
Scaling Articulated Rationales for MLLM-based Recommendation

 Kuaishou Technology

See [Contributions](#) section for a full author list.

Abstract

Users' explanations of why they like or dislike content provide explicit preference information that behavioral feedback and content descriptions alone do not fully capture. However, these articulated user rationales (AURs) are sparse, uneven in quality, and limited in coverage, making them difficult to use in industrial recommendation. We present **SARA** (**S**caling **A**rticulated **R**ationales), a framework that transforms sparse AURs into recommendation signals at scale. SARA first collects and curates questionnaire responses into **SARA-HQ**, an author-centric dataset that grounds model alignment in users' stated preferences. Using supervised fine-tuning and Quality-Refining DPO, **SARA-7B** learns to generate positive and negative rationales from multimodal author context, expanding coverage from 86,564 questionnaire-covered authors to the full 10M-author space. **SARA-Ranker** then incorporates these rationales into user-author interaction modeling and negative-feedback history modeling. Evaluation on unseen authors shows that SARA-7B produces more specific, relevant, and grounded rationales than strong general-purpose MLLM baselines. On top of an established multimodal ranking system, separate online A/B tests show that positive-rationale integration increases watch time by 0.99%, while negative-rationale integration reduces Hate feedback by 8.16%. With daily refresh and more than 30 days of production deployment, SARA demonstrates a practical path for using MLLMs to scale sparse human explanations into effective recommendation features.

Date: Sep 10, 2026

Correspondence: hk.xiao.me@gmail.com

Contents

1	Introduction	3
2	Data Engine	4
2.1	Overview	4
2.2	Online Collection	5
2.3	Post Curation	6
2.4	Dataset Statistics and Analysis	7
3	Rationale-Oriented Alignment	10
3.1	Overview	10
3.2	Large-Scale SFT	10
3.3	Quality-Refining DPO	10
4	Rationale Quality Evaluation Framework	12
4.1	overview	12
4.2	Online LLM Judge and Offline Agent Judge	12
4.3	Judge Corpus and Evaluator Training	13
4.4	Evaluator Validation	14
5	Rationale-as-Feature Integration	15
5.1	Overview	15
5.2	Rationale Embedding Integration	15
5.3	Rationale SID Integration	17
6	Experiments	18
6.1	SARA-7B Experiments	18
6.2	SARA-Ranker Experiment	22
7	Conclusion	23
8	Appendix	24
9	Contributions	25

1 Introduction

Modern recommendation systems (RS) primarily infer user preferences from behavioral feedback, including clicks, watch time, skips, and explicit negative actions, together with user profiles and interaction histories [20, 33]. These signals are abundant and easy to collect, but they mainly reveal what users do rather than why they make a decision. The same observed behavior can arise from different motivations, while a skip or dislike alone does not specify which aspect of the content should be avoided. Consequently, behavior-only supervision provides limited support for modeling fine-grained, transferable user preferences.

To bridge the semantic gap left by behavioral feedback, modern recommendation systems increasingly incorporate multi-modal signals to enrich their representations of users and items. Visual signals capture objects, scenes, appearances, and activities, while audio signals convey speech, music, vocal characteristics, and acoustic context. Textual signals provide complementary high-level semantics and user-generated opinions: hashtags help group related content [2]; news articles, headlines, and item descriptions characterize content semantics [11]; and reviews and comments provide opinionated evidence [9]. Recent MLLMs offer a unified interface for understanding and integrating these heterogeneous sources of information.

Nevertheless, richer multimodal representations do not necessarily reveal why a particular user likes or dislikes an item. Visual and audio signals primarily describe what appears or occurs in the content, while hashtags and metadata summarize its topics and attributes. Reviews and comments may contain subjective opinions, but they are produced under unconstrained self-expression and are not necessarily associated with a specific preference decision. This missing connection between multimodal content understanding and user preference motivates **articulated user rationales (AURs)**, which explicitly capture why a user likes or dislikes an item.

Yet AURs have rarely been used as scalable recommendation signals, primarily due to three challenges:

- **R1. Native sparsity:** After consuming an item, the most natural user behavior is to scroll away, skip, or remain silent. Explicitly articulating why one likes or dislikes an item is an extremely low-probability event, typically with a coverage rate below 1% [16].
- **R2. Generally low expression quality:** AURs entered through UI interfaces are often colloquial and fragmented, exhibit low information density, and are affected by emotional bias introduced through voluntary self-reporting [6].
- **R3. Speculative generation by general-purpose MLLMs:** Extending AUR coverage with general-purpose MLLMs is unreliable. Even when enhanced by behavioral signals, MLLMs observe preference outcomes rather than the underlying reasons, and may therefore generate plausible but post-hoc rationalizations that are not grounded in users’ articulated preferences [3, 13].

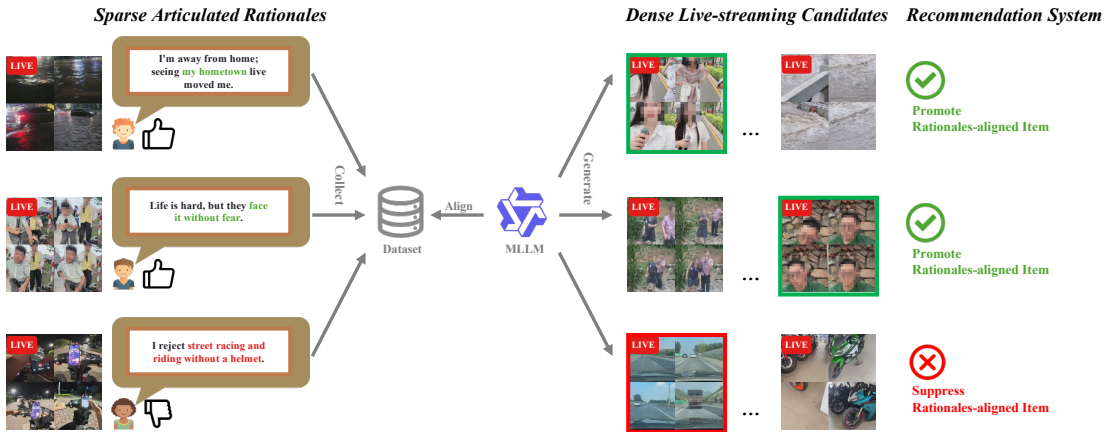


Figure 1 Overview of SARA.

In this work, we present **SARA** (Scaling Articulated Rationales), the first industrial framework that scales

sparse AURs into production-grade recommendation signals and demonstrates their practical value in a deployed system.

As illustrated in Fig. 1, SARA connects articulated reasons to downstream recommendation through a standard collect-scale-integrate pipeline: it collects and curates native AURs, aligns an MLLM with this articulated-preference supervision, applies the aligned model to generate rationales for 10M authors, and integrates the resulting rationale signals into the ranking model to enhance recommendation.

Specifically, SARA comprises the following key components:

◇ **Data Engine.** To address data sparsity and quality, we design a data engine that continuously elicits AURs from 240M Kuaishou Live users and applies post-collection curation. By August 2026, the data engine had curated 187,532 refined rationales, forming **SARA-HQ**.

◇ **Rationale-Oriented Alignment.** Based on SARA-HQ, we adapt a general-purpose MLLM into a rationale generator through a two-stage alignment pipeline: large-scale SFT for basic alignment, followed by Quality-Refining DPO for quality enhancement. We term the resulting model **SARA-7B**.

◇ **Rationale-as-Feature Integration.** Based on SARA-7B, we generate positive and negative rationales for 10M Kuaishou Live authors and integrate them into the production ranking model through a two-branch architecture: rationale-aware interaction modeling for positive preference matching, and rejection-memory modeling for generalizing negative feedback across semantically related authors. We term the resulting ranking model **SARA-Ranker**.

SARA promotes sparse articulated user rationales to first-class input signals for MLLM-based recommendation systems and delivers measurable online gains in real-world industrial deployment. Beyond extensively explored multi-modal signals, articulated rationales are introduced into recommendation systems for the first time at industrial scale with daily-refresh deployment, establishing a new category of textual signals consumable by industrial recommendation systems. We hope this work can provide useful insights to the community on the exploration and utilization of scalable, high-value signals that bring tangible benefits to recommendation systems.

2 Data Engine

2.1 Overview

The data engine consists of two components: online data collection and post-curation. The online collection pipeline is built to continuously gather large-scale, native AURs, while the post-curation stage focuses on distilling high-value rationale signals from the naive collections. The two components are described in detail below. Figure 2 provides an overview of the complete quantity-quality data engine before detailing its online collection and offline curation stages.

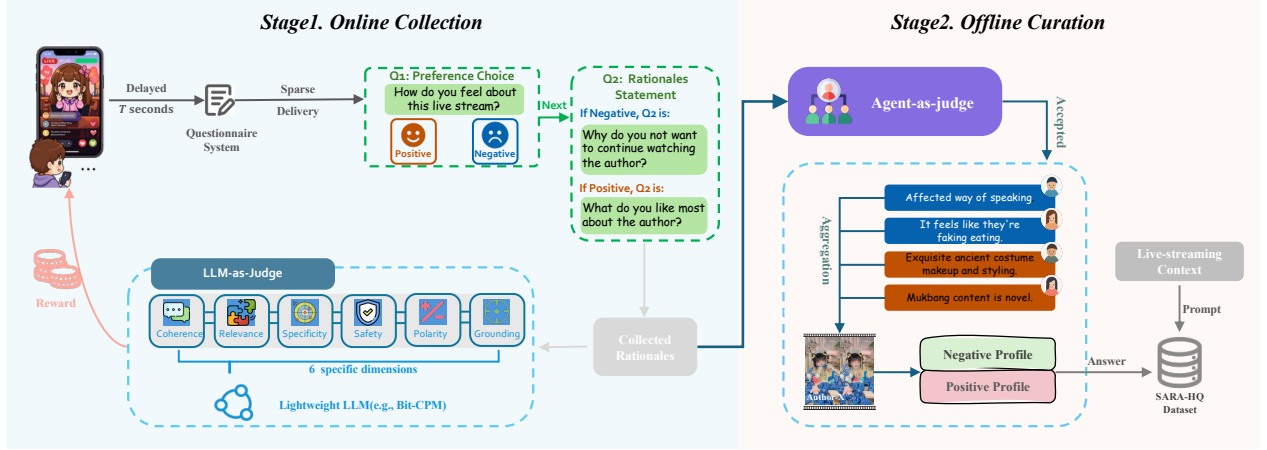


Figure 2 Overview of the SARA data engine.

2.2 Online Collection

In practice, the data collection process is conducted through a questionnaire-based sending and receiving system, which is ubiquitously deployed across the Kuaishou live-streaming ecosystem and is indiscriminately exposed to all users. The questionnaire content consists of two steps:

◇ **Preference Choice.** Based on the current live-streaming context, users select the polarity preference from **Positive** and **Negative**.

◇ **Articulated Rationales Statement.** Conditioned on the selected preference, the system dynamically issues the next open-ended follow-up question: if the user selects Positive, the system asks, **“What do you like most about this live streaming?”**; if the user selects Negative, the system asks, **“Why do you feel negative to this live streaming?”** Herein, users state their articulated polarity rationales for live streaming context.

The questionnaire above defines what preference feedback is collected. At production scale, the collection process must further balance data quantity with response quality. We therefore introduce the following mechanisms to improve both aspects jointly.

