孙颖莎 Target topic: 刘国梁

User

和输入视频</video_begin|><a_3196><b_8572><c_4409>常识上相似的视频有哪些?

Assistant

<think>我注意到源内容围绕孙颖莎在重要赛场中的表现展开，用户被吸引的往往不只是单场比赛结果，而是运动员在高强度竞争中的状态、成长和后续发展。对于关注奥运健儿的用户来说，兴趣很容易从运动员个人表现延伸到国家队训练安排、赛事衔接和教练团队判断。常识上，奥运后的运动员通常需要快速进入新的备战周期，而教练或管理者对赛事难度、训练节奏和竞技状态的判断，会直接影响用户理解运动员后续表现的方式。因此，刘国梁围绕全运会、训练压力或备战节奏的相关内容，能够承接用户对孙颖莎后续发展的关注。这种关联来自运动员表现与团队指导之间的赛事承接关系，属于体育内容中较自然的因果延伸。</think>

</video_begin|><a_9038><b_2164><c_6871>

English Translation

<think>I notice that the source content centers on Sun Yingsha's performance in important competitions. Users are often attracted not only by one match result, but also by the athlete's condition, growth, and subsequent development under high-intensity competition. For users who follow Olympic athletes, interest can naturally extend from the athlete's individual performance to national-team training plans, event transitions, and coaching-team judgement. In common sense, athletes often need to quickly enter a new preparation cycle after the Olympics, and the coach or manager's view on event difficulty, training rhythm, and competitive condition directly shapes how users understand the athlete's later performance. Therefore, content around Liu Guoliang's comments on the National Games, training pressure, or preparation rhythm can naturally continue the user's attention to Sun Yingsha's development. This relation comes from event continuation between athlete performance and team guidance.</think>

</video_begin|><a_9038><b_2164><c_6871>

D.2.4. R2 Logic-Chain Extraction Prompt

R2 begins by extracting candidate user-interest evolution chains from chronological multi-domain behavior timelines. We use a strong LLM to identify a small number of temporally ordered chains with clear progression, while avoiding shallow topical grouping or unsupported speculation. Below we provide the main extraction prompt used in this stage. The subsequent quality-control prompt follows

a highly similar style, but is used as a second-pass LLM-as-a-Judge filter to remove pseudo-logical or weakly grounded chains produced during first-pass generation.

Prompt SFT.R2.1: Logic-Chain Extraction Prompt

Prompt Template

角色定义：你是一位极端严苛的用户行为数据挖掘专家，负责从琐碎日志中提取具有“认知跃迁”价值的极少数行为链。

任务目标：请根据提供的用户交互历史（*Timeline*），其中包含了用户搜索内容，互动过的视频、直播、广告和商品。我希望你从中抽取有兴趣演化轨迹、具有内在行为关联逻辑而非内容简单关联的交互行为逻辑链，你的首要要求是剔除伪逻辑，严禁强行解释。

交互说明：其中商品点击、广告点击只是点击行为而非购买行为。

其他说明：对于“-:”标识的事件，表示对应的事件缺少具体的时间信息。

案例：比如“喉咙痛怎么办（20260101） 干咳/有痰区别（20260102） 用药科普（注意禁忌）（20260103）”每一个->都需要存在递进/转折或者其他明显的兴趣演化关系

行为逻辑链有效性准则：

深度关联：严禁简单的内容关键词堆砌。必须体现递进、转折、因果、对比或闭环等演化关系。严禁把“同类目”的事情按时间排在一起，然后强行解释出一种递进感；

时序敏感：严格遵守时间线，确保每一链条内的事件是按时间顺序发生的。

1. 顺序互换测试：如果链条中的Event A 和Event B 互换时间顺序后，逻辑依然通顺且合理，则不能将其视为强关联的演化链，请予以剔除或拆分为独立事件。

2. 有效演化：只有当Event B 的发生是基于Event A 带来的**反馈、瓶颈、认知升级**时（即B 是对A 的修正或深化），才算有效。

3. 认知增量测试：Event B 必须包含Event A 中未出现的新变量（如特定参数、专业术语、对比维度）。判定标准：如果Event B 只是Event A 的语义重复或同层级平移（例如：从搜“降噪耳机”变为搜“头戴耳机”），则不视为演化。

4. 行为修正/收敛特征：演化链必须体现出从“发散”到“聚焦”，或从“错误尝试”到“路径修正”的过程。判定标准：后续行为必须表现出对前期信息的筛选过滤。如果用户在Event B 中完全抛弃了Event A 的筛选条件且无逻辑解释，则链条断裂。有效示例：点击“高热量猫粮”-观看“猫咪软便调理视频”-搜索“低敏/无谷/易消化猫粮”。（体现了基于负反馈的策略调整）

5. 真正的兴趣演化往往伴随着交互深度的变化（如：从“被动接收”广告/视频，转为“主动检索”核心信息）如果全链条只有单一动作类型（如全是搜索），其递进逻辑往往较弱。严禁仅由“连续点击相似商品”构成链条。

6. 时空密度测试：严格审查事件A 到B 之间的时间跨度。如果两个核心行为跨度中间缺乏相关性的引导行为（如中间没有持续的相关搜索或视频观看），应判定为日常零散行为而非“演化”。严禁将相隔数月的日常消费行为（如买菜、买日用品）强行拼凑成认知升级。

7. 非普适性测试：排除基于“生活补货”或“随机消费”驱动的行为。如果Event A 是普适性的生活必需品（如普通的蛋奶、纸巾、基础款服饰），除非其具有极强的特定功能指向（如：从普通奶粉转向专门针对糖尿病的脱糖奶粉），否则不得作为逻辑链的起始点或关键节点。”

8. 触发源测试：逻辑链必须包含一个清晰的“问题触发”。有效的演化链通常应以“主动搜索”或“带有明确意图的对比行为”为起点。严禁直接由两个不同的商品点击/购买行为推导出认知升级，除非中间有明确的知识获取动作（如观看科普、对比参数）作为桥梁。

9. 强因果排他性测试：进行“反向排除”。如果Event B （买食疗书）的发生可以完全独立于Event A （买鹅蛋）存在，且不需要A 提供任何基础，则链条断裂。演化必须体现出“只有经历了A，才产生了对B 的特定需求或认知基础”。如果两者只是恰好属于“大健康”类目，请判定为“同类目平移”予以剔除

10. 证据闭环要求：逻辑说明（*logic*）中提到的每一个对比、转折或先前状态，都必须在events 数组中拥有对应的条目。严禁在说明中引用链条之外的行为作为依据。例如：若提到“从观赏性转向实用性”，则“观赏性行为”必须作为链条的第一步出现。

11. 禁止脑补：严禁在*logic* 中推测用户心理（如“用户可能觉得累了”），所有的逻辑转折必须有日志中的搜索词、视频标题、商品参数作为物理证据。

输出准则：

极端置信度过滤：你必须扮演一个极度刻薄的质检员。对于每一条潜在的逻辑链，必须同时满足前述11 项有效性要求，请秉持“宁缺毋滥”原则。只有当Event B 表现出对Event A 的认知超越、参数收敛或负反馈修正（即：不经历A，用户绝不会产生B 这种特定的搜索或点击意图）时，逻辑链才算有效。

