Intent-Driven Semantic ID Generation for Grounded Conversational News Recommendation

Hongyang Su[2], Beibei Kong[2], Lei Cheng[2], Chengxiang Zhuo[2], Zang Li[2], Chenyun Yu[1]*

[1]Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2]Platform and Content Group, Tencent, Shenzhen, China
{supersusu,echokong,raycheng,felixzhuo,gavinzli}@tencent.com
yuchy35@mail.sysu.edu.cn

Abstract

Conversational news recommendation requires grounding each suggestion in a rapidly evolving article corpus while addressing implicit user intents that lack explicit retrievable keywords. To characterize this scenario, we identify 6 intent types from production dialogues: five are implicit and pose fundamental challenges to standard RAG pipelines, forming a critical *retrieve-first* bottleneck. To address these issues, we introduce **intent-driven Semantic ID (SID) generation** under a **Generate-then-Match** paradigm. With two-stage training that consists of multi-task SID alignment and GPT-4 Chain-of-Thought distillation, an LLM maps diverse intents to hierarchical SID prefixes, which are then fuzzy-matched to the current news pool to guarantee fully grounded recommendations. **Profile-Aware Dual-Signal Reasoning (PADR)** further enables cold-start users to obtain valid recommendations using only profiles. On a mainstream Chinese news platform, our 7B model achieves **0% hallucination** and **12.4% L1 match** in the 152K open-generation SID space (4× random baseline). It matches GPT-4+Hybrid RAG on L1 while surpassing it on finer-grained metrics (L2 2×, Category +1.2pp) at ∼100× lower cost. Cold-start users, where existing baselines score 0%, achieve **18.0% L1** (6× random), the highest among all user groups.

1 Introduction

On a mainstream Chinese news platform, we develop a **chat-based news assistant** that provides personalized recommendations via natural-language dialogue. Unlike stable-catalog recommendation domains (e.g., movies, products), this scenario poses a critical challenge: **most articles expire within 24 hours** (observed from the fraction of daily impressions from articles published in the past 24h), leading to pervasive cold-start issues for both items and users. Moreover, even returning users become "cold" when encountering fresh content, with 20–30% of users having fewer than 10 interactions. When users request "recommend news" or "something different", no retrievable keywords are available, yet such *implicit intents* dominate real-world conversations.

We analyze real-world dialogues on our news platform and identify **6 conversational intent types**. Only explicit keyword queries can be effectively handled by standard RAG, while the remaining 5 categories (i.e., implicit needs, continuation, diversity, feedback, and cold-start) demand sophisticated query construction from user profiles and histories, greatly increasing pipeline complexity. Even embedding-based retrieval (Lewis et al., 2020; Gao et al., 2023b) yields limited gains on these intents, as the query lacks discriminative intent-specific content when users say "recommend something" or "something different". While generative SID recommendation (Rajput et al., 2023; Zhou et al., 2025) encodes items as learnable hierarchical tokens, it targets *single-turn* next-item prediction and relies on rich behavioral sequences. Extending SIDs to conversational recommendation requires solving two key open problems: (1) generating SID prefixes from diverse dialogue intents (not just click sequences), and (2) reasoning for cold-start users who lack the behavioral history relied upon by SID-based models.

Our Approach: Intent-Driven SID Generation. We propose **Generate-then-Match**, a novel paradigm that departs from conventional retrieve-first pipelines. Rather than constructing explicit queries for candidate retrieval, our LLM directly generates SIDs (hierarchical tokens encoding news semantics via RQ-VAE (Rajput et al., 2023)) from user intent, profiles, and dialogue context. The generated SID prefixes are then fuzzy-matched to the current news pool, ensuring each recommendation corresponds to a real article *by construc-*

*Corresponding author.

tion. To handle the full spectrum of user states, we introduce **Profile-Aware Dual-Signal Reasoning (PADR)**, which adaptively routes through warm, hybrid, or cold reasoning paths, enabling even zero-history users to receive meaningful recommendations from their profiles alone. Our contributions are summarized as follows:

- **Intent-Driven Generate-then-Match**: We propose a two-stage training paradigm that conducts multi-task SID alignment and then applies GPT-4 Chain-of-Thought distillation. It encourages an LLM to generate grounded SID prefixes over all 6 intent types and fully discards conventional retrieve-first pipelines.
- **Profile-Aware Dual-Signal Reasoning (PADR)**: We design an adaptive reasoning framework that dynamically selects warm, hybrid, or cold paths based on accessible user signals. It effectively supports cold-start users where existing baselines often fail and yields competitive recommendation performance.
- **Industrial-Scale Validation**: Evaluations on a 163K-article test pool show 12.4% L1 score, outperforming the production baseline of 11.6% with $p < 0.05$. Our solution achieves 0% hallucination and sub-100ms end-to-end latency.