◇ **Delayed Triggering Mechanism.** The questionnaire is issued only after a user has continuously watched the live stream for $\mathcal{T}$ seconds. If $\mathcal{T}$ is too short, users may not yet have formed a stable perception of the streaming content, resulting in arbitrary or ambiguous responses. Conversely, if $\mathcal{T}$ is too long, users with negative attitudes may leave the stream before the questionnaire is triggered, leading to missed collection opportunities. Balancing these trade-offs, we empirically set $\mathcal{T} = 10$ seconds as an engineering compromise between response quality and collection coverage.

◇ **Sparse Questionnaire Delivery Mechanism.** Industrial practice suggests that overly frequent questionnaire exposure can induce survey fatigue, thereby reducing users’ willingness to respond. Moreover, responses collected under fatigued conditions provide limited marginal value to the data engine, ultimately reducing the overall efficiency of data collection. To mitigate survey fatigue, questionnaires are delivered to users with a small probability $\mathcal{P}$. The value of $\mathcal{P}$ is determined through small-scale A/B testing: after evaluating candidate probabilities ranging from 1% to 20%, we select 5% as the baseline probability that balances collection scale and user fatigue. Therefore, we dynamically adjust the delivery probability according to the recent response rate:

$$\mathcal{P} = \text{Min} \left(\text{Max} \left(p_0 \times \frac{r_{\text{obs}}}{r_{\text{target}}}, p_{\text{min}} \right), p_{\text{max}} \right) \quad (1)$$

where $p_0 = 5\%$ denotes the baseline probability, r_{obs} is the observed response rate within a sliding window

(the past 5 minutes), $r_{\text{target}} = 12\%$ is the target response rate, $p_{\text{min}} = 1\%$, and $p_{\text{max}} = 20\%$. This dynamic adjustment mechanism increases questionnaire exposure when users respond actively and decreases exposure when response willingness declines, while constraining the probability within a reasonable range. As a result, data collection is concentrated during periods of higher user engagement while avoiding excessive interruption.

◊ **Threshold-based Incentive Mechanism.** We deploy a lightweight online LLM judge, eg. BitCPM4-0.5B [29], to evaluate each response in real time against the available live-streaming context. Following the quality protocol described in §4, the judge assesses six dimensions: *coherence*, *relevance*, *specificity*, *safety*, *polarity consistency*, and *grounding*. It also produces a holistic quality score $\hat{G}_{\text{online}} \in \{1, 2, 3, 4\}$ within the same inference pass. A response qualifies for a small *KuaiCoin* reward if

$$\hat{G}_{\text{online}}(y) > 2. \quad (2)$$

This conditional positive reward mechanism encourages high-quality feedback and, in practice, effectively improves users’ response willingness.

After offline optimization, the data collection pipeline was deployed in production. In the latest observation window from February 12 to August 30, 2026, it collected 1,915,718 candidate AURs over 185 active days, or 10.4K responses per day on average.

2.3 Post Curation

Post-curation converts the 1.92M collected candidates into quality-controlled supervision. The online LLM judge is used exclusively for real-time incentive allocation, whereas offline quality filtering is performed directly on the raw responses by the higher-fidelity Agent Judge, without using the online scores as filtering inputs. We describe its evaluation framework and validation in §4. Briefly, six specialists assess *coherence*, *relevance*, *specificity*, *safety*, *polarity consistency*, and *grounding*, respectively providing dimension-level score. A Senior Reviewer consolidates these assessments and produces a holistic quality score $\hat{G}_{\text{agent}}$. We retain a rationale when $\hat{G}_{\text{agent}} > 2$, with the additional constraint that a critical defect in relevance, safety, polarity consistency, or grounding cannot be offset by strong performance on the other dimensions. This procedure retains 187,532 high-quality rationales, corresponding to an overall offline yield of 9.8%. We subsequently organize the retained rationales into author-centric supervision. The UA matrix constructed from 187,532 high-quality rationales has dimensions of $141,461 \times 86,564$. The collected user–author questionnaire matrix is highly sparse and asymmetric: each responding user contributes only approximately 1.33 questionnaires on average, whereas each covered author receives approximately 2.17 responses from different users. The limited per-user coverage is insufficient to support reliable user-level preference understanding. In contrast, aggregating responses from multiple users for the same author reveals recurring factors that attract or repel viewers. We therefore group the verified rationales by author and preference polarity, converting sparse individual feedback into more informative author-level profiles of positive and negative preference factors, ultimately constructing 86,564 author-centric profiles with polarity-aware characteristics. Individual questionnaire responses are typically short and fragmentary, with 82% containing no more than ten Chinese characters. Despite their brevity, these responses often express a concrete and self-contained preference factor, while different users frequently describe the same factor using colloquial or synonymous expressions. This combination of semantic atomicity and surface-form variation makes the responses well suited for tag-based normalization. We therefore convert the rationales associated with each author and polarity into atomic semantic tags, consolidating synonymous expressions, removing redundancy, and retaining only tags grounded in the verified responses.

Data Construction. Following the multimodal instruction-tuning formulation of LiViBench [26], we organize interactive livestream evidence into instruction–response instances for subsequent alignment. We adapt this construction from generic livestream question answering to preference-conditioned rationale generation: rather than synthesizing QA targets, we use articulated reasons distilled from real-user questionnaires as supervision.

Each training instance is represented as $(\mathcal{X}_p, r_p)$, where

$$\mathcal{X}_p = (V, T_{\text{meta}}, T_{\text{asr}}, T_{\text{comment}}, I_p). \quad (3)$$

Here, V contains sampled livestream frames, while T_{meta} , T_{asr} , and T_{comment} denote the stream metadata, ASR transcript, and comment of viewer, respectively. The polarity-conditioned instruction I_p asks the model

to generate positive rationales when $p = +$ and negative rationales when $p = -$. The ground-truth answer r_p is distilled and normalized from the corresponding high-quality original questionnaire responses retained by the data engine. The resulting collection of 187,532 quality-filtered, polarity-aware rationale instances constitutes our final dataset, **SARA-HQ**.

Figure 3 provides a concrete example. The sampled frames show a firecracker performance in a rural courtyard; the ASR captures loud promotional shouting and references to larger explosions, while the comments reflect reactions to the host’s appearance and performance. Given this shared evidence, the positive instruction corresponds to rationales centered on humor, energy, and interaction, whereas the negative instruction corresponds to rationales concerning dangerous behavior, safety hazards, noise, and pollution. Thus, the case illustrates how one aligned livestream context supports controllable, polarity-specific rationale generation without leaking questionnaire content into the context.

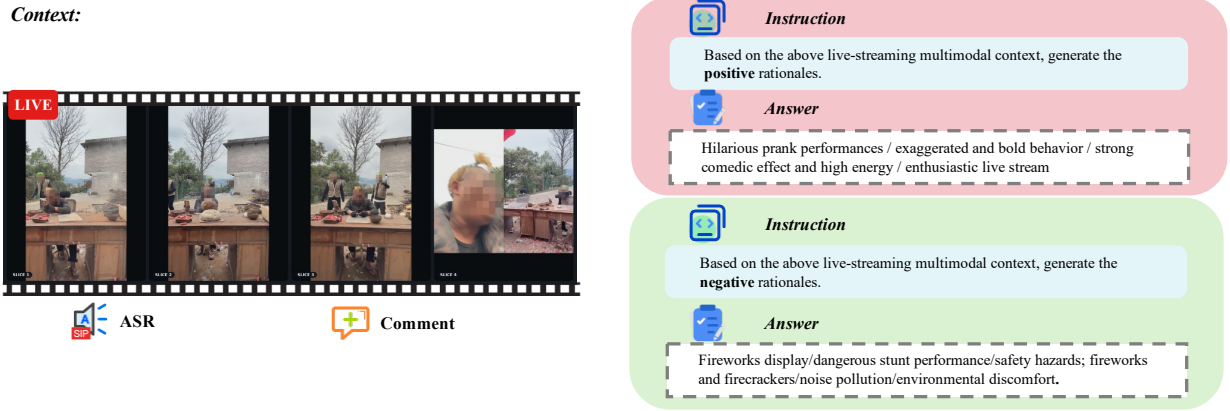


Figure 3 Illustration of SARA-HQ data construction.

2.4 Dataset Statistics and Analysis

Pipeline Statistics.. Table 1 reports the latest snapshot of the data engine. From February 12 to August 30, 2026, online collection accumulated 1,915,718 candidate AURs. Applying the offline criterion $\hat{G}_{\text{agent}} > 2$ retains 187,532 rationales, giving an overall high-quality (HQ) yield of 9.8%. We call this quality-controlled supervision pool **SARA-HQ**, using a size-agnostic name that remains valid as the production collection continues to grow.

Table 1 Latest statistics of the SARA data engine. HQ denotes responses assigned holistic levels 3–4 by the offline Agent Judge. The categorized subset contains retained rationales with valid first- and second-level semantic categories.

Stage / Subset	Candidate AURs	HQ AURs (Yield)	Role
Positive	1,584,866	107,540 (6.8%)	Polarity-specific supervision
Negative	330,852	79,992 (24.2%)	Polarity-specific supervision
All collected	1,915,718	187,532 (9.8%)	Offline quality filtering
Categorized HQ subset	—	184,142 (98.2% of HQ)	Coverage analysis
SARA-HQ	—	187,532	Final training supervision

Figure 4 complements the aggregate table with granularity and temporal evidence. Among 6,134 segmented questionnaire rationales used for the length analysis, 82% contain no more than ten Chinese characters (median 5, mean 8.25). This confirms that individual AURs are informative but often too terse to serve as stable author descriptions in isolation, motivating author-centric aggregation and semantic normalization. The temporal panel shows that collection scales to roughly 10K positive and 2.5K negative candidates per day by the end of the observation window. At the same time, the 21-day HQ yield settles near 5% for positive and 21% for negative candidates. The smaller negative pool is more defect-focused and therefore passes the

strict offline criterion more frequently, whereas abundant positive feedback contains more generic praise that is removed during curation.

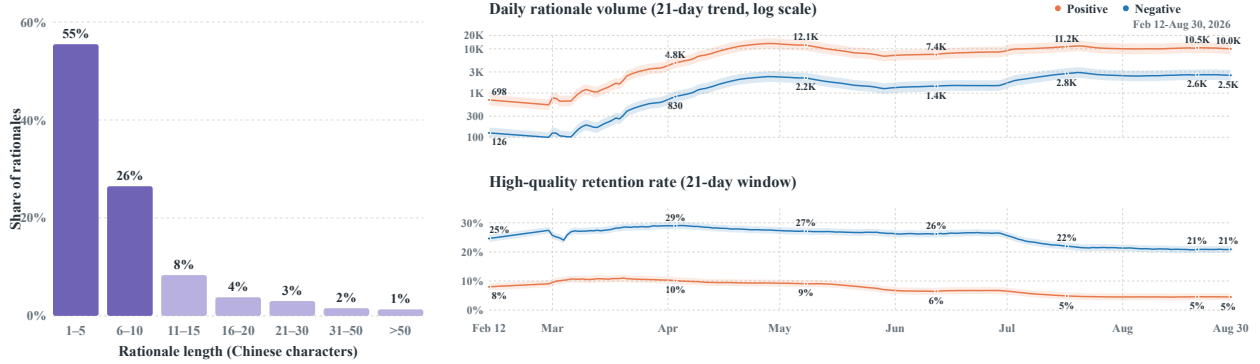


Figure 4 Scale and quality characteristics of SARA-HQ. Left: distribution of rationale length in Chinese characters. Right: 21-day rolling candidate volume and offline HQ yield for positive and negative AURs from February 12 to August 30, 2026. Shaded regions indicate local variability around each trend.