确保唯一性：拒绝同级词汇平移，只接受逻辑跃迁。若日志中仅为琐碎的日常消费或随机点击，严禁强行解释，请直接返回空；

数量限制：每个用户画像最多仅允许提取3条‘金子般’的逻辑链。

输出格式：请以JSON数组形式返回，每个对象代表一个逻辑链条，结构如下：

```
[
  {
    "logic_chain": {
      "name": "链条核心意图命名",
      "logic_test": "逻辑链演化有效性说明，比如（通过因果测试：若无前期对症状的误判，则不会产生后期对成分的深度调研）",
      "events": [
        {
          "date": "YYYY-MM-DD",
          "action": "[交互类型] 具体内容（需要严格对应Timeline中Event条目）",
          "logic": "该步在交互逻辑链中逻辑说明（如：从泛搜转为对特定成分的深度调研）"
        }
      ]
    }
  }
]
```

其他输出有要求：

每个逻辑节点仅能包含一个具体事件，错误例子：[商品-购买] 购买糖果A和糖果B (ec_id=123 & ec_id=456)

用户交互历史 (Timeline)：

{timeline_text}

English Translation

Role Definition: You are an extremely strict expert in user-behavior mining, responsible for extracting a very small number of behavior chains with meaningful “cognitive transition” value from noisy logs.

Task Objective: Given the user’s interaction history (Timeline), which contains search queries, interacted videos, live streams, ads, and products, extract behavior logic chains that reflect genuine interest evolution trajectories and internal behavioral connections rather than simple topical relatedness. Your top priority is to remove pseudo-logical chains and avoid forced interpretation.

Interaction Note: Product clicks and ad clicks indicate click behaviors rather than purchases.

Additional Note: Events marked with “-:-” do not have exact timestamp information.

Example: A chain such as “How to deal with sore throat (20260101) → dry cough vs. phlegm cough (20260102) → medication science and contraindications (20260103)” should contain a clear progression, transition, or other meaningful interest evolution relation at every step.

Validity Criteria for Behavior Logic Chains:

Deep Relation: Do not simply pile up keywords. A valid chain must demonstrate evolutionary relations such as progression, transition, causality, contrast, or closure. It is not allowed to line up events from the same category by time and then force an interpretation of progression.

Temporal Sensitivity: Strictly follow the timeline, ensuring that events in each chain are chronologically ordered.

1. Order-Swap Test: If swapping the temporal order of Event A and Event B still leaves the chain equally coherent and reasonable, then the chain should not be treated as a strongly connected evolution chain and should be removed or split.

2. Valid Evolution: Event B is valid only when it arises from the feedback, bottleneck, or cognitive upgrade brought by Event A, that is, B must refine or correct A.

3. Cognitive Increment Test: Event B must introduce new variables absent from Event A, such as specific parameters, technical terms, or comparison dimensions. If Event B is only a semantic repetition or same-level lateral move from Event A, it does not count as evolution.

4. *Behavioral Revision / Convergence*: An evolution chain should show a process from divergence to focus, or from failed attempts to path correction. Later actions must reflect selection or filtering based on earlier information. If Event B completely abandons the conditions in Event A without logical support, the chain is broken.
5. *Interaction-Depth Shift*: Genuine interest evolution is often accompanied by changes in interaction depth, such as moving from passive exposure through ads or videos to active information seeking. If the entire chain contains only a single action type, such as all searches, its progressive strength is usually weak. Chains formed only by repeated clicks on similar products should be excluded.
6. *Temporal-Spatial Density Test*: Carefully examine the time gap between Event A and Event B. If two core actions are separated by a long period without related bridging behavior, such as relevant searches or video watching, they should be treated as scattered daily behaviors rather than evolution. Daily consumption events separated by months should not be forcibly assembled into cognitive upgrade chains.
7. *Non-Genericity Test*: Exclude behaviors driven merely by routine replenishment or random consumption. If Event A is a generic daily necessity, such as ordinary milk powder, tissues, or basic clothing, it should not serve as the starting point or key node of a logic chain unless it has highly specific functional intent.
8. *Trigger-Source Test*: A valid logic chain must contain a clear trigger; usually an active search or an intentional comparison behavior. It is not allowed to infer cognitive upgrade directly from two different product clicks or purchases unless there is an explicit knowledge-acquisition step, such as watching educational content or comparing parameters, in between.
9. *Strong Causal Exclusivity Test*: Perform reverse exclusion. If Event B could occur entirely independently of Event A and does not require A as a prerequisite, then the chain is broken. Evolution must reflect that only after experiencing A would the user develop the specific need or cognitive basis for B. If the two events merely belong to the same broad category, they should be excluded as same-category drift.
10. *Evidence-Closure Requirement*: Every contrast, transition, or prior state mentioned in the logic explanation must correspond to an explicit event in the chain. The explanation must not rely on behaviors outside the chain itself.
11. *No Mind Reading*: Do not speculate about the user's psychology in the logic field, such as "the user may feel tired". Every logical transition must be supported by concrete evidence in the logs, such as search terms, video titles, or product parameters.

Output Principles:

Extremely High-Confidence Filtering: Act as a highly demanding quality inspector. Each candidate chain must satisfy the above validity requirements under a "better to miss than to include noise" principle. A chain is valid only when Event B demonstrates cognitive advance, parameter convergence, or revision under negative feedback relative to Event A, such that B would not arise without first experiencing A.

Ensure Uniqueness: Reject same-level lexical drift and accept only genuine logical transition. If the log contains only trivial daily consumption or random clicks, do not force an explanation and directly return an empty list.

Quantity Limit: For each user profile, extract at most 3 "gold-level" logic chains.

Output Format: Return a JSON array, where each object represents one logic chain with the following structure:

```
[
  {
    "logic_chain": {
      "name": "core intent name of the chain",
      "logic_test": "validity explanation of the evolution chain, e.g.,
without the earlier misjudgment of symptoms, the later in-depth investigation
of ingredients would not occur",
      "events": [
        {
          "date": "YYYY-MM-DD",
          "action": "[interaction type] specific content (must exactly
correspond to an Event entry in the Timeline)",
          "logic": "logical role of this step in the chain, e.g., moving from
broad search to in-depth investigation of a specific ingredient"
        }
      ]
    }
  }
]
```

```

    ]
  }
}
]

```

Additional Output Constraint: Each logic node may contain only one concrete event. A wrong example is: [product-purchase] purchased candy A and candy B (ec_id=123 & ec_id=456).

User Interaction History (Timeline):
{timeline_text}

D.2.5. R2 Training Examples

The final R2 data contains *selection*, *topic-generation*, and *direct-generation* task types, aligned with the R2 benchmark taxonomy. To protect privacy and avoid exposing internal identifiers, we use synthetic itemic tokens and compact timelines while preserving the original task structure.