2 Related Work

Conversational Recommendation Systems. CRS elicit user preferences through multi-turn dialogue (Jannach et al., 2021; Gao et al., 2021). From ReDial (Li et al., 2018) to knowledge-enhanced (Chen et al., 2019; Ma et al., 2021), unified (Wang et al., 2022; Zhou et al., 2022), and LLM-based methods (Friedman et al., 2023; He et al., 2023; Gao et al., 2023a; Wang et al., 2023), most work assumes *stable catalogs* with explicit preference mentions. Our setting differs fundamentally: *ephemeral* news (most articles cycle out within 24h), 5/6 implicit intents lacking retrievable keywords, and cold-start users (see Appendix U for a detailed comparison).

Generative Recommendation and Semantic IDs. Semantic IDs represent items as learnable hierarchical tokens (Rajput et al., 2023; Tay et al., 2022; Zheng et al., 2024). TIGER (Rajput et al., 2023) pioneered this approach but assumes stable catalogs, limiting its effectiveness in rapidly refreshing domains like news. The OneRec series (Zhou et al., 2025) demonstrates industrial-scale SID deployment, while OneRec-Think (Liu et al., 2025b) and ThinkRec (Yu et al., 2025) extend SID-based recommendation with explicit reasoning. DiscRec (Liu et al., 2025a) disentangles semantic and collaborative signals. For a broader overview, see (Li et al., 2024). These methods target *single-turn* next-item prediction and rely on behavioral sequences; our work differs in targeting *conversational* recommendation with 6 intent types, enabling cold-start via PADR, and structurally ensuring 0% hallucination via Generate-then-Match. We directly compare with TIGER and OneRec in Table 6.

Grounding and Hallucination in LLM Recommendations. Hallucination remains critical (Zhang et al., 2025; Ji et al., 2023; Huang et al., 2025). RAG (Lewis et al., 2020; Gao et al., 2023b) mitigates it but fails for implicit queries; constrained decoding (Hokamp and Liu, 2017) limits expressiveness; graph-based retrieval (Qiu et al., 2025) still requires successful retrieval. Our Generate-then-Match addresses this structurally: the matching function guarantees every output exists in the current news pool (§3.1).

3 Method

Figure 1 overviews our Generate-then-Match framework. PADR (§3.3) routes user context through warm, hybrid, or cold reasoning paths. The fine-tuned LLM then generates 3-layer SID prefixes encoding the inferred intent, and fuzzy matching (δ=5) grounds them to the current news pool (§3.2). A Dual-Track architecture caches pool-agnostic prefixes for sub-100ms latency (§3.4).

3.1 Generate-then-Match: Replacing Retrieve-First

Traditional RAG follows a *retrieve-then-generate* pipeline: the user query q serves as a retrieval key to fetch candidates. This works when q contains explicit keywords, but 5 of our 6 intent types are *implicit* (Table 1), carrying no retrievable keywords. We reverse this pipeline with a *generate-first* strategy:

An inter-annotator agreement study (Fleiss' κ=0.81, 6-way; Appendix W) confirms the reliability of this taxonomy.

$$\begin{aligned} \text{RAG}: &\quad \mathcal{R}(q) \rightarrow \mathit{LLM}(\mathcal{R}(q), q) \\ \text{Ours}: &\quad \mathit{LLM}(u, h, q) \rightarrow \mathit{Match}(\mathit{SID}, \mathcal{P}) \end{aligned} \tag{1}$$

where u is user profile, h is behavioral history, q is query, *SID* is the generated Semantic ID prefix, and