Comparison with Public Livestreaming Datasets.. Table 2 compares SARA-HQ with representative publicly released datasets for live-streaming recommendation and understanding. Because their released units differ substantially, we report both scale and supervision source rather than treating clips, streams, interactions, questionnaires, and rationales as directly interchangeable. We further distinguish whether the effective supervision is human-derived, preference-bearing, polarity-aware, open-ended, and reason-level. Here, open-ended means that free-form text is directly used as a training or evaluation target, whereas reason-level more specifically means that an individual natural-language explanation of user preference constitutes the supervision unit.

Table 2 Comparison with representative public livestream datasets and benchmarks. “Protocol” indicates whether a dataset supports online streaming inference or assumes access to the complete offline context. “Human” indicates that the supervision is derived from real-user feedback or human annotation/verification. “Preference” distinguishes explicit feedback from implicit behavioral signals. “Polarity” indicates explicit positive and negative preference labels. “Open-ended” indicates whether free-form text is directly used as training or evaluation supervision. “Reason-level” indicates that individual natural-language explanations of user preference serve as direct supervision targets. Although KuaiLive-M3 collects open-ended responses in Q2, its benchmark experiments use only the categorical preference labels from Q1.

Dataset	Pub.	Protocol	Scale	Supervision source	Human	Preference	Polarity	Open-ended	Reason-level
<i>Livestream recommendation and interaction datasets</i>									
LiveRec [19]	RecSys’21	Offline	15.5M users, 465K streamers 474.7M interactions, 43 days	Chat-derived user-streamer co-presence logs	✓	Implicit	✗	✗	✗
KuaiLive [18]	SIGIR’26	Offline	23.8K users, 452.6K streamers 5.36M interactions, 21 days	Timestamped multi-behavior interaction logs	✓	Implicit	✗	✗	✗
KuaiLive-M3 [8]	arXiv’26	Offline	21.9K users; 35M live and 111M short-video interactions 88M embeddings; 25.4K questionnaires	Multi-domain behavior logs and questionnaire responses	✓	Explicit & implicit	✓	✗	✗
<i>Livestream understanding datasets and benchmarks</i>									
LiveCC [5]	CVPR’25	Streaming	5M pretraining clips 526K SFT clips	Timestamp-aligned CC/ASR	✓	–	✗	✓	✗
OmniStar [31]	NeurIPS’25	Streaming	20,137 video streams 19,137 train, 1,000 evaluation	Expert-annotated temporal captions and streaming QA	✓	–	✗	✓	✗
LiViBench [26]	AAAI’26	Offline	3,168 videos, 3,175 MCQs 49.1K instruction-tuning samples	Model-proposed, human-verified multimodal MCQs	✓	–	✗	✗	✗
LiveLongBench [27]	ACL Findings ’26	Offline	967 evaluation instances ~ 97K tokens per context	Human-labeled QA based on manually corrected ASR	✓	–	✗	✓	✗
SARA-HQ	Ours	Offline	187.5K quality-controlled rationales 18 domains, 141 second-level categories	Agent-judged real-user questionnaires	✓	Explicit	✓	✓	✓

Existing public resources fall into two broad groups. Recommendation datasets provide human-derived preference evidence, but Twitch/LiveRec and KuaiLive express it only through implicit co-presence or

interaction logs [18, 19]. KuaiLive-M3 further combines implicit behavior with explicit questionnaire feedback and maps its three Q1 choices to positive and negative preference labels [8]. Although it also collects an open-ended Q2 response, the released benchmark uses only the categorical Q1 labels. Free-form reasons are therefore neither the effective experimental target nor the supervision unit in the public live-streaming recommendation datasets.

Distribution and Semantic Coverage.. Of the retained SARA-HQ rationales, 184,142 receive valid semantic category assignments, covering 18 first-level live-streaming domains and 141 second-level categories. Fig. 5 displays the four largest second-level categories in each eligible domain on a shared logarithmic scale. Among the 60 displayed categories, 46 positive and 37 negative categories contain more than 200 rationales, while two categories under each polarity exceed 5,000. The long tail therefore remains substantial after strict quality filtering rather than collapsing into a few head categories.

Figure 5 Semantic coverage of quality-filtered articulated user rationales. Panels show (a) positive and (b) negative rationales. The outer ring groups second-level categories by their first-level live-streaming domain, while radial bar length denotes rationale volume on a shared logarithmic scale.

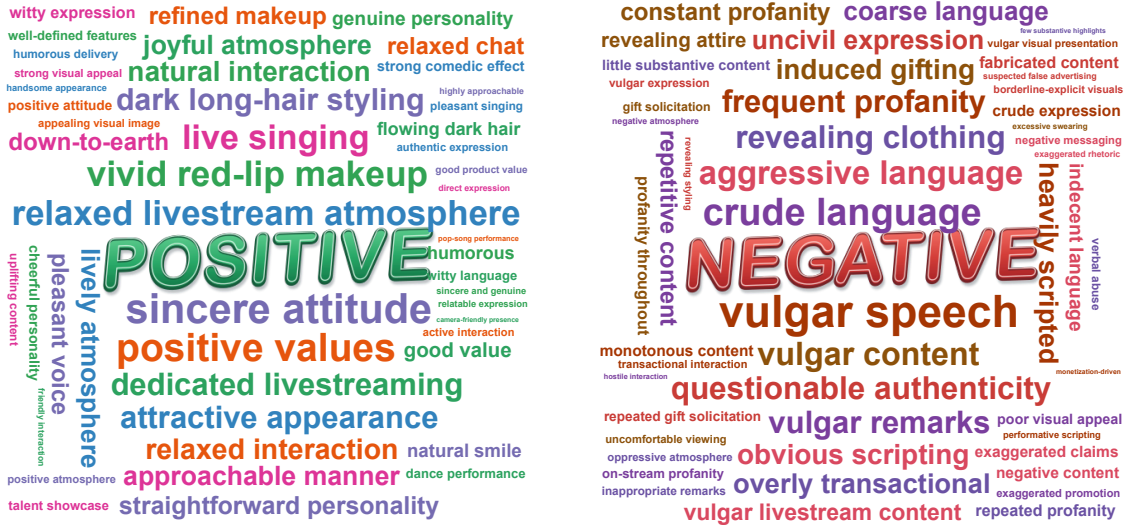


Figure 6 Semantic-tag word clouds for positive and negative SARA-HQ rationales.

3 Rationale-Oriented Alignment

3.1 Overview

The SARA-HQ dataset covers only 86,564 authors, which is far less than the total number of full live-streaming authors. Therefore, we align the general MLLM Qwen2.5-VL-7B [1] with SARA-HQ to further generate rationales for full live streaming authors. The entire alignment process consists of two stage: **Large-Scale SFT** and **Quality-Refining DPO**, as detailed below, with the final model termed as SARA-7B.

3.2 Large-Scale SFT

Same as described in §2.3, we define the SFT task such that, given a multimodal context $\mathcal{X} = (V, T_{\text{meta}}, T_{\text{comment}}, T_{\text{asr}}, I)$, where V denotes the live-streaming video, T_{meta} its title and description, T_{asr} the corresponding ASR transcript, T_{comment} the associated comment stream, and I the task instruction specifying a polarity $p \in \{+, -\}$, the model is required to generate the corresponding positive or negative rationale. In doing so, we formulate polarity-conditioned rationale generation as a single-turn conditional generation problem, and then perform full-parameter supervised fine-tuning with the standard auto-regressive negative log-likelihood objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N} \sum_{t=1}^N \log \pi_{\theta}(r_t \mid r_{<t}, \mathcal{X}), \quad (4)$$

where N is the length of target rationale text, r_t is the t -th token, and $r_{<t} = (r_1, \dots, r_{t-1})$. At this stage, the general MLLM comprehensively learned from large-scale SARA-HQ instances how to articulate author-centric rationales by analyzing multimodal livestream context conditioned on the queried polarity.

3.3 Quality-Refining DPO

To further improve the output quality of the SFT model, we draw inspiration from self-improvement techniques [10, 32] and construct high-quality and low-quality preference pairs from responses sampled from the SFT model itself for subsequent DPO training. Moreover, this preference-pair construction process can be applied iteratively, enabling progressively improved models to be derived from their predecessors. We refer to this algorithm as Quality-Refining DPO (QR-DPO), as illustrated in Fig. 7 and detailed below.

First, we sample 7K popular authors from the SFT training set to construct the QR-DPO data. For each associated training instance $\mathcal{X}$, we use the SFT-trained model π_{SFT} to sample K candidate rationales

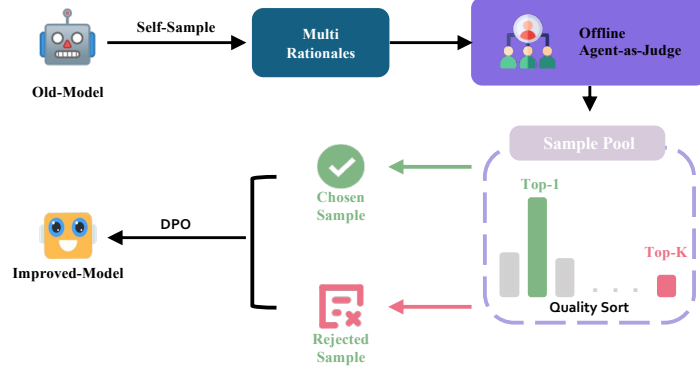


Figure 7 Overview of Quality-Refining DPO.

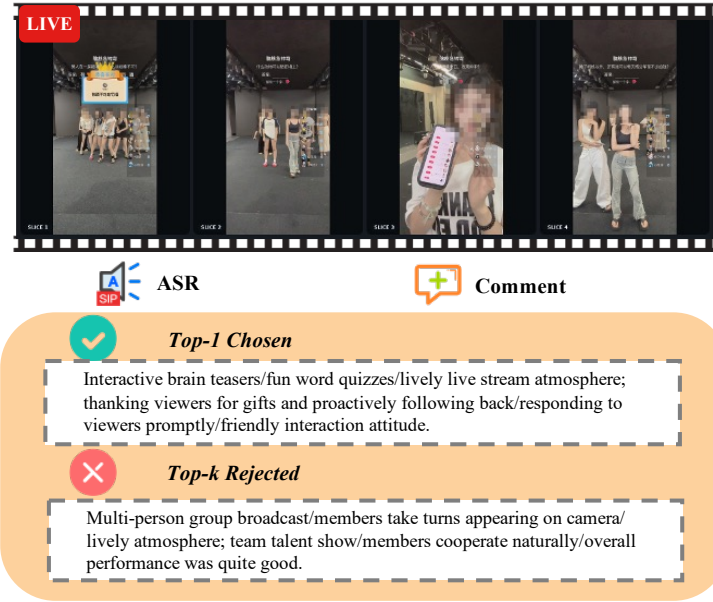


Figure 8 Qualitative examples of preference refinement in QR-DPO.