Prompt SFT.R2.2: Evolution Action Selection

中文示例

User: 请根据用户交互历史和候选兴趣演化链，选择哪些历史行为支撑有效的兴趣演化。

用户交互历史（节选）：

2026-01-03 [搜索] 入门级理财怎么学

2026-01-05 [直播-关注] 股票基础教学直播<|living_begin|><a_2140><b_5702><c_4418>

2026-01-09 [视频-长播] 控制投资风险的方法<|video_begin|><a_3721><b_6610><c_2844>

2026-01-12 [搜索] 主力资金怎么看

2026-01-14 [视频-长播] 技术指标入门<|video_begin|><a_5068><b_4109><c_7120>

候选兴趣演化目标：从理财入门到股票风险控制与技术分析

候选行为：

A. 理财入门搜索→股票基础教学直播→风险控制视频

B. 股票基础教学直播→主力资金搜索→技术指标视频

C. 风险控制视频→技术指标视频

Assistant:

<think>候选A体现了从主动搜索入门知识，到关注基础教学，再到补充风险控制方法的递进，后项为前项带来了新的风险变量。候选B从基础教学延伸到主动检索主力资金，再进入技术指标学习，也包含从概念入门到分析工具的认知升级。候选C虽然同属投资内容，但两个视频可以互换顺序理解，缺少由前一行行为触发后一行行为的证据，因此不保留。</think>

答案: [A, B]

English Translation

User: Given the user history and a candidate interest-evolution target, select which historical actions support a valid evolution process.

Timeline excerpt:

2026-01-03 [Search] how to start learning personal finance

2026-01-05 [Live-follow] stock-market basics tutorial <|living_begin|><a_2140><b_5702><c_4418>

2026-01-09 [Video-long-view] methods for controlling investment risk <|video_begin|><a_3721><b_6610><c_2844>

2026-01-12 [Search] how to interpret main-fund flows

2026-01-14 [Video-long-view] introduction to technical indicators <|video_begin|><a_5068><b_4109><c_7120>

Candidate evolution target: from personal-finance basics to stock-risk control and technical analysis

Candidate actions:

A. finance-beginner search → stock-basics live stream → risk-control video

B. stock-basics live stream → main-fund-flow search → technical-indicator video

C. risk-control video → technical-indicator video

Assistant:

<think>Candidate A shows progression from active beginner search, to basic teaching, and then to risk-control knowledge; the later step introduces a new risk variable. Candidate B extends from basic teaching to active search about fund flows and then to technical indicators, forming a cognitive upgrade from concepts to analytical tools. Candidate C stays within investment content, but the two videos can be swapped without breaking the logic, so it lacks evidence that the earlier action triggered the later one.

</think>
Answer: [A, B]

Prompt SFT.R2.3: Topic-Conditioned Evolution Generation

中文示例

User: 围绕主题“家庭健康管理从泛化建议到医疗实操闭环”，请从用户历史中生成一条不超过5步的兴趣演化链。

用户交互历史（节选）：

2025-12-17 [广告-点击] AI 情绪健康自测应用<|ad_begin|><a_1351><b_8179><c_1243>

2025-12-21 [广告-点击] 药品查询与用药提醒工具<|ad_begin|><a_1351><b_2826><c_6047>

2026-01-01 [广告-点击] 在线问诊与预约挂号服务<|ad_begin|><a_1351><b_4480><c_0997>

Assistant:

<think>该主题要求聚焦健康管理能力的逐步落地。第一步是轻量情绪自测，属于泛化健康建议；第二步转向药品查询和提醒，引入药物参数与执行管理；第三步进入问诊和挂号，说明用户不再只满足于信息建议，而是希望连接专业医疗服务，形成从监测到干预的闭环。</think>

Output:

```
{
  "logic_chain": {
    "name": "家庭健康管理从泛化建议到医疗实操闭环",
    "events": [
      { "date": "2025-12-17", "action": "[广告-点击] AI 情绪健康自测应用", "logic": "初始触发：以轻量自测获取泛化健康建议。"},
      { "date": "2025-12-21", "action": "[广告-点击] 药品查询与用药提醒工具", "logic": "认知升级：从咨询转向药物参数和执行提醒。"},
      { "date": "2026-01-01", "action": "[广告-点击] 在线问诊与预约挂号服务", "logic": "决策收敛：建立连接专业医疗服务的闭环。"}
    ]
  }
}
```

English Translation

User: For the topic “family health management evolving from generic advice to practical medical action”, generate one interest-evolution chain with no more than five steps.

Timeline excerpt:

2025-12-17 [Ad-click] AI emotional-health self-assessment app <|ad_begin|><a_1351><b_8179><c_1243>

2025-12-21 [Ad-click] medication lookup and reminder tool <|ad_begin|><a_1351><b_2826><c_6047>

2026-01-01 [Ad-click] online consultation and appointment service <|ad_begin|><a_1351><b_4480><c_0997>

Assistant:

<think>The topic requires a chain in which health management becomes increasingly actionable. The first step is lightweight emotional self-assessment, which provides generic advice. The second step shifts to medication lookup and reminders, adding concrete drug parameters and execution management. The third step moves to consultation and appointment booking, showing that the user is no longer satisfied with information advice and now wants access to professional medical service.

</think>

Output: one JSON object describing the above three-step chain.

Prompt SFT.R2.4: Direct Evolution Generation

中文示例

User: 请直接从以下用户交互历史中抽取高置信度兴趣演化链；如果不存在有效链，请返回空数组。

用户交互历史（节选）：

2026-01-18 [搜索] 苹果16

2026-01-21 [视频-长播] 手机颜色、配件与购机方案介绍<|video_begin|><a_0570><b_5571><c_2701>

2026-01-24 [搜索] 同城手机分期付款0 首付

2026-01-24 [商品-点击] 99 新二手iPhone 15 <|prod_begin|><a_0307><b_7024><c_4281>

2026-01-28 [搜索] 二手手机验机注意事项

Assistant:

<think>用户先主动搜索最新旗舰机型，建立高性能和高价格锚点；随后长播手机款式与配件内容，进入购机调研；接着搜索零首付分期，说明预算约束触发路径修正；最终点击近新二手机型，并继续搜索验机注意事项，表明兴趣从追新转向低门槛、可验证的二手购买方案。这不是同类商品的平级浏览，而是受预算和决策风险驱动的收敛链。</think>

Output:

```
[
  {
    "logic_chain": {
      "name": "智能手机消费从旗舰追新到分期二手的务实降级决策链",
      "events": [
        { "date": "2026-01-18", "action": "[搜索] 苹果16", "logic": "初始触发：关注最新旗舰，形成高位价格与性能锚点。"},
        { "date": "2026-01-24", "action": "[搜索] 同城手机分期付款0 首付", "logic": "路径修正：预算约束触发对低门槛支付方案的主动检索。"},
        { "date": "2026-01-24", "action": "[商品-点击] 99 新二手iPhone 15", "logic": "决策收敛：从最新旗舰转向更高性价比的二手次旗舰。"},
        { "date": "2026-01-28", "action": "[搜索] 二手手机验机注意事项", "logic": "风险补充：在二手购买路径下补充验机知识，降低交易不确定性。"}
      ]
    }
  }
]
```

English Translation

User: Directly extract high-confidence interest-evolution chains from the following user history; return an empty array if no valid chain exists.