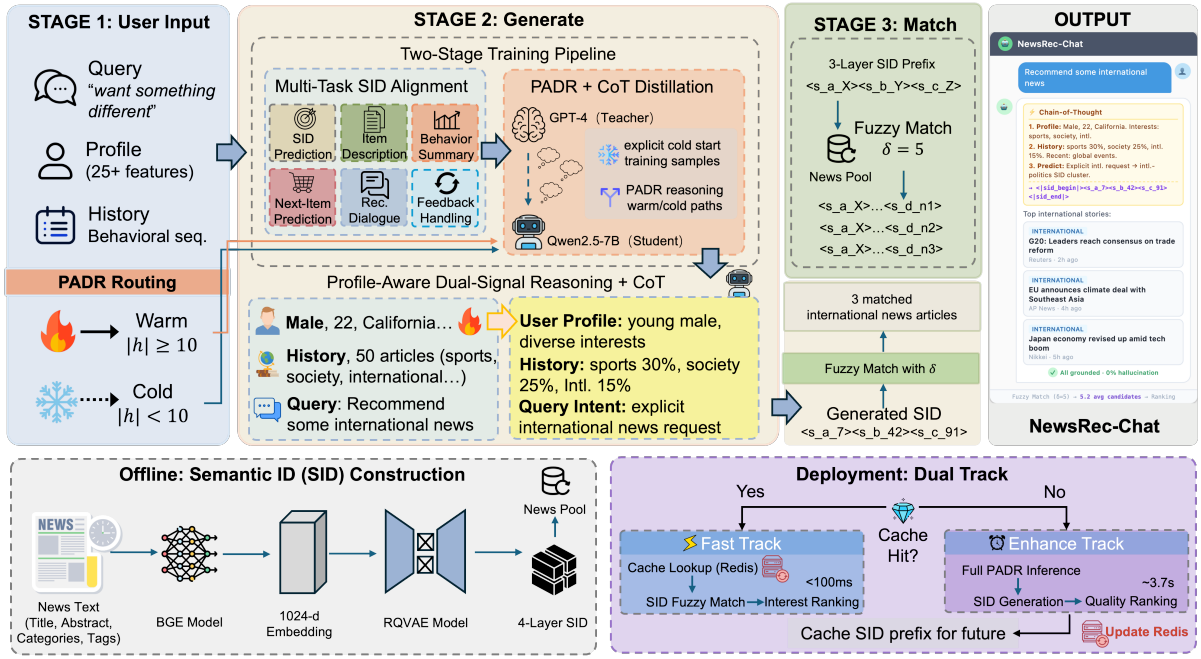

Figure 1: Overview of the Generate-then-Match framework. **Left**: PADR routes user context through warm/cold/hybrid reasoning paths. **Center**: The two-stage trained LLM (Stage 1: SID alignment, Stage 2: CoT distillation) generates 3-layer SID prefixes from intent and context. **Right**: Fuzzy matching grounds prefixes to the current news pool. Bottom: offline SID construction (RQ-VAE) and Dual-Track deployment with SID-prefix caching.

Intent Type	Example Query	Explicit?
Candidate Selection	"recommend tech news"	✓
Next-Item Continuation	"what else?"	×
Diversity Seeking	"something different"	×
Feedback Adjustment	"not sports"	×
Cold-Start (PADR)	"recommend news" (sparse history)	×
Pure Cold-Start	"recommend news" (new user)	×

Table 1: Six conversational intents mapped to evaluation tasks. "Explicit?" indicates whether the query contains retrievable keywords. 5/6 intents are implicit, posing a structural challenge for retrieval-first pipelines.

$\mathcal{P}$ is the current news pool. Rather than retrieving first, the LLM *internalizes* user context (u, h, q) and directly generates a structured SID that encodes the inferred intent. The matching function then provides a *pool-existence guarantee*:

$$Match(SID, \mathcal{P}) \subseteq \mathcal{P} \tag{2}$$

That is, every recommendation is restricted to $\mathcal{P}$ by construction, eliminating hallucination at the architectural level. This paradigm shift introduces two new challenges: (1) the model must learn to generate *semantically meaningful* SID prefixes from diverse intents, addressed by two-stage training and PADR (§3.3); (2) generated prefixes must be grounded to a *dynamically refreshed* news pool, addressed by fuzzy matching (§3.2).