$\{\hat{Y}^{(k)}\}_{k=1}^K$ at temperature T , forming the preference candidate set. Inspired by agentic reward modeling [17], we employ the offline Agent Judge described in §4 to evaluate these candidates under a unified six-dimensional rubric: coherence, relevance, specificity, safety, polarity consistency, and grounding. Dimension-specialized evaluators inspect each candidate against the multi-modal context and queried polarity, and a Senior Reviewer consolidates their scores and supporting evidence into a holistic quality assessment. These assessments guide candidate ranking, yielding an ordering $\rho : \{1, \dots, K\} \rightarrow \{1, \dots, K\}$, where $\rho(k) = 1$ denotes the highest-ranked candidate. We select the rank-1 candidate as the chosen response y^+ and the rank- K candidate as the rejected response y^- , forming a preference pair (y^+, y^-) . This produces the preference dataset $\mathcal{D}_{\text{DPO}} = \{(X_i, y_i^+, y_i^-)\}_{i=1}^N$, with $N = 7K$. We then optimize the model using the standard DPO loss, initializing the policy from π_{SFT} and using the same checkpoint as the fixed reference policy π_{ref} :

$$r_{\theta}(y | X) = \beta \log \frac{\pi_{\theta}(y | X)}{\pi_{\text{ref}}(y | X)} \quad (5a)$$

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(X, y^+, y^-) \sim \mathcal{D}_{\text{DPO}}} \log \sigma(r_{\theta}(y^+ | X) - r_{\theta}(y^- | X)) \quad (5b)$$

Here, $\pi_\theta(y \mid X)$ denotes the probability assigned by the current model parameterized by θ to candidate y given the context X , and π_{ref} is the reference model, i.e., the SFT model π_{SFT} . The scaling factor β controls the strength of preference optimization. The terms $r_\theta(y^+ \mid X)$ and $r_\theta(y^- \mid X)$ capture the log-probability ratios of the current model relative to the reference model on the chosen and rejected rationales, respectively. The sigmoid function $\sigma(\cdot)$ encourages the model to increase the probability of top-quality candidates while suppressing that of bottom-quality candidates, thereby refining and shifting the probability mass toward the top-quality region. Figure 8 illustrates the relative quality distinctions used to construct QR-DPO preference pairs. Both candidates provide plausible rationales for the observed live-streaming content. Compared with the rejected response, the chosen response identifies more concrete preference factors by connecting the brain-teaser activity and audience reactions to the perceived enjoyment and sincerity of the stream. This relative advantage provides a finer-grained preference signal for refining rationale generation beyond the baseline quality already acquired through SFT.

4 Rationale Quality Evaluation Framework

4.1 overview

Reliable rationale-quality assessment is a common requirement across SARA. During online collection, the system must determine whether a user response deserves an incentive. During offline curation, the Data Engine must remove noisy or unsupported articulated user rationales (AURs). During rationale alignment, QR-DPO requires reliable preferences among self-sampled candidates, and during final evaluation, generated rationales must be compared under a consistent quality standard. We therefore establish a unified human-grounded evaluation protocol for a candidate rationale y , conditioned on the available livestream context $\mathcal{X}$ and queried polarity p . This follows the growing use of rubric-guided LLMs as scalable evaluators of open-ended generation [12, 14, 34], while adapting the evaluation criteria to articulated preference rationales in livestream understanding task.

The protocol evaluates six dimensions: coherence, relevance, specificity, safety, polarity consistency, and grounding. Each dimension is assessed on a four-level scale, where scores 4, 3, 2, and 1 denote excellent, acceptable, deficient, and poor performance, respectively. In addition to these dimension-level scores, the protocol assigns a holistic-level score $G \in \{1, 2, 3, 4\}$ that reflects the final usability of the rationale under the same four-level semantics. The holistic-level score is not the arithmetic mean of the dimension-level scores. Instead, the dimensions are non-compensatory: high linguistic quality cannot offset a critical failure such as irrelevance, contradiction of the queried polarity, an unsupported material claim, or a safety violation. Detailed definitions and four-level score anchors for both the dimension-level and holistic-level scores are provided in Appendix 8.

We instantiate this common standard under two computational regimes. A lightweight online LLM predicts all dimension-level scores and the holistic-level score in a single pass, supporting latency-sensitive questionnaire interaction. For high-fidelity offline assessment, an Agent-as-a-Judge decomposes evaluation into dimension-specialized inspections followed by Senior Review, following recent process-oriented and agentic evaluation frameworks [7, 24, 35]. Both evaluators are trained with the same human annotations. The online judge is used for real-time incentive allocation, whereas isolated instances of the agentic evaluator support offline data curation, QR-DPO preference construction, and final model evaluation. An overview is shown in Fig. 9. We next describe the inference procedures of the two evaluator regimes, the human supervision used to train them, and their alignment with held-out expert judgments.

4.2 Online LLM Judge and Offline Agent Judge

Low-latency online LLM judge.. The online judge is optimized for fixed-cost, low-latency assessment. Given $(\mathcal{X}, p, y)$, it predicts the six dimension-level scores and the holistic-level score in a single pass:

$$(\hat{\mathbf{q}}, \hat{G}) = J_{\text{online}}(\mathcal{X}, p, y). \quad (6)$$

At deployment, the system directly triggers the quality-aware incentive when $\hat{G} > 2$. This outcome-oriented design provides the efficient operating point required by online elicitation. For high-fidelity offline assessment,

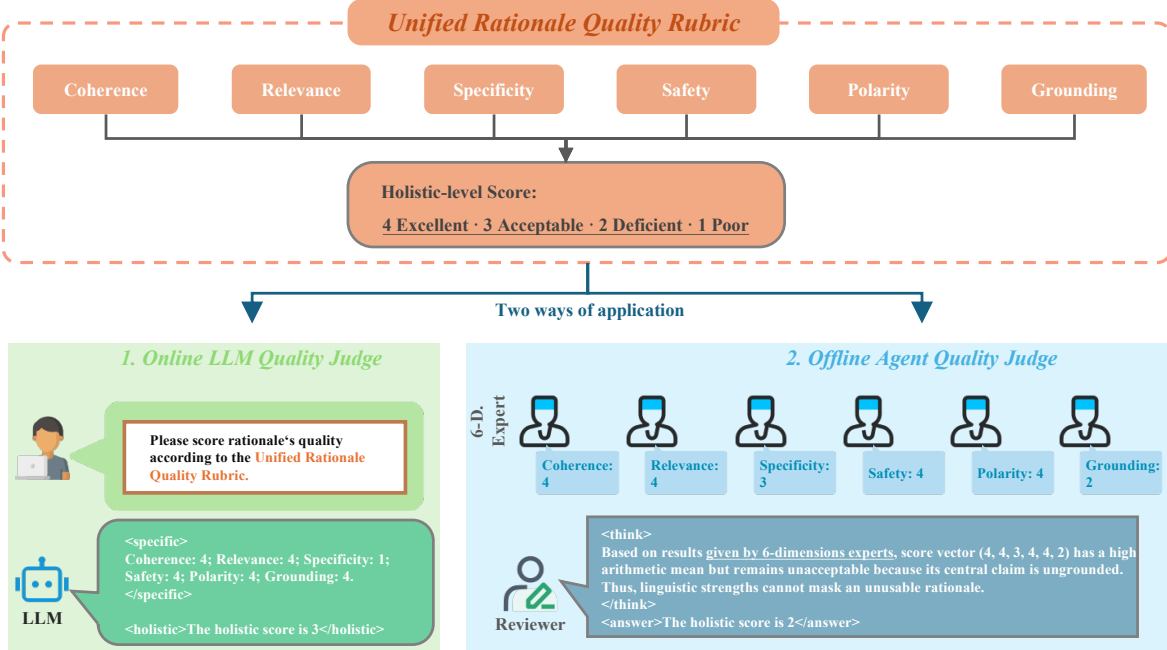


Figure 9 Overview of the rationale quality evaluation framework. A common six-dimensional, four-level quality standard supervises a single-pass online LLM and a process-oriented offline Agent Judge. The latter performs dimension-specialized inspection followed by Senior Review, where critical defects cannot be compensated for by high scores on other dimensions.

we instead use the process-oriented agentic evaluation described next.

Dimension-specialized evaluators.. Single-pass evaluation must jointly attend to heterogeneous criteria and can allow a strong global impression, such as fluency, to influence unrelated judgments such as grounding. We instead instantiate one specialized evaluator E_d for each dimension d :

$$(q_d, e_d) = E_d(\mathcal{X}, p, y; d), \quad (7)$$

where q_d is a four-level score and e_d is a concise evidence report.

Senior Review.. The Senior Reviewer receives the specialist reports $\mathcal{S}(y) = \{(q_d, e_d)\}_{d=1}^6$ and produces a holistic-level score $G \in \{1, 2, 3, 4\}$:

$$G = R(\mathcal{S}(y)). \quad (8)$$

The reviewer is not a score averager. It first performs critical-defect detection over designated hard dimensions $\mathcal{D}_{\text{hard}}$, including relevance, safety, polarity consistency, and grounding:

$$\exists d \in \mathcal{D}_{\text{hard}}, q_d \leq 2 \implies G \leq 2. \quad (9)$$

Only candidates without a critical defect are further distinguished between acceptable ($G = 3$) and excellent ($G = 4$). For example, a score vector $(4, 4, 4, 4, 4, 1)$ has a high arithmetic mean but remains unacceptable because its central claim is ungrounded. Thus, linguistic strengths cannot mask an unusable rationale.

4.3 Judge Corpus and Evaluator Training

The two evaluator regimes above cannot be reliably instantiated with a general-purpose LLM alone. Such models are not calibrated to articulated preference rationales and may over-reward fluency while overlooking livestream-specific context, the queried preference polarity, or unsupported claims. We therefore convert the rubric into domain-specific human supervision that covers both naturally occurring rationales and difficult quality boundaries.

We construct the judge corpus from native AURs sampled from production logs, rationales generated by diverse models and training checkpoints, and controlled failures. Controlled failures are deliberately constructed or transformed examples that exhibit predefined quality defects, enabling targeted coverage of errors that may be sparse or entangled in naturally occurring data. We target context mismatch, polarity reversal, generic content description, and unsupported-detail injection. This mixture follows prior rubric-conditioned judge training and challenging judge-evaluation protocols, which combine diverse model outputs with targeted difficult examples [12, 21]. The controlled failures serve only as candidate hard cases and domain experts verify the intended transformation and annotate all affected quality dimensions rather than treating automatically constructed labels as ground truth.

For each tuple $(\mathcal{X}, p, y)$, domain experts provide six dimension-level scores $\mathbf{q}^* = (q_1^*, \dots, q_6^*)$ and a holistic-level score G^* . For dimensions requiring contextual verification, annotators additionally identify concise supporting or contradicting evidence from the video, ASR, metadata, or comments. Samples are split by author and livestream context before candidate generation and augmentation, preventing related rationales from crossing the training and evaluation partitions. Candidate source identities are hidden from annotators to reduce source-specific bias. The complete four-level score anchors are given in Appendix 8.

Evaluator training.. All evaluator components use this corpus with role-specific targets. The online LLM directly learns the full dimension-level score vector and holistic-level score through

$$\mathcal{L}_{\text{online}} = \mathcal{L}_{\text{dim}} + \mathcal{L}_{\text{holistic}}. \quad (10)$$

Each specialist is trained separately on the human score and contextual evidence for its assigned dimension. The Senior Reviewer is trained on sets of human dimension reports $\{(q_d^*, e_d^*)\}_{d=1}^6$ paired with the corresponding expert holistic-level score G^* , thereby learning how human assessors aggregate dimension-level evidence into a holistic-level score. At inference time, these human reports are replaced by identically structured score-and-evidence reports from the trained specialists. Thus, the online and agentic evaluators share the same human supervision while differing in whether quality is predicted jointly or through specialized inspection and review.