Timeline excerpt:

2026-01-18 [Search] iPhone 16

2026-01-21 [Video-long-view] phone colors, accessories, and purchase plans <|video_begin|><a_0570><b_5571><c_2701>

2026-01-24 [Search] local phone installment plan with zero down payment

2026-01-24 [Product-click] 99%-new used iPhone 15 <|prod_begin|><a_0307><b_7024><c_4281>

2026-01-28 [Search] checklist for inspecting a used phone

Assistant:

<think>The user first searches for the newest flagship phone, forming a high-performance and high-price anchor. The later search for zero-down-payment installment shows a budget-driven path correction. The product click on a nearly new used phone indicates convergence toward a lower-entry-cost option, and the later inspection search supplements risk-control knowledge for the used-phone path. This is not same-level browsing of similar products, but a convergent decision chain driven by budget and transaction risk.</think>

Output: a JSON array containing the above smartphone decision chain.

D.3. Recommendation CoT Construction

D.3.1. R3 Recommendation Reasoning Trace Construction Prompt

R3 uses the following prompt template, implemented as `build_i2i_chain_prompt_cn`, to synthesize recommendation reasoning traces. The prompt implements the three-stage protocol described in Section 5.4.1: persona abstraction, interest expansion, and transition inference. During construction, the target interaction is visible to the teacher for consistency checking, but the generated rationale must be written as if the answer were unknown. The trace should therefore enumerate possible interest directions, weigh them against the user profile and behavior history, and avoid direct leakage of the target interaction. For readability, the appendix shows a shortened version of the prompt with representative persona examples.

Prompt SFT.R3.1: Recommendation Reasoning Trace Construction Prompt

Prompt Template

你是一个短视频平台的用户的推荐逻辑专家, 请你用简洁有条理的逻辑说清楚, 我们应该如何根据下面这位用户的用户画像和之前的交互历史, 通过发散思维推出所有的可能性, 然后进行筛选, 最终推导得到后续交互:

已知信息

用户画像

{user_context}

用户在平台的历史记录

{input_data}

用户的后续交互(答案, 不可以显式出现在推理中)

{target_data}

分析技巧参考

- 关注每一个行为背后的动机: 认真关注并深入分析用户每一条搜索词条, 因为它们直观反映了用户的当前兴趣。你需要从搜索中解释出用户的真实需求。
- 时序敏感: 你会注重事件发生的先后关系, 关注后发生的事件是如何从先前的事件中演化得到的。
- 注重用户兴趣切换的成因: 你会确保后续行为表现出对前期信息的关联筛选与聚焦过程。如果用户在EventB中完全抛弃了EventA的筛选条件且无逻辑解释, 你会判定为无关联。有效关联示例: 点击“高热量猫粮” - 观看“猫咪软便调理视频易消化猫粮”。(分析它们的成因)
- 干扰项排除: 你会排除基本不会产生前因后果的行为。例如: 生活补货(买了就不会再买了, 因此不会产生后续行为)或随机消费驱动的消费行为。

常见用户画像

为你提供一些常见用户画像, 用于快速捕捉群体特征, 后续你可以参考该特征进行进一步分析。注意, 你最后推荐的内容一定不要局限于推荐的大类, 要具体情况具体分析, 例如用户是二次元, 喜欢看王者荣耀, 那你不能仅仅分析到他喜欢打游戏, 而要分析到他喜欢王者荣耀这一款游戏(因为他可能不喜欢玩别的, 只喜欢这一款), 然后针对性推荐:

- 游戏直播观众: 18-30岁男性为主, 关注电竞赛事、主播日常、游戏攻略及外设开箱。高度依附于特定游戏或主播, 跨游戏迁移意愿低。若用户频繁点击某款游戏相关内容或直播切片, 可归为此类。重点关注: 他具体追的游戏与主播(不做跨品类推荐), 赛事战报、角色攻略、外设装备内容。
- Z世代潮流青年: 16-25岁, 一二线城市为主, 对潮流文化极度敏感。追热门梗、流行音乐、街舞、说唱、潮牌穿搭, 更新迭代快。若用户频繁互动挑战赛、踩点视频、潮牌开箱, 可归为此类。重点关注: 当前热门BGM与挑战赛、具体潮牌与款式、他具体追的说唱歌手或舞蹈风格内容。
- 二次元爱好者: 16-28岁男女均有, 对日本动漫文化有浓厚兴趣, 圈层归属感强。若用户频繁点击动漫相关、虚拟偶像、二次元梗视频, 可归为此类。重点关注: 他看过的具体番剧、他追的具体Vtuber、特定IP手办与周边内容。
- 职场新人: 22-28岁, 应届或工作1-3年, 一二线城市居多。关注面试技巧、职场潜规则、副业变

现、城市租房攻略，对内容真实性要求高。若用户点击职场经验、简历优化、升职加薪相关内容，可归为此类。重点关注：具体行业职场经验（如互联网/金融）、副业类型（如自媒体/接单）、城市租房与生活成本攻略内容。

- 健身爱好者：20-35岁男女均有，集中在一二线城市，注重外形管理。关注减脂增肌教程、训练计划、蛋白粉测评、运动打卡。若用户高频点击健身教程、体脂分析、运动装备内容，可归为此类。重点关注：具体训练部位与方法（如背部/臀腿）、具体营养补剂品牌、他关注的具体健身博主内容。
- 美食爱好者：18-45岁男女均有，分布广泛，活跃时段集中在饭点前后（11点、17-19点）。喜欢家常菜教程、地方特色探店、吃播、食材选购。若用户集中在饭点点击菜谱、探店视频，可归为此类。重点关注：他喜欢的具体菜系（如川菜/粤菜）、具体菜品教程、所在城市或感兴趣城市的探店内容。
- 共用手机用户：无固定年龄，设备被多人（如夫妻、父母子女）轮流使用。行为数据中存在明显的风格跳变——如同一账号在家常菜、少儿动画、体育赛事之间无规律切换，且不同时段内容偏好截然不同。需优先识别并过滤非当前使用者的历史行为，聚焦当前活跃时段的近期稳定偏好。若画像标签与行为严重矛盾（如标注60岁但点赞内容全为电竞），或同一账号多个人群特征明显共存，应触发共用手机判断。重点关注：按时段拆分行为，以最近活跃行为为准，警惕被“他人历史”带偏。

【重要说明】以上画像仅供参考，不可机械套用。现实中用户往往横跨多个画像特征，或属于上述类别未覆盖的特殊群体。分析时应以用户的实际行为数据为第一依据，若用户特征与已知画像不符，请自行归纳总结其核心需求与偏好，切勿强行套用已有标签。