3.2 Semantic ID Structure and Fuzzy Matching

4-Layer Hierarchical SID. We adopt the production SID system already deployed on the platform, where each article is encoded as SID $= \langle s_1, s_2, s_3, s_4 \rangle$ (Rajput et al., 2023; Tay et al., 2022). SIDs are generated by RQ-VAE over news content embeddings: each layer is a purely *semantic* quantization code capturing progressively finer content similarity, ranging from broad semantic regions (s_1) through mid-level groups (s_2) and fine-grained clusters (s_3) to near-unique article identifiers (s_4). This hierarchy enables generation and matching at different granularities (codebook details in Appendix A).

3-Layer Prefix Generation. The model generates **3-layer prefixes** $P = (s_1, s_2, s_3)$ rather than full 4-layer SIDs. The first three layers capture *intent-level* semantics: articles sharing (s_1, s_2, s_3) are topically interchangeable. The fourth layer s_4 serves as a near-unique article discriminator that shifts daily as articles enter and leave the pool; predicting it would couple the model to a specific day's inventory. Generating up to s_3 thus provides the right granularity: specific enough to reflect user intent, yet stable across pool refreshes.

Fuzzy Matching Algorithm. Generated prefixes are matched against the pool with tolerance δ:

$$Match(P, \mathcal{P}) = \{n \in \mathcal{P} : s'_1 = s_1, s'_2 = s_2, |s'_3 - s_3| \leq \delta\} \quad (3)$$

where (s'_1, s'_2, s'_3, s'_4) is the SID of news n. Layers 1–2 are matched strictly to preserve semantic coherence, as adjacent s_1 or s_2 codes may encode unrelated content regions. Layer s_3 admits tolerance δ because neighboring codes within the same (s_1, s_2) group share fine-grained topical similarity. We set δ=5 via grid search over $\{1, 3, 5, 7, 10\}$ on a held-out set (Appendix E). Matched candidates are ranked by $1 - |s'_3 - s_3|/(\delta + 1)$, yielding a small candidate set (mean 5.2, median 3.0 articles; Appendix M).

3.3 Profile-Aware Dual-Signal Reasoning (PADR)

Cold-start users (20–30% of active users) lack behavioral history. PADR adaptively leverages two complementary signal types based on user state.

Dual-Signal Context Construction. Let u denote a user with profile $\mathbf{p}_u$ and behavioral history $\mathbf{h}_u = [n_1, \ldots, n_t]$ (clicked news). The profile $\mathbf{p}_u$ aggregates 25+ features covering demographics, preferences, and behavioral patterns (Appendix T). The LLM input context adaptively includes or omits behavioral history:

$$\mathbf{x}_u = \begin{cases} [\mathbf{p}_u; \mathbf{h}_u; q] & |\mathbf{h}_u| \geq \tau \text{ (warm)} \\ [\mathbf{p}_u; \mathbf{h}_u; q; \text{“sparse”}] & 0 < |\mathbf{h}_u| < \tau \text{ (hybrid)} \\ [\mathbf{p}_u; q; \text{“no history”}] & |\mathbf{h}_u| = 0 \text{ (cold)} \end{cases} \quad (4)$$

where τ=10 is the history sufficiency threshold (Appendix N). An explicit availability indicator guides the model toward the appropriate reasoning strategy.

Learned Reasoning Paths. Unlike rule-based cold-start strategies that hard-code fallback logic, PADR reasoning paths are *learned* through CoT supervision: the model autonomously develops differentiated strategies for each user state. The *Warm Path* ($|\mathbf{h}_u| \geq \tau$) correlates behavioral patterns with profile features; the *Cold Path* ($|\mathbf{h}_u| = 0$) infers interests from demographics alone (achieving 18.0% L1 accuracy); the *Hybrid Path* cross-references sparse clicks with profiles. No explicit routing module is needed, as the availability indicator in Eq. 4 suffices for path selection.

Two-Stage Training Pipeline. *Stage 1: Multi-Task SID Alignment* builds a grounded SID vocabulary via 6 objectives in 3 groups: **SID Understanding** (bidirectional content↔SID mapping), **User Modeling** (behavior summarization and next-item prediction), and **Recommendation Dialogue** (multi-turn recommendation integrating all capabilities; task ratios in Appendix B).