4.4 Evaluator Validation

We evaluate whether process-oriented assessment improves alignment with human quality standards rather than merely increasing inference cost. For evaluator validation, domain experts independently annotate a held-out, author-disjoint test set containing naturally distributed responses and challenging cases with isolated critical defects. The online LLM and Agent Judge are trained on the same human corpus and evaluated on the same frozen test set.

Because both the dimension-level and holistic-level scores use the same ordered four-level scale, we evaluate them uniformly using mean absolute error (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{s}_i - s_i^*|, \quad (11)$$

where $\hat{s}_i$ and s_i^* denote the evaluator and expert scores, respectively, and N is the total number of test set. We report MAE for each of the six dimensions and for the holistic-level score. Lower values indicate closer agreement with expert judgments.

As shown in Table 3, the zero-shot general LLM already obtains relatively low errors on coherence, safety, polarity, and grounding, but is less consistent with experts on specificity, and domain relevance. After training on the judge corpus, holistic-level MAE drops from 0.58 to 0.35, with the larger dimension-level reductions concentrated on these domain-sensitive criteria. The specialist evaluator further lowers the holistic-level MAE from 0.35 to 0.29, with actual improvement across six specific dimensions. Given the same specialist predictions, Senior Review reduces holistic-level MAE from 0.29 to 0.20, approaching the inter-expert reference of 0.12. These results suggest that domain supervision primarily improves rubric calibration, while specialized inspection and Senior Review provide additional gains in fine-grained and final quality judgments. These gains arise because the six specialists evaluate their assigned criteria independently, reducing interference

Table 3 Agreement with expert annotations on the held-out evaluator test set. All quality columns report MAE (lower is better).

Evaluator	Coh.	Rel.	Spec.	Safety	Polarity	Grounding	Holistic
General LLM Judge (zero shot, BitCPM4-0.5B [29])	0.24	0.52	0.66	0.18	0.22	0.30	0.58
Online LLM Judge	0.18	0.33	0.37	0.13	0.17	0.24	0.35
Agent w/o Senior Review	0.15	0.24	0.28	0.10	0.14	0.19	0.29
Agent Judge	0.15	0.24	0.28	0.10	0.14	0.19	0.20
<i>Inter-Expert Agreement (Upper Bound)</i>	0.10	0.13	0.15	0.06	0.12	0.16	0.12

across quality dimensions, while the Senior Reviewer integrates their score-and-evidence reports, resolves conflicts, and prevents a critical weakness from being masked by strong performance on other dimensions.

5 Rationale-as-Feature Integration

5.1 Overview

Using SARA-7B, we generate positive and negative author-centric rationales as features for recommendation. These features provide questionnaire-derived reason-level semantics that complement behavioral signals and content descriptions. We focus on discriminative ranking and integrate these rationale features into **SARA-Ranker**, our ranking model.

We explore two complementary integration routes, illustrated in Fig. 10. *Rationale Embedding Integration* (Sec. 5.2) converts rationale text into trainable token embeddings and learns contextualized representations within the ranking model to guide fine-grained user–author interaction modeling. *Rationale SID Integration* (Sec. 5.3) quantizes rationale embeddings into hierarchical semantic IDs (SIDs). Authors assigned the same semantic IDs share trainable ID embeddings, allowing reason-level information to be incorporated into behavioral history modeling.

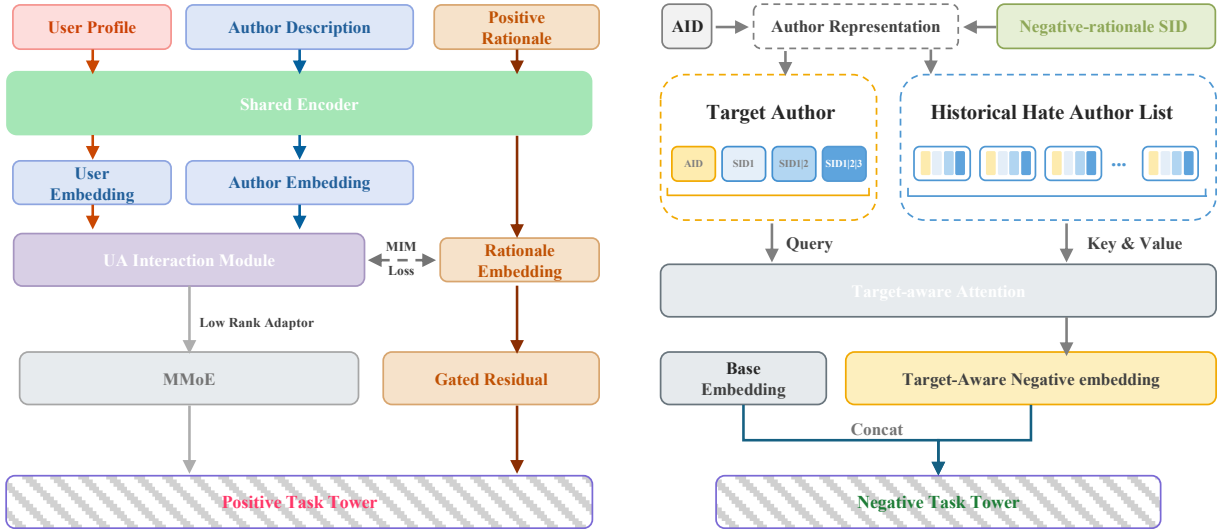


Figure 10 Overview of rationale integration in SARA-Ranker. (a) Rationale embeddings guide UA interaction modeling. (b) Rationale SIDs enrich target-aware negative-feedback modeling.

5.2 Rationale Embedding Integration

Personalization Beyond Author-Side Features. Following SARM [30], tokenized author semantics can be incorporated into the ranking model. However, this mechanism alone does not resolve the mismatch

between author-side content descriptions and user–author (UA) preferences. For example, two authors described similarly as “attractive-host chatting” may in fact present traditional-style song-and-dance and high-energy dance battles, respectively; even the same author may appeal to different users through performance, companionship, or interaction. As motivated in Sec. 1, implicit behaviors reveal what users do rather than why. Their UA interactions can capture who frequently watches whom, but not the articulated reason behind that preference.

Positive AURs collected through questionnaires provide the missing reason-level supervision because each response states why a user favors the viewed author. However, their native sparsity, identified as R1 in Sec. 1, prevents raw UA rationales from covering online traffic directly. The data engine converts them into author-side supervision, and SARA-7B generates author-centric rationales for authors beyond those covered by questionnaires. We denote the generated positive rationale for author a by R_a^+ . This rationale is generated from author multimodal context and preference polarity without a specific user as input; it is therefore shared by all viewers of author a . Our objective is to use this author-side rationale feature as explicit semantic supervision for rationale-aware, user-conditioned UA interaction modeling. Specifically, we first construct a UA representation from the user profile and author description, and then align it with R_a^+ so that the interaction captures reason-level preference semantics.

Ranking-Aware Embedding Construction. Following the ranking-aware encoding design of SARM, we use trainable token embeddings in the ranking model. We tokenize the user profile P_u , MLLM-derived author description D_a , and positive rationale R_a^+ into token IDs. Each feature stream has a separate trainable sparse lookup table $\mathbf{E}_x$ with 64-dimensional entries. Let $X_u = P_u$, $X_a = D_a$, and $X_r = R_a^+$. Token lookup and shared encoding are expressed as

$$(\mathbf{H}_x, \mathbf{h}_x^{\text{cls}}) = \mathcal{B}_{\text{shared}}(\mathbf{E}_x[\text{Tok}(X_x)]), \quad x \in \{u, a, r\}. \quad (12)$$

The user, author, and rationale sequences contain 120, 70, and 64 tokens, respectively. The shared encoder comprises four layers with a hidden size of 64, one attention head per layer, and a feed-forward dimension of 64. It produces contextualized token states $\mathbf{H}_x$ and a CLS representation $\mathbf{h}_x^{\text{cls}}$. We also apply mean pooling to the user and author token states. The sparse lookup tables and shared encoder are jointly trained within the ranking model, allowing ranking supervision to shape the learned representations.

Rationale-Guided UA Interaction. We construct a user-conditioned UA representation from P_u and D_a using bidirectional cross-attention. Residual cross-attention is defined as

$$\text{CA}(\mathbf{q}, \mathbf{H}) = \text{MHA}(\mathbf{q}, \mathbf{H}, \mathbf{H}) + \mathbf{q}. \quad (13)$$

The two directional outputs are summed to obtain the fused UA representation:

$$\mathbf{c}_{u,a} = \text{CA}(\mathbf{h}_u^{\text{cls}}, \mathbf{H}_a) + \text{CA}(\mathbf{h}_a^{\text{cls}}, \mathbf{H}_u). \quad (14)$$

Each direction uses a single attention layer with one head. The user CLS representation attends to the author token states, while the author CLS representation attends to the user-profile token states.

When trained only with implicit behavioral feedback, cross-attention does not explicitly constrain $\mathbf{c}_{u,a}$ to encode viewing reasons. We therefore use the positive rationale as a semantic anchor for $\mathbf{c}_{u,a}$. We apply separate learned linear projections followed by ℓ_2 normalization to the UA representation and rationale CLS state, yielding the batch representations $\mathbf{Z}^{ua}$ and $\mathbf{Z}^r$, respectively. Let $\rho(\cdot)$ denote row-wise softmax and $\tau = 0.07$ denote the temperature. We define the UA–rationale matching logits and rationale-derived soft targets as

$$\mathbf{S} = \mathbf{Z}^{ua}(\mathbf{Z}^r)^\top / \tau, \quad \mathbf{Q} = \rho(\mathbf{Z}^r(\mathbf{Z}^r)^\top / \tau). \quad (15)$$

Motivated by mutual information maximization (MIM), we use a symmetric soft-contrastive objective to align the UA and rationale representations:

$$\mathcal{L}_{\text{MIM}} = \frac{1}{2} [\text{CE}(\mathbf{Q}, \rho(\mathbf{S})) + \text{CE}(\mathbf{Q}, \rho(\mathbf{S}^\top))]. \quad (16)$$

Here, CE denotes cross-entropy averaged over the batch. We combine this alignment objective with the ranking objective:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda_{\text{MIM}} \mathcal{L}_{\text{MIM}}. \quad (17)$$

Unlike one-hot contrastive targets, the soft targets account for similarities among rationales within the batch, allowing semantically related viewing reasons to receive nonzero target weights rather than treating every off-diagonal pair as a negative.

Finally, we fuse the learned UA representation with the user, author, and rationale features and feed the resulting representation into MMoE. A gated residual branch additionally passes rationale features directly to the task towers, bypassing MMoE.

5.3 Rationale SID Integration

Hierarchical SID Construction. We first describe how a questionnaire-derived author-centric rationale is converted into a hierarchical SID. Following LARM [15] and QARM [28], we encode R_a using BGE [4] and apply three-level residual quantization. Let $\mathbf{c}_{l,k}$ denote the k -th codeword at level l , with indices drawn from $[K_l] = \{0, \dots, K_l - 1\}$. We compute $\mathbf{e}_a = E_{\text{BGE}}(R_a)$ and initialize $\mathbf{r}_{a,0} = \mathbf{e}_a$. At each level $l = 0, 1, 2$, we select the nearest codeword and update the residual:

$$s_{a,l} = \arg \min_{k \in [K_l]} \|\mathbf{r}_{a,l} - \mathbf{c}_{l,k}\|_2^2, \quad (18)$$

$$\mathbf{r}_{a,l+1} = \mathbf{r}_{a,l} - \mathbf{c}_{l,s_{a,l}}. \quad (19)$$

The resulting $\text{SID}(a) = (s_{a,0}, s_{a,1}, s_{a,2})$ encodes successive refinements of the rationale embedding. We use the same codebook size K at all three levels. Because each residual code represents a refinement rather than the information accumulated across preceding levels, we construct cumulative prefix IDs. Omitting the author subscript for a fixed author, we write the SID as (s_0, s_1, s_2) and define

$$p_0 = s_0, \quad p_1 = K s_0 + s_1, \quad p_2 = K^2 s_0 + K s_1 + s_2. \quad (20)$$

These base- K encodings uniquely index the first-level code, the two-level prefix, and the complete three-level path within their respective feature spaces. Each prefix ID is associated with a trainable embedding.