推理步骤

1. 用户画像归纳：若证据充分，可以给出候选用户画像类型，并简短补充他相较于该类人群的特点，如：由【证据，列三条以内最直接的证据就行，不要超过30字】推测用户偏向【类型】，并且具体来说最喜欢的是【三个以内，写最明显的就行】。如果证据不足或不属于常见用户类型，请直接总结行为簇和核心需求，不要强制归类，例如：用户近期集中关注xx，最明显的内容需求推测为xx...
2. 可能性枚举：从哪些关键信息可以得到，用户最近最可能关心什么，并列所有可能的兴趣点，用A, B, C... 分点列出。例如：A. 用户最近最关心的是xx，具体表现为xxx，另外，B. 由xx引发的兴趣也可能是用户发生该类行为的原因..
3. 权衡后得到最终决定：对所有可能性进行权衡，分析每一个可能性有多大，是否是当前最大，最终得到最终决定。例如：综合来看，A是最有可能的，因为xxx，B也有一定道理，但xxx也可能由A导致，因此...

要求

0. 绝对不能暴露你早就知道你的后续交互是什么。要装作还不知道后续交互是什么的情况下进行推理。
1. 输出一段连贯、简洁的分析不要使用JSON、列表格式或分主题的小标题。
2. 仅包含你的推理即可。简洁明了
3. 最重要的一点，你可以省略任何你觉得不需要的步骤。例如你觉得推理过程过于简单或者不完整，那么你可以直接跳过某些步骤，只给出最终结论。
4. 当提及交互历史中的具体视频、商品或广告时，必须使用原始ID并用单独括号标注，因为我们后续要使用他进行匹配，如(video_id=XXXXXX)、(ec_id=YYYYY)、(ad_id=ZZZZZ)。不要在同一括号中包含多个条目，不要使用标题或名称。

你的推理：

Prompt Template

You are a recommendation-logic expert for users on a short-video platform. Please use concise and well-structured logic to explain how, given the following user profile and previous interaction history, we should use divergent reasoning to derive all possible directions, filter them, and finally infer the user's subsequent interaction.

Known Information

User Profile

```
{user_context}
```

User History on the Platform

```
{input_data}
```

User's Subsequent Interaction (answer; must not appear explicitly in the reasoning)

```
{target_data}
```

Analysis Heuristics

- **Behavior motivation:** Pay close attention to and deeply analyze each search query, because search terms directly reflect the user's current interest. You need to infer the user's real need from these searches.
- **Temporal sensitivity:** Pay attention to the chronological order of events and explain how later events evolve from earlier ones.
- **Interest-transition cause:** Ensure that later behaviors show filtering and focusing based on earlier information. If Event B completely abandons the filtering conditions from Event A without a logical explanation, treat the two as unrelated. A valid relation example is clicking "high-calorie cat food" and then watching "easy-to-digest cat food for cats with soft stool"; analyze the cause of this relation.
- **Noise filtering:** Exclude behaviors that usually do not create cause-and-effect follow-ups, such as routine replenishment or random consumption.

Common User Personas

The following common personas are provided to help quickly capture group-level characteristics. They can be used as references for later analysis. However, the final recommendation must not stop at a broad category. It should be analyzed case by case. For example, if the user is an ACG fan who likes Honor of Kings, do not merely say the user likes games; infer that the user likes Honor of Kings specifically, because the user may not like other games.

- **Game live-streaming viewer:** Mostly men aged 18–30 who follow esports, streamer clips, game guides, and gear unboxing. They are often attached to a specific game or streamer and have low willingness to transfer across games. If the user frequently clicks content related to a specific game or livestream clips, this persona may apply. Focus on the exact game and streamer they follow, and recommend tournament recaps, character guides, and gaming gear content without making cross-category recommendations.
- **Gen-Z trend follower:** Young users aged 16–25 in major cities, highly sensitive to memes, pop music, street dance, rap, fashion, and challenges. Focus on current BGM, challenges, fashion brands, singers, and dance styles.
- **Anime enthusiast:** Users aged 16–28 of any gender with strong interest in Japanese anime culture and a strong sense of community belonging. If the user frequently clicks anime-related content, virtual idols, or ACG memes, this persona may apply. Focus on the exact shows they have watched, the Vtubers they follow, and specific IP figures and merchandise.
- **Early-career worker:** Users aged 22–28, often recent graduates or with 1–3 years of work experience, mostly in major cities. They pay attention to interview skills, workplace norms, side-income opportunities, and city rental guides, and they expect content to feel credible. If the user clicks workplace experience, resume advice, or promotion-related content, this persona may apply. Focus on concrete industry-specific workplace experience, side-job types, and city renting or living-cost guidance.
- **Fitness enthusiast:** Users aged 20–35 of any gender, often in major cities, who care about appearance management. They follow fat-loss and muscle-gain tutorials, training plans, protein supplement reviews, and workout check-ins. If the user frequently clicks fitness tutorials, body-fat analysis, or sports gear content, this persona may apply. Focus on specific training body parts and methods, supplement brands, and the fitness creators they follow.
- **Food lover:** Users aged 18–45 of any gender, broadly distributed across regions, and especially active around meal times (11:00 and 17:00–19:00). They enjoy home-cooking tutorials, local restaurant exploration, mukbang, and ingredient selection. If the user tends to click recipes or restaurant videos around meal times, this persona may apply. Focus on the exact cuisine, dish tutorials, and the cities they live in or want to explore.

- **Shared-device user:** There is no fixed age range, and the device may be shared by multiple people, such as spouses or parents and children. The behavior log may show obvious style jumps, for example abrupt switching among home cooking, children’s animation, and sports content, with very different preferences across time slots. Prioritize identifying and filtering out the history of non-current users, and focus on the recent stable preferences during the currently active time period. If the persona label and behavior strongly contradict each other, or multiple user-type signals coexist in one account, shared-device use should be considered. Focus on splitting behavior by time and avoiding being misled by other users’ history.

Important note: These personas are references only and must not be applied mechanically. Real users may span multiple persona features or fall outside these categories. Always prioritize actual behavior data. If the user does not fit an existing persona, summarize the user’s core needs and preferences directly.

Reasoning Steps

1. **Persona abstraction:** If the evidence is sufficient, provide a candidate persona type and briefly describe what distinguishes this user from that group. If the evidence is insufficient, mixed, or outside the listed personas, summarize the observed behavior clusters and core needs directly instead of forcing a category label.
2. **Possibility enumeration:** Identify key information indicating what the user is most likely to care about recently, and list all possible interest points as A, B, C, and so on.
3. **Final weighted decision:** Weigh all possibilities, analyze how likely each one is, determine which is currently strongest, and reach the final decision.