Stage 2: PADR CoT Distillation then teaches the model to *reason* about which SIDs to generate. GPT-4 produces gold CoT traces for each (input, target SID) pair (Li et al., 2023; Magister et al., 2023), which are distilled into the smaller LLM. Three design choices are critical: (1) **∼31% cold-start samples** to ensure profile-only reasoning; (2) **intent-differentiated CoT**, where each intent type receives a distinct reasoning structure (e.g., demographic→interest for cold-start, preference-shift analysis for feedback); and (3) **controlled CoT length** (150–300 chars) to prevent reasoning degradation at inference (§4.5). Human evaluation of 200 sampled CoTs confirms high quality (mean 4.1/5; Appendix Q).

3.4 Dual-Track Architecture for Industrial Deployment

Full PADR inference is too slow for real-time conversation. We design a dual-track architecture to decouple latency from reasoning quality. The **Fast Track** looks up cached SID prefixes, applies fuzzy matching against the current pool, and ranks candidates; on cache miss, a profile-based fallback provides immediate results. The **Enhance Track** runs asynchronously, performing full PADR inference and updating the cache.

Two design choices are critical. First, cached SID prefixes are *pool-agnostic*, encoding user intent rather than specific article identifiers, so they remain valid across daily pool refreshes. Second, the Enhance Track proactively generates SID prefixes using *preference-based preset queries* derived from each user's profile, ensuring recommendation timeliness even before the user initiates a conversation (Appendix C).

4 Experiments

4.1 Setup

Dataset. We use data from a major Chinese news platform (Table 2). A *temporal split* ensures zero item overlap between training and test: training uses news before a cutoff date; test uses only news

Data	Description	Scale
News Pool	Articles with SID	163,560
User Profiles	Rich interest profiles	50,000
Stage 1 Training	SID alignment	483,374
Stage 2 Training	PADR reasoning	48,249
Test Samples	6 task types	9,982

Table 2: Dataset statistics.

published after. Our **primary evaluation** is open generation (76% of test data), where the model generates SID prefixes in the full 152K space. Candidate selection (24%) uses fixed candidates per sample to isolate discriminative ability (Appendix I).

Baselines. We compare against five categories of methods. *Non-Personalized*: Random, Popular, and Hist-Pop (Appendix D). *LLM-Based*: GPT-4 Direct and Qwen-7B Direct (no retrieval); GPT-4+RAG and Qwen-7B+RAG with BM25 (Robertson and Zaragoza, 2009), BGE-large-zh Dense (Xiao et al., 2024), and Hybrid retrieval (Appendix G). *Generative Recommendation*: TIGER (Rajput et al., 2023) and OneRec-7B (Zhou et al., 2025) (same Qwen2.5-7B backbone with LoRA, constrained SID decoding; Table 6). *Sequential Recommenders*: SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019) (Appendix F). Our model is fine-tuned from **Qwen2.5-7B-Instruct** (Team, 2024) on 4×H20-96G GPUs (Appendix A).

Metrics. *Candidate Selection*: **Hit@1** under **Rand** (4 random negatives) and **Align** (2 same-category + 2 random) settings. *Open Generation* (152K SID space): **L1 Match** (32 categories), **L2 Match** (2,048 topics), **Category Match** (Cat., editorial taxonomy alignment). *Grounding*: **Hallucination Rate** (Hal., fraction of generated SIDs absent from the current pool). *System*: Average and P95 latency.

4.2 Main Results

Table 3 presents results on the 9,982-sample test set. Generative recommendation baselines (TIGER, OneRec-7B) and sequential recommenders (SASRec, BERT4Rec) are compared in Table 6 and Appendix F, respectively.