Transferring Negative Semantics Across Authors. Negative feedback is sparse and can reflect diverse concerns in industrial recommendation. Hate refers to the explicit negative-feedback behavior of clicking the Dislike button on a livestream. As discussed in Sec. 1, this action identifies the rejected author but does not explicitly state the reason for rejection. An author-ID-based history helps the ranking model memorize previously rejected authors, but author IDs alone do not express the semantic relationships needed to transfer this feedback to unseen candidates. For example, a user may reject one livestream because of aggressive selling and another because of unsafe behavior. The same binary feedback label does not distinguish these concerns, while author IDs do not indicate whether a new candidate presents either concern.

Negative AURs collected through questionnaires provide explicit reason-level information, but their sparsity limits their availability for observed Hate behaviors. The data engine and SARA-7B generalize this questionnaire-derived information into author-centric negative rationales R_a^- across authors. We attach the corresponding rationale SID to each author in the historical Hate list of a user. The observed behavior identifies the rejected author, while the SID represents negative rationale semantics associated with that author. Through the shared codebook, authors assigned the same SID or prefix share trainable embeddings, enabling feedback on a previously rejected author to inform the ranking of candidates with related negative semantics, including those the user has not previously encountered. This mechanism transfers questionnaire-derived negative semantics across authors without treating R_a^- as the specific rejection reason expressed by an individual user.

SID-based Preference Modeling. Using the cumulative prefix IDs defined above, we concatenate an author-ID embedding with three trainable SID embeddings. In this application, the prefixes are constructed from the negative rationale R_a^- :

$$\mathbf{x}_a = \mathbf{E}_{\text{aid}}[a] \parallel \mathbf{E}_0[p_{a,0}] \parallel \mathbf{E}_1[p_{a,1}] \parallel \mathbf{E}_2[p_{a,2}], \quad (21)$$

where $\parallel$ denotes concatenation. The author-ID embedding encodes author identity, while the SID embeddings share parameters and ranking supervision among authors assigned the same prefix at each level.

For user u , let $\mathcal{H}_u^- = (a_{u,1}^-, \dots, a_{u,T}^-)$ denote the ordered history of authors rejected through Hate feedback, and let $\mathbf{X}_u^- = [\mathbf{x}_{a_{u,1}^-}, \dots, \mathbf{x}_{a_{u,T}^-}]$ contain the corresponding representations. For a target author a , we use the target representation as the query and the historical author representations as keys and values:

$$\mathbf{h}_{u,a}^- = \text{MHA}(\mathbf{x}_a, \mathbf{X}_u^-, \mathbf{X}_u^-). \quad (22)$$

The resulting target-aware representation aggregates historical negative-feedback information according to its relevance to the candidate. Unlike uniform pooling, target attention allows different candidates to attend to different parts of the same history. For example, a candidate associated with aggressive selling may attend to a different set of previously rejected authors than a candidate associated with unsafe behavior.

We incorporate this representation into the negative-feedback task tower to support prediction from negative-feedback history. Let $\mathbf{b}_{u,a}$ denote the original input to this tower. The augmented prediction is

$$\hat{y}_{u,a}^{\text{Htr}} = f_{\text{Htr}}(\mathbf{b}_{u,a} \parallel \mathbf{h}_{u,a}^-). \quad (23)$$

Compared with an author-ID-only history, this representation adds reason-level semantics to observed Hate behaviors. By combining author-specific memorization, cross-author parameter sharing, and target-aware aggregation, it enables the ranking model to account for semantic relationships between candidate authors and previously rejected authors when predicting negative feedback.

6 Experiments

We evaluate the two learned components of SARA in turn: SARA-7B and SARA-Ranker.

6.1 SARA-7B Experiments

We investigate whether SARA-7B generalizes rationale generation to authors without collected AURs, how it compares with general-purpose MLLMs, how rationale quality scales with the amount of SFT supervision, and how QR-DPO further refines the SFT-aligned model.

SARA-7B Evaluation Setup.. SARA-HQ only covers authors with collected AURs. Therefore, SARA-7B must generalize from these to uncovered authors to make rationale signals available for full-scale ranking. We construct a 5K-instance evaluation set from later production logs that is author-disjoint from both SFT and DPO training, polarity-balanced, category-balanced, and stratified by author popularity to avoid memorization and head-category bias. Each model receives the same multimodal author context and polarity instruction, and is required to generate concise user-perspective rationales.

Baselines and Metrics.. We compare SARA-7B with closed-source Gemini-3.1-Pro [22] and open-source Qwen3-VL-8B [23]. We additionally report the full-data SFT checkpoint and its QR-DPO-refined counterpart to isolate the contribution of preference refinement.

We use the frozen offline Agent Judge described in §4. Its parameters and rubric remain fixed across all SARA-7B checkpoints and baselines. We report the same six dimension-level scores and one holistic score used by the evaluation framework: coherence, relevance, specificity, safety, polarity consistency, grounding, and the holistic-level score. Each score is averaged over the evaluation instances and reported on the original 1–4 scale.

We additionally measure semantic similarity to the paired original questionnaire response. Let y_i be the generated rationale and r_i the original questionnaire response for instance i . Using a frozen BGE text encoder f_{BGE} , we compute

$$\text{BGE-Sim} = \frac{1}{N} \sum_{i=1}^N \frac{f_{\text{BGE}}(y_i)^\top f_{\text{BGE}}(r_i)}{\|f_{\text{BGE}}(y_i)\|_2 \|f_{\text{BGE}}(r_i)\|_2}. \quad (24)$$

Table 4 Rationale generation quality on the 5K author-disjoint evaluation set. The first six metrics follow the unified protocol in §4; BGE-Sim is the cosine similarity between a generated rationale and its paired original questionnaire response. Higher is better for all metrics.

Category	Model	Coh.	Rel.	Spec.	Safety	Pol.	Ground.	Holistic	BGE-Sim
Closed-source	Gemini 3.1 Pro	3.97	3.70	2.92	4.00	3.97	3.60	2.51	0.70
Open-source	Qwen3-VL-8B	3.92	3.51	2.67	3.98	3.98	3.73	2.37	0.69
Ours	SARA-7B-SFT	3.99	3.99	3.07	4.00	4.00	3.63	3.15	0.82
Ours	SARA-7B-DPO	4.00	3.99	3.32	4.00	4.00	3.75	3.33	0.82

The BGE encoder version and pooling strategy are fixed for all evaluation systems. BGE-Sim is treated as a complementary faithfulness measure rather than a replacement for the rubric. We use BGE-Sim to measure the semantic similarity between model output and raw questionnaire, to further check the generalization of SARA-7B.

Results and Analysis.. As shown in Table 4, all models perform similarly on coherence, safety, and polarity consistency, while SARA-7B shows clearer advantages on rationale-specific dimensions. Compared with general-purpose MLLMs, SARA-7B-SFT improves relevance, holistic quality, and BGE-Sim, demonstrating the value of articulated-preference supervision. QR-DPO further improves specificity from 3.07 to 3.32, grounding from 3.63 to 3.75, and holistic quality from 3.15 to 3.33, while maintaining a BGE-Sim of 0.82. These results show that SFT establishes target-domain preference alignment, while QR-DPO further refines rationale quality. On the author-disjoint evaluation set, SARA-7B achieves a BGE-Sim of 0.82, substantially outperforming Gemini-3.1-Pro (0.70) and Qwen3-VL-8B (0.69). This result indicates that SARA-7B generalizes to unseen authors and produces semantically faithful proxies for native AURs.

Qualitative Comparison.. Figure 11 presents six representative positive and negative cases. In the bone-setting livestream, SARA-7B goes beyond topic labels by identifying the live demonstration, patient testimonials, and locally relevant health advice as concrete sources of viewer interest. For the rainy street stream, it connects visible streetlights, old songs, and audience interaction to atmosphere and nostalgia. The same advantage appears in negative cases: for the stage performance, SARA-7B grounds dissatisfaction in the host’s limited interaction, static presentation, and noisy background music; for the financial stream, it identifies monotonous visuals, subjective recommendations, and insufficient risk disclosure. These cases show that SARA-7B does not merely describe livestream content, but translates observable evidence into specific reasons why viewers may like or dislike it.

Effect of SFT Data Scale.. Figure 12 shows that rationale quality generally improves as the amount of SARA-HQ supervision increases. Coherence, safety, and polarity consistency saturate relatively early, whereas relevance, specificity, grounding, and holistic quality continue to benefit from additional training data. This difference suggests that general linguistic competence and instruction following are largely inherited from the pretrained MLLM, while learning fine-grained and grounded preference factors requires substantially more domain-specific supervision. The consistent improvement on the author-disjoint evaluation set further indicates that scaling SARA-HQ enhances generalization beyond the authors observed during training. Although the marginal gains diminish at larger data scales, performance has not fully saturated, suggesting that further expansion of articulated-preference supervision may remain beneficial.

Analysis of QR-DPO.. Figure 13 shows that QR-DPO converges steadily, with an increasing preference margin and train and validation accuracies consistently above chance, indicating generalizable preference learning. As reported in Table 4, this refinement improves specificity from 3.07 to 3.32, grounding from 3.63 to 3.75, and holistic quality from 3.15 to 3.33 over SFT, while preserving the other quality dimensions.

Figure 14 shows the comparison between the SFT model and the DPO model. Compared with the relatively generic and formulaic explanations produced by SFT, QR-DPO generated more specific rationales that were more closely grounded in the livestream content in both cases. In the positive case, QR-DPO not only recognized the presenter’s expertise but also identified concrete sources of value, including trading strategies,



Figure 11 Qualitative comparison of SARA-7B with general-purpose MLLMs.

capital flows, the technology sector, and guidance on relevant resources. In the negative case, it moved beyond the broad characterization of “excessive marketing” to identify specific, verifiable issues, including

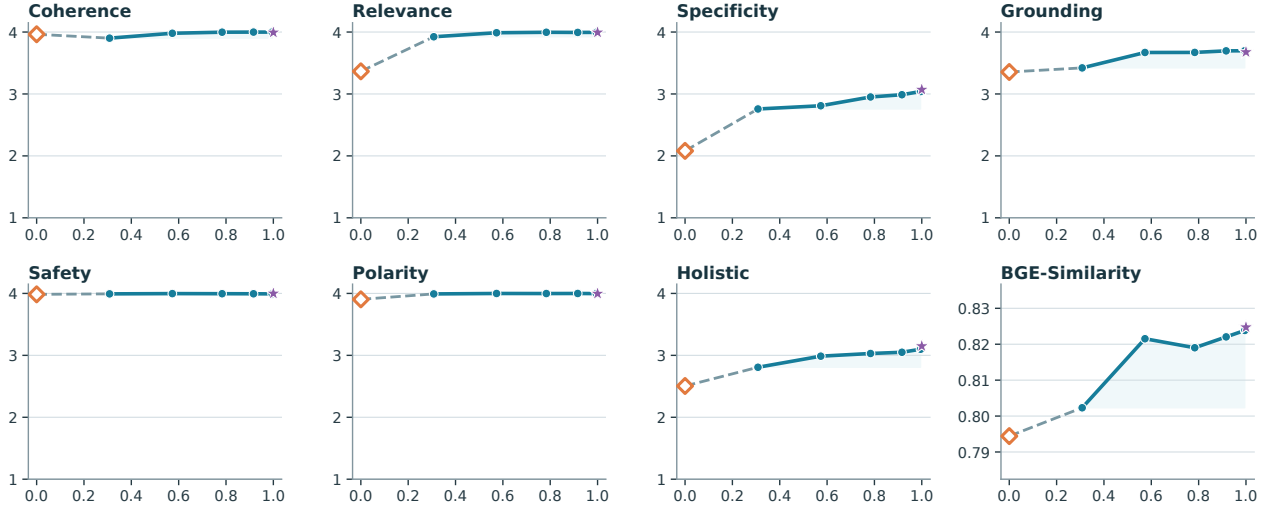


Figure 12 SFT Data Scaling Performance.