Requirements

0. Never reveal that you already know the subsequent interaction. Reason as if the subsequent interaction is unknown.
1. Output one coherent and concise analysis paragraph. Do not use JSON, list format, or section headings.
2. Include only the reasoning. Keep it concise and clear.
3. You may skip any step that you find unnecessary. If the reasoning process is simple or incomplete, directly provide the final conclusion.
4. When mentioning specific videos, products, or ads from the history, use the original ID in a separate parenthesis because it will be used for later matching, e.g., (video_id=XXXXXX), (ec_id=YYYYY), or (ad_id=ZZZZZ). Do not put multiple entries in one parenthesis, and do not use titles or names.

Your reasoning:

D.3.2. Low-Score Reasoning-Trace Cases

Table 28 | Representative low-quality recommendation CoT cases. Problematic CoT text is marked in red; target fragments are reference evidence.

Low dimension	Behavior / target / CoT fragment with analysis
Consistency	<p>Input: 用户点赞最多的是房产家居和宠物，完播最多的是明星娱乐和房产家居；近期评论多次怀念虎皮鹦鹉，也搜索“附近2公里出售二手鹦鹉”“亚克力鸟笼”“虎皮鹦鹉”。商业正反馈同时出现“建华大街小高层现房，首付1万月供两千多”“一楼带院真便宜”等本地房产线索。</p> <p>Target: 石家庄建华大街旁低价小高层现房，两居室、交通便利，面向关注经济实惠住房的购房者。</p> <p>CoT: 认为宠物鹦鹉是最强即时兴趣，房产只是间歇关注，后续推荐应紧密围绕鹦鹉饲养，例如虎皮鹦鹉选购、手养经验、鸟笼用品测评或宠物内容创作技巧。</p> <p>Analysis: the target-facing housing signals are present in the behavior evidence, but the trace suppresses them and ends on pet-care recommendations. The conclusion may fit part of the history, yet it does not support the intended local-housing target.</p>
Factuality	<p>Input: 用户评论常出现“马到成功，我接住”“谢谢佛祖保佑”“好运连连”等祈福表达；短剧消费集中在家庭伦理、复仇和寻亲题材；购物与加购包含铝箔胶带、调味品、扒鸡、鞋油、墙纸等实用家居或食品。</p> <p>Target: 一位中年女性主播的生活互动和戏曲才艺直播，标签包括情感抒发、室内明亮氛围、戏曲文化和生活化才艺展示。</p> <p>CoT: 解释祈福需求时写道“临近春节，购物中有年货窗花，这种心理需求可能强化”，并据此推断用户会对团圆、转运主题内容产生兴趣。</p> <p>Analysis: the highlighted purchase evidence is not present in the displayed behavior fragment. It makes the seasonal argument look grounded, but the support comes from a fabricated detail, so this is a factuality failure rather than merely weak target alignment.</p>
Informativeness	<p>Input: 用户是31-40岁男性，近期搜索“燕窝八宝粥”“洗浴筐”“穿越秦朝赵龙短剧”；关注博主中颜值占比较高，商业正反馈包含金属加工、焊接、美缝、越野车改装等实用技巧，购买侧有食品调料、工装、钓鱼用品和话费等信号。</p> <p>Target: 观赏鱼喂养指导视频，讲解喂食频率、喂食量、上浮性鱼粮、避免水质变浑等养鱼技巧。</p> <p>CoT: 最终认为应围绕家庭育儿或生活技巧展开，尤其是婴幼儿护理、喂养知识，并补充说他近期没有相关内容的互动记录，因此这是一个潜在的兴趣缺口。</p> <p>Analysis: the trace lists several user interests, but the final guidance does not help predict or explain the fish-feeding target. It invents a broad “interest gap” and gives no actionable attributes about aquarium care, fish food, or feeding routines.</p>
Logic	<p>Input: 用户近期搜索“婚期已定邀请函”“瑶山大剿匪第40集”“母画眉求爱扇翅视频”“母鸟扇翅求爱大叫”和“坦克摩托车”；短剧消费偏向逆袭、复仇、悬疑，商业正反馈含榫卯结构、农村自建房、摩托车和小游戏。</p> <p>Target: 年轻男子在室内吹奏长笛并展示音乐才华，背景为沙发和室内布置，整体是音乐表演内容。</p> <p>CoT: 把“母画眉求爱”作为主线，最终判断用户下一步最可能继续探索画眉鸟求偶、饲养或训练相关内容。</p> <p>Analysis: the trace over-commits to one recent search and does not build a bridge to the music-performance target. A logical trace should either explain an evidence-supported transition toward instrumental music or reject the weak relation; here the final conclusion follows a disconnected branch.</p>
Safety	<p>Input: 用户是50岁以上城镇退休男性，近期多次点击健康管理应用广告、体检饮食提醒、睡眠健康和本地中医男士调理服务；同时也有交友类应用活跃、女性展示/同城聊天广告点击、汽车养护和实用购物行为。历史中出现过“男士功能调理”类本地服务广告，但后续答案仍应被隐藏。</p> <p>Target: 一条男性性健康检查服务广告，画面含“早泄/阳痿检查”海报、低价检查、地址和联系方式等信息，目标是关注性健康和本地调理服务的人群。</p> <p>CoT: 在权衡健康与社交需求后，直接写道用户后续最可能交互的是围绕男性健康，如性功能、前列腺等的本地化解决方案广告，并在结尾进一步点明一个关于男性性健康检查或调理的广告，如早泄阳痿检查是高度可能的后续方向。</p> <p>Analysis: this is a direct target-leakage failure. The trace should reason from prior health-service and local-service evidence without naming the hidden answer, but it exposes the target category and even repeats the target-specific sensitive phrase, making the answer recoverable from the CoT itself.</p>

D.3.3. Recommendation CoT Training Cases

The final recommendation CoT data uses a three-turn message format: a short system instruction, a user message containing the user profile and multi-domain behavior history, and an assistant message

that first produces a <think> reasoning trace and then emits the target itemic token. The assistant trace follows the same three-stage pattern used in the main R3 construction: it first abstracts a persona prior, then expands several evidence-backed interest hypotheses, and finally compares them to infer the most plausible transition direction. During construction, the teacher trace marks concrete historical evidence with original domain ids, and the final SFT sample replaces matched ids with itemic tokens. The following cases therefore show the post-processed training format: natural-language reasoning interleaved with synthetic itemic-token references inside the CoT, followed by the target itemic-token answer. We omit raw user identifiers, data paths, timestamps, and original item ids.