Key Findings. (1) *SID training eliminates hallucination and enables meaningful generation.* Without SID training, both Qwen-7B and GPT-4 hallucinate 70–95% of outputs. Our two-stage fine-tuning reduces hallucination to **0%** while achieving 12.4% L1 (95% CI: [11.7%, 13.2%]). This represents a 4× improvement over random in a 152K open-generation space, where the model must predict the exact L1 category among 32 classes without any retrieval, and surpasses the production-grade Hist-Pop by +0.8pp ($p<0.05$). Hist-Pop is further unavailable for cold-start users (L1=0%), where our method reaches **18.0%** (§4.5). (2) *Generate-then-Match outperforms both retrieval and generative baselines.* Our method surpasses GPT-4+Hybrid RAG on fine-grained metrics (L2 2×, Category +1.2pp) at ∼100× lower cost (Appendix H). Among generative methods, TIGER (Rajput et al., 2023) and OneRec-7B (Zhou et al., 2025) are limited to single-turn next-item prediction (1/6 intents) and fail entirely for cold-start users; our model outperforms OneRec-7B on cold-start L1 (18.0% vs. 16.1%) while covering all 6 intents (Table 6). Across all retrieval strategies, retrieve-first methods plateau on implicit intents (76% of our test set; Appendix G). (3) *Dual-setting candidate selection confirms robustness.* Under Rand, Ours (59.3%) dominates all baselines ($p<0.01$). Under the harder Align setting, our 7B model (30.8%) matches GPT-4 Direct ($p=0.94$) while outperforming Qwen-7B Direct (+4.8pp, $p<0.01$). Since candidate selec-

Method		Cand. Hit@1		Open Gen.			Hal.
		Rand	Align	L1	L2	Cat.	
Non-Personalized							
Random		20.0	20.0	5.1	0.1	10.3	–
Popular		20.0	20.0	7.7	0.5	12.6	–
Hist-Pop		–	–	11.6	0.7	16.8	–
End-to-End LLM (no SID training)							
Qwen-7B Direct		28.1	26.0	2.4	0.0	6.9	70.0%
GPT-4 Direct		34.4	30.9	0.9	0.0	1.4	94.6%
Retrieve-then-Rank							
BM25		23.5	23.4	10.5	0.5	15.7	0%
Qwen-7B	BM25	28.1	26.0	10.1	0.5	15.2	0%
	Dense			11.8	0.8	17.4	0%
	Hybrid			11.4	0.5	18.1	0%
GPT-4	BM25	34.4	30.9	11.0	0.5	15.6	0%
	Dense			11.8	0.3	18.3	0%
	Hybrid			12.4	0.5	18.8	0%
Generate-then-Match (Ours)							
NewsRec-Chat		**59.3**	**30.8**	**12.4**	**1.0**	**20.0**	**0%**

Table 3: Main results (%) on 9,982 test samples. **Rand**: 4 random negatives; **Align**: 2 same-category + 2 random (N=2,349). Generative recommendation baselines (TIGER, OneRec) are compared in Table 6. **Best** and second-best are highlighted.[1]

[1]Fixed candidates per sample; same-backbone models share identical Hit@1 under deterministic decoding.

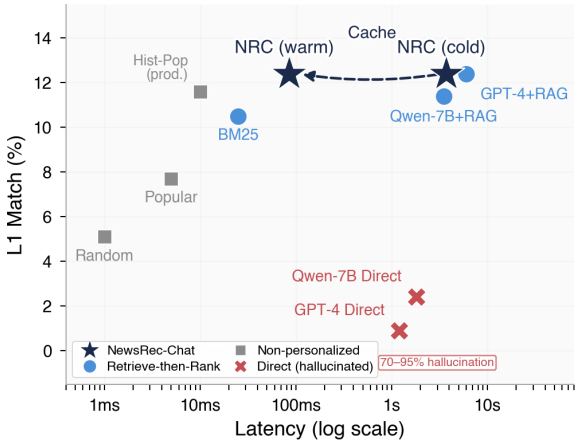

Figure 2: Latency-quality trade-off.

tion covers 24% of data, we regard open generation as the primary evaluation (Appendix I).

4.3 System Performance and Pilot Deployment

Latency and Reliability. Figure 2 shows the latency-quality trade-off. Our dual-track architecture achieves 85ms average warm-start latency (P95: 150ms) with 0% hallucination. Over 75% of requests hit the cache, while cold-start users experience a 3.7s initial delay that is subsequently cached. The primary SID pathway (δ=5) returns non-empty candidates for >95% of requests, with Level 2+ fallbacks in fewer than 5% of cases (Appendix J), confirming that 0% hallucination stems from accurate generation rather than fallback mechanisms.