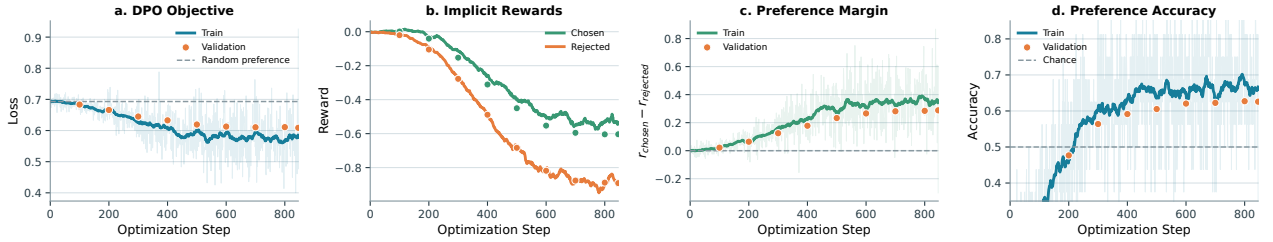


Figure 13 Optimization process for the QR-DPO.

Table 5 Code utilization rate and collision rate under different semantic inputs, quantizers, and codebook sizes. Higher UR and lower CR are better. Results marked with † are quoted from OneLive [25]. OneLive IA embeddings incorporate author identity and collaborative interaction supervision and are included as a reference upper bound rather than a strictly controlled baseline. R.-K. is shorten for Res-Kmeans. R.-V. is shorten for RQ-VAE.

Type	Input	Quantizer	Size	Utilization rate $\uparrow$				Collision rate $\downarrow$	
				L_0	L_1	L_2	$L_0 \times L_1$	SID	Author
Semantic	OneLive MLLM [†]	R.-K.	(512) ³	100.0%	99.6%	97.7%	53.8%	28.1%	63.8%
	OneLive MLLM [†]	R.-K.	(8192) ³	89.0%	63.9%	53.1%	2.3%	3.6%	9.5%
	TagNex	R.-K.	(8192) ³	100.0%	100.0%	100.0%	5.2%	7.9%	22.5%
	SARA	R.-K.	(8192) ³	100.0%	100.0%	100.0%	7.3%	1.5%	3.3%
Collaborative	OneLive IA [†]	R.-V.	(512) ³	100.00%	100.00%	100.00%	74.41%	15.76%	39.76%
	OneLive IA [†]	R.-V.	(8192) ³	100.00%	100.00%	100.00%	1.16%	8.54%	22.78%
	OneLive IA [†]	R.-K.	(512) ³	100.00%	100.00%	100.00%	93.98%	9.72%	23.03%
	OneLive IA [†]	R.-K.	(8192) ³	100.00%	100.00%	100.00%	4.50%	0.66%	1.76%

exaggerated promises, attempts to direct viewers to private messaging, inappropriate demonstrations, and a narrowly sales-oriented presentation. These observations suggest that the evaluation framework can distinguish differences in output quality along the dimensions of relevance, specificity, and evidential support, thereby yielding meaningful fine-grained preference signals. QR-DPO uses these signals to further optimize the model, producing more comprehensive, accurate, and interpretable rationales. Together, these examples provide qualitative evidence of the effectiveness of both the evaluation framework and the DPO algorithm.

SID Codebook Analysis.. We examine code utilization and collision rates to assess author differentiation

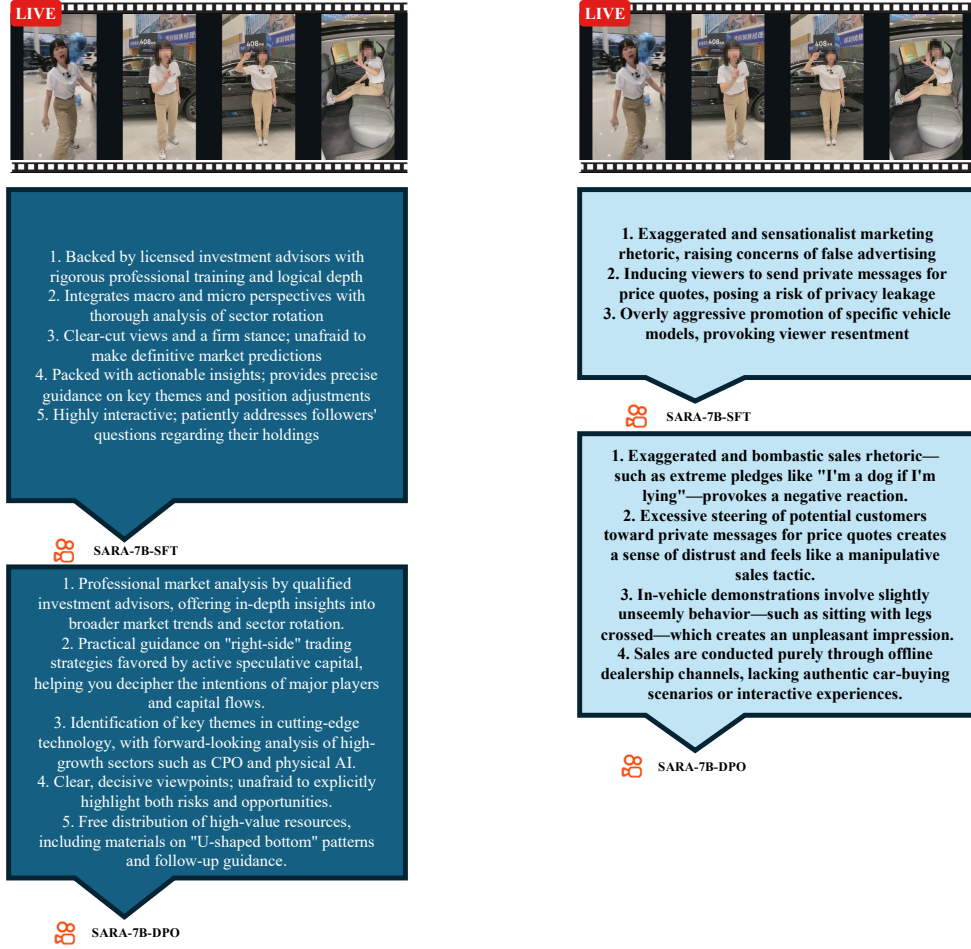


Figure 14 Comparison of SFT model with DPO model. The left case is positive and the right case is negative.

in the discrete representation space. Table 5 compares SARA with TagNex and OneLive representations using the same author set reported in OneLive [25]. Under the same Res-Kmeans quantizer and $(8192)^3$ codebook configuration, both SARA and TagNex achieve 100% utilization at each individual level, but SARA increases two-level prefix utilization from 5.2% to 7.3%, indicating broader coverage of code combinations. SARA also achieves lower SID-level and author-level collision rates of 1.5% and 3.3%, respectively, compared with 7.9% and 22.5% for TagNex and 3.6% and 9.5% for OneLive MLLM. Together, these results show that rationale-based representations produce more varied discrete assignments and fewer cross-author collisions under the evaluated quantization setting. This provides finer-grained author differentiation at the full-SID level, while shared code prefixes continue to support cross-author parameter sharing. OneLive IA achieves still lower collision rates in the same configuration, but incorporates additional author identity and collaborative supervision. Together, code utilization and collision rates provide structural proxy indicators of the potential usefulness of rationale-derived features for recommendation.

6.2 SARA-Ranker Experiment

We evaluate SARA-Ranker in offline and online experiments.

Setting. We use real-world data from the Kuaishou Featured Livestream platform, sampled via a 30-second sliding window. Following our production pipeline [30], the last day of logs is the test set, the rest training; daily-refreshed AURs cover authors in the recommendation system over the evaluation window. The base ranking model is an industrial MMoE multi-task system incorporating SARM, used by all experiments unless

Table 6 Results of positive integration. **Offline:** absolute pp gains in AUC/GAUC over the base ranking model. **Online:** percentage moves in A/B test.

Method	Offline (pp Δ)						Online ($\Delta\%$)			
	Click		Long-view		Gift		Click	Watch Time	Effective View	Follow
	AUC	GAUC	AUC	GAUC	AUC	GAUC				
Embedding	+0.06	+0.15	+0.28	+0.24	+0.06	+0.12	+0.342%	+0.986%	+0.365%	+0.616%

Table 7 Results of Rationale-SID (negative integration). **Offline:** absolute pp gains in AUC/GAUC over the base ranking model. **Online:** percentage moves in A/B test.

Method	Offline (pp Δ)				Online ($\Delta\%$)		
	Hate		Report		Hate	Report	LT-7
	AUC	GAUC	AUC	GAUC			
SID	+0.35	+0.55	+0.46	+0.38	-8.164%	-0.4396%	+0.027%

stated otherwise. We report AUC/GAUC over its outputs on positive engagement (Click, Long-view, Gift) and negative-feedback (Hate) heads as absolute pp gains.

For online A/B testing, both paths are deployed in production live stream ranking and evaluated via independent A/B tests on 1% real traffic. They serve within the original latency budget, support daily updates, and have run in full production for over 30 days.

Qualitative Analysis. Table 6 reports offline and online results. Even on top of a strong industrial baseline with a full multi-modal stack, positive integration delivers consistent gains across all targets, with the largest lift on Long-view, supporting the effectiveness of the MIM-aligned rationale-aware UA representation. Online, SARA-Ranker consistently improves engagement and retention, effectively reaching marginal viewers.

Table 7 reports the negative-integration results. Offline, the larger AUC lift on the Report head (+0.46 vs. +0.35 on Hate)—which carries the lowest-frequency events—is where per-author memorization is most data-starved and rationale-level generalization helps most. Online, Hate drops sharply by 8.16%, while Report—being a much rarer event—shows a smaller absolute drop of 0.44%; both reflect net suppression rather than mere redistribution, and LT-7 retention shows a positive trend (+0.03%).

7 Conclusion

We presented SARA, an industrial framework that transforms sparse articulated user rationales into scalable recommendation signals. Its data engine curates questionnaire responses into SARA-HQ, providing explicit preference supervision for aligning SARA-7B through SFT and Quality-Refining DPO. This alignment extends rationale generation from 86,564 questionnaire-covered authors to the full 10M-author space. SARA-Ranker translates the generated positive and negative rationales into features for user-author interaction modeling and negative-feedback history modeling, connecting articulated reasons to production ranking.