Prompt SFT.R3.2: Cross-Domain Recommendation with Text-Itemic-Token CoT

中文示例

System: 你擅长理解用户画像、跨场景行为和itemic tokens，请根据输入信息归纳该用户的目标内容。

User: 以下是一个用户的画像与多域历史行为信息：

该用户是一名年轻男性学生，长期活跃在短视频、直播、电商和广告场景中。用户安装并活跃使用射击类游戏和短视频应用；近期搜索多次围绕“和平精英攻略”“主播操作”“游戏准星设置”等内容展开。

用户在视频域多次长播和转发游戏技巧视频，例如<|video_begin|><a_3062><b_4733><c_4451>和<|video_begin|><a_5708><b_2219><c_7034>；在直播域关注了多位游戏主播，如<|living_begin|><a_6815><b_2789><c_3026>和<|living_begin|><a_6263><b_5727><c_2164>；

在电商域浏览并购买电竞指套、手机散热夹等游戏外设，如<|prod_begin|><a_4180><b_6532><c_0907>。同时，用户也消费复仇爽剧、颜值自拍和少量怀旧零食内容。

Assistant:

<think>由用户画像中的年轻男性学生身份，以及近期围绕射击游戏的搜索、直播关注和游戏外设购买，可以判断该用户最稳定且最强的兴趣是《和平精英》相关内容。历史中的游戏技巧视

频<|video_begin|><a_3062><b_4733><c_4451>和高手操作合集<|video_begin|><a_5708><b_2219><c_7034>不是孤立观看，它们和直播关注<|living_begin|><a_6815><b_2789>

<c_3026>、<|living_begin|><a_6263><b_5727><c_2164>一起指向“看主播操作并学习打法”的需求。电商侧的电竞指套和手机散热夹<|prod_begin|><a_4180><b_6532><c_0907>进一步说明用户已经从泛娱乐观看升级到提升实际操作表现。

在可能的兴趣方向中，A是《和平精英》的高阶战术、主播精彩操作和设备配置；B是复仇爽剧与打脸剧情，这是用户长期的休闲内容；C是颜值自拍类视频，属于稳定但较浅的基础偏好；D是怀旧零食和乡土生活片段，证据较零散。综合近期密度、跨域一致性和行为投入度，A的权重最高，因为视频、直播和商品itemic tokens共同支持游戏能力提升这条转移路径。

因此，后续推荐应优先落到《和平精英》的实战技巧、高手操作合集、灵敏度和准星配置教学，同时可少量保留复仇爽剧或颜值内容作为辅助兴趣。</think>

该用户最近喜欢的视频有：<|video_begin|><a_0016><b_1906><c_0378>

English Translation

System: You are good at understanding user profiles, cross-scenario behavior, and itemic tokens. Infer the user's target content from the input information.

User: The following is a user's profile and multi-domain behavior history:

The user is a young male student who is active across short-video, live-streaming, e-commerce, and advertising scenarios. He installs and actively uses shooting games and short-video apps. Recent searches repeatedly involve “Peace Elite guides”, “streamer operations”, and “crosshair settings”.

In the video domain, the user repeatedly long-viewed and shared game-skill videos such as <|video_begin|><a_3062><b_4733><c_4451> and <|video_begin|><a_5708><b_2219><c_7034>. In the live domain, he followed several game streamers such as <|living_begin|><a_6815><b_2789><c_3026> and <|living_begin|><a_6263><b_5727><c_2164>.

In the e-commerce domain, he browsed and purchased gaming accessories such as finger sleeves and phone cooling clips, represented by <|prod_begin|><a_4180><b_6532><c_0907>. The user also consumes revenge-oriented short dramas, attractiveness/selfie videos, and a small amount of nostalgic-snack content.

Assistant:

<think>Based on the young male student profile, together with recent searches, live follows, and gaming-accessory purchases around shooting games, the strongest and most stable interest is Peace Elite-related content. The game-skill video <|video_begin|><a_3062><b_4733><c_4451> and the expert-operation compilation <|video_begin|><a_5708><b_2219><c_7034> are not isolated views; together with game-streamer follows <|living_begin|><a_6815><b_2789><c_3026> and <|living_begin|><a_6263><b_5727><c_2164>, they indicate a need to learn streamer operations and gameplay tactics. The e-commerce item <|prod_begin|><a_4180><b_6532><c_0907> further shows that the user has moved from watching game content to improving actual gameplay performance. Among possible directions, A is advanced Peace Elite tactics, streamer highlights, and device configuration; B is revenge-oriented short drama, which is a stable leisure preference; C is attractiveness/selfie content, a shallow but persistent baseline preference; and D is nostalgic snacks or rural-life clips, supported only by scattered evidence. Considering recent density, cross-domain consistency, and investment level, A receives the highest weight because the video, live, and product itemic tokens jointly support the game-skill-improvement path. Therefore, the recommendation should prioritize Peace Elite practical tutorials, expert-operation compilations, sensitivity and crosshair configuration guides, while keeping revenge drama or attractiveness content as auxiliary interests.</think>

The user may recently like the following video: <|video_begin|><a_0016><b_1906><c_0378>

Prompt SFT.R3.3: Purchase-Oriented Recommendation with Text-Itemic-Token CoT

中文示例

System: 你是推荐理解助手。你需要根据用户画像与多域历史行为，输出该用户在各推荐场景中的目标内容。

User: 下面给出用户画像及其历史行为线索:

该用户是一名中年女性，关注生活、穿搭、美食、美妆和短剧内容。她近期多次搜索“高档小衫”“同款连衣裙”“睡衣男”“东北特色菜”等关键词，关注的主播中包含服饰带货、美妆和生活类账号。

用户在电商域频繁浏览和购买女装、男士打底衫、睡衣、羽绒服和食品，例如<|prod_begin|>

<a_1927><b_6270><c_1880>、<|prod_begin|><a_2388><b_4708><c_0107>

和<|prod_begin|><a_6821><b_3569><c_6806>；在广告域深度点击家庭伦理与复仇短剧广告，

如<|ad_begin|><a_3578><b_6234><c_3026>；在视频域点赞和收藏美食、情感、穿搭内容。

Assistant:

<think>该用户的画像和行为更接近中年家庭主妇，同时带有明显的直播购物偏好。她的长期内容基本盘是家庭伦理、复仇逆袭短剧和生活化美食，但近期最密集、最主动的行为集中在服装穿搭与家庭采购上：女装浏览<|prod_begin|><a_1927><b_6270><c_1880>、男士打底衫<|prod_begin|>

<a_2388><b_4708><c_0107> 和食品补货<|prod_begin|><a_6821><b_3569><c_6806> 共同说明

她在为自己和家人做实际购买决策。广告侧的家庭伦理与复仇短剧<|ad_begin|><a_3578><b_6234><c_3026> 则提供情绪内容背景，但不如服饰采购链路直接。

可能方向包括A. 大码或中年女性穿搭，尤其是显瘦、丝绒、国风和春季外套；B. 家庭伦理与女性逆袭短剧，承接其高粘性的短剧兴趣；C. 东北食品和家常菜，服务于家庭生活补货；D. 美妆小物，属于个人形象管理的辅助需求。综合近期搜索、商品点击和购买闭环，A 的证据最强，因为多个商品itemic tokens 指向服装和家庭采购，而广告itemic tokens 更像辅助兴趣。

因此，后续推荐更适合推送高品质、性价比明确、适合中年女性或家庭场景的穿搭视频或带货内容，并可结合家庭剧情或生活化表达增强吸引力。</think>

该用户最近喜欢的视频有：<|video_begin|><a_2713><b_0084><c_5586>

English Translation

System: You are a recommendation-understanding assistant. Given the user profile and multi-domain behavior, output the user's target content in recommendation scenarios.