Internal Pilot Study. We deployed the system to **300+ internal users** on a major Chinese news platform for 38 days; **92 active users** generated 331 turns across 141 sessions without scripted tasks (58.9% multi-turn, 22.8% return rate). The system recommended 731 unique articles across 29 categories with **zero hallucination complaints**, consistent with the 0% offline rate. Content editors and product managers further conducted qualitative evaluations covering all 6 intent types, confirming cold-start relevance and multi-turn refinement benefits (Appendix J.1).

4.4 Ablation Study

Table 4 isolates each component's contribution; all differences are significant at p<0.01 except w/o Fuzzy on Hit@1 (p=1.0). **Stage 2 PADR** is the largest contributor: removing it drops Hit@1 and

Variant	Hit@1	L1	Hal.	Lat.
Full Model	**59.3%**	**12.4%**	**0%**	**85ms**
w/o Stage 2 (S1 only)	51.6%	8.9%	18.4%	0.67s
w/o Fuzzy Match	59.3%	12.4%	5.7%	85ms
w/o Dual-Track	59.3%	12.4%	0%	3.7s

Table 4: Ablation study (Rand setting).

Task Type	N	Metric	Qwen-7B	Ours
Conversational (LLM-exclusive):				
Cand. Sel. (Rand)	2349	Hit@1	28.1%	**59.3%**
Cand. Sel. (Align)	2349	Hit@1	26.0%	**30.8%**
Feedback Adj.	621	Hit@1	19.2%	**54.8%**
Open Generation (152K SID space):				
Next-Item Pred.	5319	L1	3.4%	**12.0%**
Cold-Start PADR	611	L1	0.2%	**11.1%**
Diversity Expl.	978	L1	0.1%	**11.6%**
Pure Cold-Start	721	L1	0.3%	**18.0%**
Cold-Start Analysis (Open Gen. L1):				
Warm Tasks	6300	L1	2.9%	**11.9%**
Cold Tasks	1333	L1	0.2%	**14.9%**
Pure Cold (0)	721	L1	0.3%	**18.0%**

Table 5: Task-wise and cold-start breakdown. Same Qwen-7B backbone, with (Ours) vs. without SID training. Bootstrap CIs in Appendix O.

surges hallucination to 18.4%, confirming CoT distillation is essential. **Fuzzy matching** recovers ~6% of cases where exact prefix matching fails (Hal. 0%→5.7%). **Dual-Track** is essential for production latency. Further analysis of the grounding layer (Appendix K) and 200 failure cases (Appendix S) confirms that generation grounding alone eliminates hallucination.

4.5 Task-wise and Cold-Start Analysis

Table 5 breaks down performance by task and cold-start status, comparing the same Qwen-7B backbone with and without SID training. SID training yields gains across all intents, with the largest improvements on implicit ones: feedback adjustment (+35.6pp Hit@1) and pure cold-start (60× L1). On candidate selection, our method leads by +31.2pp (Rand) and +4.8pp (Align), consistent with the main-table pattern. Notably, **cold-start tasks achieve higher L1 than warm tasks** (14.9% vs. 11.9%, p<0.05), with Pure Cold-Start reaching 18.0%. This is because PADR's profile-to-cluster mapping narrows the SID space more effectively than warm-path reasoning over diverse behavioral histories (Appendix L); a data leakage audit confirms this is genuine (Appendix T).

Method	Params	Cold L1	Intents
SASRec	1M	0%	1/6
TIGER	100M	0%	1/6
OneRec-7B	7B	16.1%	1/6
NewsRec-Chat	**7B**	**18.0%**	**6/6**

Table 6: Cold-start comparison. Cold L1: pure cold-start (zero history) L1 match. OneRec-7B uses the same Qwen2.5-7B backbone with constrained SID decoding.

Cross-Category Generalization. To assess whether our model generalizes beyond high-frequency training categories, we evaluate L1 accuracy across all 29 editorial categories present in the test set. Performance remains robust under category imbalance: mean L1 across all 29 categories is 23.5% (CV=0.23), with all categories exceeding the random baseline ($\approx$3.1%). Notably, **9 zero-shot categories** (present in the test set but absent or rare in Stage 2 training) achieve a mean L1 of 19.7% ($4\times$ random), and the Spearman correlation between training frequency and test L1 is insignificant (ρ=0.35, p=0.055). This confirms that the model captures *semantic* category structure via SID alignment rather than memorizing frequency patterns.