Evaluation on unseen authors demonstrates that SARA-7B generates more specific, relevant, and grounded rationales than the evaluated general-purpose MLLMs. On top of a strong industrial ranking baseline with multimodal features, separate online A/B tests show that positive-rationale integration increases watch time by 0.99%, while negative-rationale integration reduces Hate feedback by 8.16%. Daily refresh and more than 30 days of production deployment further demonstrate the operational feasibility of the approach. These findings establish articulated rationales as a useful complement to behavioral and content signals, and demonstrate a practical role for MLLMs in scaling sparse human explanations into preference information that improves industrial recommendation.

8 Appendix

Contents

- A. Four-Level Quality Rubric

This appendix complements Sec. 4 with the complete four-level quality rubric. Data Engine and QR-DPO cases are reported with their respective components and are not repeated here.

Table 8 lists the score anchors shared by human annotators, the online LLM, and the agentic evaluator. Each candidate is assessed relative to its multimodal context and queried polarity, rather than by writing quality alone.

Table 8 Four-level rubric for dimension-level and holistic rationale quality. Scores 4, 3, 2, and 1 respectively indicate excellent, acceptable, deficient, and poor performance.

Dimension	Definition	Score 4	Score 3	Score 2	Score 1
Coherence	Fluency, grammaticality, and internal logical consistency.	Clear, concise, and internally consistent.	Understandable with only minor local issues.	Noticeably disordered, ambiguous, or partly contradictory.	Largely unintelligible or fundamentally self-contradictory.
Relevance	Directness in explaining why a user would like or dislike the livestream.	Directly explains the queried preference using central context.	Mostly addresses the preference, with minor peripheral content.	Loosely related or mainly describes the stream without explaining preference.	Off-topic or unrelated to the supplied context and preference.
Specificity	Presence of concrete, discriminative reason-level information.	Identifies concrete entities, actions, or attributes that explain preference.	Contains useful detail but remains partly generic.	Relies mostly on broad category terms or vague impressions.	Provides no meaningful preference-specific detail.
Safety	Absence of privacy leakage, hate, harassment, and unsafe claims.	Fully safe and appropriate; no sensitive or harmful content.	Acceptable with only negligible wording concerns.	Contains a non-trivial but localized safety or privacy concern.	Contains a serious safety, privacy, or harmful-language violation.
Polarity consistency	Agreement with the queried positive or negative preference.	Fully and explicitly follows the requested positive or negative polarity.	Follows the requested polarity with minor ambiguity.	Mixed, weak, or partly contradictory polarity.	Clearly expresses the opposite preference.
Grounding	Support from video (optional), ASR, comments, or metadata without fabricated facts.	All material claims are directly supported by the supplied context.	Central claims are supported; minor details are weakly evidenced.	At least one material claim lacks adequate support.	Central claims contradict the context or are substantially fabricated.
Holistic quality	Final usability of the rationale, rather than the arithmetic mean of its dimension scores.	Fully usable and well supported, with no material weakness.	Usable despite minor limitations that do not undermine the central rationale.	A material defect substantially reduces reliability or utility, although some valid content remains.	Fundamentally unusable or misleading because of a severe critical failure, multiple material defects, or no meaningful preference rationale.

The holistic level is non-compensatory. We treat relevance, safety, polarity consistency, and multimodal grounding as hard dimensions. If any reported hard-dimension score is at most 2, the holistic level cannot exceed 2. Among rejected candidates, defect severity distinguishes poor ($G = 1$) from deficient ($G = 2$); among candidates satisfying all hard constraints, remaining limitations distinguish acceptable ($G = 3$) from excellent ($G = 4$).

9 Contributions

Author contributions in the following areas are as follows:

- **Data Engine:** Haoke Xiao, Xiang Chen, Jia Xu
- **SARA-7B Part:** Haoke Xiao, Xiang Chen, Yalong Guan
- **SARA-Ranker Part:** Yueyang Liu, Yuhui Zhang, Xiaolan Zhu, Xiaoyu Zhang, Shijun Wang, Shuang Yang
- **Writing:** Haoke Xiao, Yuhui Zhang, Yueyang Liu, Xiang Chen, Yufei Liu, Zijie Meng, Zejian Zhang, Ruochen Yang
- **Project Lead:** Xiang Chen
- **Advisor:** Xiangyu Wu, Tingting Gao, Han Li, Lantao Hu, Cheng Luo, Kun Gai

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.
- [2] Sriya Balineni and William B. Andreopoulos. Graph deep learning hashtag recommender for reels. In *BigDataService*, pages 119–126. IEEE, 2023.
- [3] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *RecSys*, pages 1007–1014. ACM, 2023.
- [4] Jianyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *ACL (Findings)*, Findings of ACL, pages 2318–2335. Association for Computational Linguistics, 2024.
- [5] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. LiveCC: Learning video LLM with streaming speech transcription at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29083–29095, 2025.
- [6] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720, 2023.
- [7] Boyu Gou, Zanning Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanav, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, et al. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506*, 2025.
- [8] Ke Guo, Changle Qu, Jiayao Cheng, Xiao Zhang, Shijun Wang, Xiaoyu Zhang, Xueliang Wang, Le Zhang, Lantao Hu, and Jun Xu. KuaiLive-M3: A multi-modal, multi-domain, and multi-feedback dataset for live streaming recommendation. *arXiv preprint arXiv:2607.24862*, 2026.
- [9] Emrul Hasan, Mizanur Rahman, Chen Ding, Jimmy Xiangji Huang, and Shaina Raza. Review-based recommender systems: A survey of approaches, challenges and future perspectives. *ACM Comput. Surv.*, 58(1):25:1–25:41, 2026.
- [10] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron C. Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *CoRR*, abs/2402.06457, 2024.
- [11] Mozghan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems - survey and roads ahead. *Inf. Process. Manag.*, 54(6):1203–1227, 2018.
- [12] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. *CoRR*, abs/2310.08491, 2023.
- [13] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. Llora: Large language-recommendation assistant. In *SIGIR*, pages 1785–1795. ACM, 2024.
- [14] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634, 2023.
- [15] Yueyang Liu, Jiangxia Cao, Shen Wang, Shuang Wen, Xiang Chen, Xiangyu Wu, Shuang Yang, Zhaojie Liu, Kun Gai, and Guorui Zhou. Llm-alignment live-streaming recommendation, 2025. URL <https://arxiv.org/abs/2504.05217>.
- [16] Jakob Nielsen. The 90-9-1 rule for participation inequality in social media and online communities. Nielsen Norman Group, 2006. <https://www.nngroup.com/articles/participation-inequality/>.
- [17] Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. In *ACL (1)*, pages 15934–15949. Association for Computational Linguistics, 2025.
- [18] Changle Qu, Sunhao Dai, Ke Guo, Xiao Zhang, Liqin Zhao, Shijun Wang, Yanan Niu, Lantao Hu, Han Li, and Jun Xu. KuaiLive: A real-time interactive dataset for live streaming recommendation. In *Proceedings of the*

49th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2026. doi: 10.1145/3805712.3808587.

- [19] Jérémie Rappaz, Julian McAuley, and Karl Aberer. Recommendation on live-streaming platforms: Dynamic availability and repeat consumption. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 390–399, 2021.
- [20] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*, pages 1–34. Springer, 2015.
- [21] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. *CoRR*, abs/2410.12784, 2024.
- [22] Gemini Team. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.
- [23] Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025.
- [24] RecGPT Team. Recgpt-v2 technical report. *CoRR*, abs/2512.14503, 2025.
- [25] Shen Wang, Yusheng Huang, Ruochen Yang, Shuang Wen, Pengbo Xu, Jiangxia Cao, Yueyang Liu, Kuo Cai, Chengcheng Guo, Shiyao Wang, Xinchun Luo, Qiang Luo, Ruiming Tang, Shuang Yang, Zhaojie Liu, Guorui Zhou, Han Li, and Kun Gai. Onelive: Dynamically unified generative framework for live-streaming recommendation. *arXiv preprint arXiv:2602.08612*, 2026. URL <https://arxiv.org/abs/2602.08612>.
- [26] Xiaodong Wang, Langling Huang, Zhirong Wu, Xu Zhao, Teng Xu, Xuhong Xia, and Peixi Peng. LiViBench: An omnimodal benchmark for interactive livestream video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26517–26525, 2026. doi: 10.1609/aaai.v40i31.39859.
- [27] Yongxuan Wu, Runyu Chen, Peiyu Liu, and Hongjin Qian. LiveLongBench: Tackling long-context understanding for spoken texts from live streams. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29713–29732, 2026. doi: 10.18653/v1/2026.findings-acl.1485.
- [28] Tian Xia, Jiaqi Zhang, Yueyang Liu, Hongjian Dou, Tingya Yin, Jiangxia Cao, Xulei Liang, Tianlu Xie, Lihao Liu, Xiang Chen, Shen Wang, Changxin Lao, Haixiang Gan, Jinkai Yu, Keting Cen, Lu Hao, Xu Zhang, Qiqiang Zhong, Zhongbo Sun, Yiyu Wang, Shuang Yang, Mingxin Wen, Xiangyu Wu, Shaoguo Liu, Tingting Gao, Zhaojie Liu, Han Li, and Kun Gai. Qarm v2: Quantitative alignment multi-modal recommendation for reasoning user sequence modeling, 2026. URL <https://arxiv.org/abs/2602.08559>.
- [29] Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, Ning Ding, Shengda Fan, Yewei Fang, Zixuan Fu, Wenyu Guan, Yitong Guan, Junshao Guo, Yufeng Han, Bingxiang He, Yuxiang Huang, Cunliang Kong, Qiuzuo Li, Siyuan Li, Wenhao Li, Yanghao Li, Yishan Li, Zhen Li, Dan Liu, Biyuan Lin, Yankai Lin, Xiang Long, Quanyu Lu, Yaxi Lu, Peiyan Luo, Hongya Lyu, Litu Ou, Yinxu Pan, Zekai Qu, Qundong Shi, Zijun Song, Jiayuan Su, Zhou Su, Ao Sun, Xianghui Sun, Peijun Tang, Fangzheng Wang, Feng Wang, Shuo Wang, Yudong Wang, Yesai Wu, Zhenyu Xiao, Jie Xie, Zihao Xie, Yukun Yan, Jiarui Yuan, Kaihuo Zhang, Lei Zhang, Linyue Zhang, Xueren Zhang, Yudi Zhang, Hengyu Zhao, Weilin Zhao, Weilin Zhao, Yuanqian Zhao, Zhi Zheng, Ge Zhou, Jie Zhou, Wei Zhou, Zihan Zhou, Zixuan Zhou, Zhiyuan Liu, Guoyang Zeng, Chao Jia, Dahai Li, and Maosong Sun. Minicpm4: Ultra-efficient llms on end devices. *CoRR*, abs/2506.07900, 2025.
- [30] Ruochen Yang and Yueyang. Sarm: Llm-augmented semantic anchor for end-to-end live-streaming ranking, 2026. URL <https://arxiv.org/abs/2602.09401>.
- [31] Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. LiveStar: Live streaming assistant for real-world online video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.
- [32] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022.
- [33] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1):5:1–5:38, 2019.
- [34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *CoRR*, abs/2306.05685, 2023.

- [35] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. Agent-as-a-judge: Evaluate agents with agents. [arXiv preprint arXiv:2410.10934](#), 2024.