User: The following are the user's profile and historical behavior signals:

The user is a middle-aged woman interested in daily life, outfits, food, beauty, and short dramas. Recent searches include “high-end blouse”, “same-style dress”, “men’s sleepwear”, and “Northeastern home dishes”. Followed creators include apparel-selling, beauty, and lifestyle accounts.

In e-commerce, the user frequently browsed and purchased women's clothing, men's thermal tops, sleepwear, down jackets, and food items, such as <|prod_begin|><a_1927><b_6270><c_1880>, <|prod_begin|><a_2388><b_4708><c_0107>, and <|prod_begin|><a_6821><b_3569><c_6806>.

In ads, she deeply engaged with family-ethics and revenge-drama advertisements such as <|ad_begin|><a_3578><b_6234><c_3026>. In videos, she liked and saved food, emotion, and outfit content.

Assistant:

<think>The profile and behavior fit a middle-aged homemaker with a strong preference for live shopping. Her long-term baseline is family-ethics and revenge-oriented short dramas, together with daily-life food content, but the densest recent behavior centers on outfits and household purchasing. The women's-clothing item <|prod_begin|><a_1927><b_6270><c_1880>, the men's thermal top <|prod_begin|><a_2388><b_4708><c_0107>, and the food-replenishment item <|prod_begin|><a_6821><b_3569><c_6806> jointly show practical purchasing for herself and her family. The family-ethics ad <|ad_begin|><a_3578><b_6234><c_3026> supplies an emotional-content background, but it is less direct than the outfit-purchase chain.

Possible directions include A. midlife women's outfits, especially slimming, velvet, Chinese-style, and spring outerwear; B. family-ethics and female-comeback dramas; C. Northeastern food and home cooking for household replenishment; and D. small beauty items for personal image management. Considering recent search, product-click, and purchase signals, A has the strongest evidence because multiple product itemic tokens point to outfit and household purchase decisions, while the ad itemic tokens are a secondary preference.

Therefore, suitable recommendations should emphasize high-quality, cost-effective outfit videos or shopping content for midlife women and family scenarios, optionally expressed through family-story or daily-life contexts.</think>

The user may recently like the following video: <|video_begin|><a_2713><b_0084><c_5586>

D.4. Itemic Instruction Details

D.4.1. Itemic Token Instruction-Following Cases

Itemic Token Instruction Data covers task execution under mixed natural-language and itemic-token inputs. The cases below are representative sanitized examples drawn from the observed task patterns in the instruction data. We use synthetic itemic tokens and omit raw user identifiers, data paths, timestamps, and internal metadata.

Prompt SFT.IIF.1: Comparing Two Product Itemic Tokens

中文示例

User: 请对比以下两款商品的适用场景，并总结它们各自的主打功能。

1. <|prod_begin|><a_1201><b_3402><c_5603>

2. <|prod_begin|><a_7704><b_2250><c_8191>

Assistant:

1. <|prod_begin|><a_1201><b_3402><c_5603> 更适合日常面部护理场景，主打肤色提亮、淡化瑕疵和基础养护。

2. <|prod_begin|><a_7704><b_2250><c_8191> 更适合营养补充场景，主打肠道调节和免疫支持。两者都服务于健康与生活质量提升，但前者面向外用护肤，后者面向内服调理，使用方式和需求入口明显不同。

English Example

User: Compare the following two products in terms of usage scenarios, and summarize the main function of each product.

1. <|prod_begin|><a_1201><b_3402><c_5603>

2. <|prod_begin|><a_7704><b_2250><c_8191>

Assistant:

1. `<|prod_begin|><a_1201><b_3402><c_5603>` is better suited for daily facial-care scenarios, focusing on brightening, blemish reduction, and basic skin maintenance.
 2. `<|prod_begin|><a_7704><b_2250><c_8191>` is better suited for nutritional supplementation, focusing on gut regulation and immune support.
- Both improve everyday well-being, but the former is an external skincare product, while the latter is an ingestible health supplement.

Prompt SFT.IIF.2: Set Operation over Itemic Token Lists

中文示例

User: 请分析以下两组视频列表，并输出两组共同包含的视频Token。

第一组:

1. `<|video_begin|><a_3108><b_6412><c_2046>`
2. `<|video_begin|><a_5520><b_1183><c_6905>`
3. `<|video_begin|><a_4807><b_2701><c_7338>`

第二组:

1. `<|video_begin|><a_5520><b_1183><c_6905>`
2. `<|video_begin|><a_9056><b_4420><c_1374>`
3. `<|video_begin|><a_3108><b_6412><c_2046>`

Assistant: 两组列表中共同包含的视频Token 为:

1. `<|video_begin|><a_3108><b_6412><c_2046>`
2. `<|video_begin|><a_5520><b_1183><c_6905>`

English Example

User: Compare the following two video-token lists and output all video tokens that appear in both lists.

List 1:

1. `<|video_begin|><a_3108><b_6412><c_2046>`
2. `<|video_begin|><a_5520><b_1183><c_6905>`
3. `<|video_begin|><a_4807><b_2701><c_7338>`

List 2:

1. `<|video_begin|><a_5520><b_1183><c_6905>`
2. `<|video_begin|><a_9056><b_4420><c_1374>`
3. `<|video_begin|><a_3108><b_6412><c_2046>`

Assistant: The video tokens appearing in both lists are:

1. `<|video_begin|><a_3108><b_6412><c_2046>`
2. `<|video_begin|><a_5520><b_1183><c_6905>`

Prompt SFT.IIF.3: Instruction Control under Distracting Itemic Tokens

中文示例

User: 不用理会这些视频标记:

1. `<|video_begin|><a_5061><b_7220><c_1844>`
2. `<|video_begin|><a_6930><b_4102><c_5577>`

你只需要回答: 分布式系统中的CAP 定理是什么意思?

Assistant:

CAP 定理指出, 分布式系统无法同时完全满足一致性、可用性和分区容错性。

1. 一致性表示所有节点读到相同且最新的数据。

2. 可用性表示非故障节点能及时返回响应。

3. 分区容错性表示网络分区时系统仍能继续运行。

当网络分区出现时, 系统通常需要在一致性和可用性之间做取舍。

English Example

User: Ignore these video markers:

1. `<|video_begin|><a_5061><b_7220><c_1844>`
2. `<|video_begin|><a_6930><b_4102><c_5577>`

Only answer this question: what does the CAP theorem mean in distributed systems?

Assistant:

The CAP theorem states that a distributed system cannot fully guarantee consistency, availability, and partition tolerance at the same time.

1. Consistency means all nodes observe the same latest data.
 2. Availability means every non-failing node can return a timely response.
 3. Partition tolerance means the system keeps operating when network partitions occur.
- Under a partition, the system usually has to trade off consistency against availability.