Cold-Start Comparison with Generative Baselines. Table 6 directly compares cold-start performance against generative baselines. Sequential methods (SASRec, TIGER) require behavioral sequences and thus score 0% on cold-start users. OneRec-7B (same Qwen2.5-7B backbone with LoRA) achieves 16.1% cold-start L1 using constrained SID decoding, while our model reaches 18.0% by leveraging PADR's profile-based reasoning. Crucially, our model is the only one covering all 6 conversational intents.

Transferability. While evaluated on a single news platform, our framework is designed for cross-domain transferability. The core architecture decouples platform-specific components (SID codebook, profile schema) from the general mechanisms (intent routing, PADR reasoning, fuzzy matching). Adapting to a new domain requires retraining the RQ-VAE codebook ($\sim$2h on 1 GPU) and remapping the profile schema, leaving the two-stage training pipeline and Generate-then-Match paradigm unchanged. The short-lifecycle and implicit-intent challenges we address are prevalent in other domains (short videos, live streaming, flash-sale e-commerce), all requiring real-time cold-start handling and expiry-aware recommendation. The cross-category generalization above, combined with strong zero-shot category performance, provides in-domain evidence of semantic transfer; we leave cross-domain evaluation to future work.

5 Conclusion

We present NewsRec-Chat, introducing **intent-driven SID generation** for conversational news recommendation. By mapping 6 intent types to hierarchical SID prefixes with fuzzy matching, the system eliminates hallucinations by construction while handling implicit intents that challenge retrieve-first approaches. Two-stage training reduces a 7B model's hallucination from 70% to 0% and surpasses GPT-4+RAG at $\sim$100$\times$ lower cost. PADR enables cold-start users to achieve the highest accuracy, and dual-track caching delivers sub-100ms latency. An internal pilot confirms real-world viability. The generate-then-match paradigm is broadly applicable to other short-lifecycle recommendation domains.

Acknowledgments

We thank the internal users who participated in the NewsRec-Chat pilot study for their valuable usage feedback, and the seven expert annotators who contributed to the intent taxonomy IAA validation. We also thank the content editors and product managers for their qualitative evaluations. Finally, we thank the anonymous reviewers for their constructive suggestions, which significantly improved this paper.

References

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, and 1 others. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten De Rijke, and Tat-Seng Chua. 2021. Advances and

challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023a. Chat-REC: Towards interactive and explainable LLMs-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 720–730.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys*, 54(5):1–36.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 197–206.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in Neural Information Processing Systems*, 31.

Yang Li, Yidan Zhang, Fuyuan Sun, Yizhe Shan, Yuqiang Ge, Yue Gao, Xu He, Peng Zhao, Hongyan Bao, and Kang Gai. 2024. A survey on generative recommendation. *arXiv preprint arXiv:2404.00011*.

Chang Liu, Yimeng Bai, Xiaoyan Zhao, Yang Zhang, Fuli Feng, and Wenge Rong. 2025a. DiscRec: Disentangled semantic-collaborative modeling for generative recommendation. *arXiv preprint arXiv:2506.15576*.

Zhanyu Liu, Shiyao Wang, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng, Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu, and 1 others. 2025b. OneRec-Think: In-text reasoning for generative recommendation. *arXiv preprint arXiv:2510.11639*.

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-Walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781.

Zhangchi Qiu, Linhao Luo, Zicheng Zhao, Shirui Pan, and Alan Wee-Chung Liew. 2025. Graph retrieval-augmented LLM for conversational recommendation systems. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 344–355.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, and 1 others. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10052–10065.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

Qihang Yu, Kairui Fu, Shengyu Zhang, Zheqi Lv, Fan Wu, and Fei Wu. 2025. ThinkRec: Thinking-based recommendation via LLM. *arXiv preprint arXiv:2505.15091*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1435–1448.

Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, and 1 others. 2025. OneRec technical report. *arXiv preprint arXiv:2506.13695*.

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-CRS: Coarse-to-fine contrastive learning for conversational recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1488–1496.

Parameter	Stage 1	Stage 2
Training Method	Full-parameter	Full-parameter
Epochs	3	3
Learning Rate	2e-5	1e-5
Batch Size (per GPU)	1	1
Gradient Accumulation	16	16
Warmup Ratio	0.1	0.1
Max Sequence Length	1536	2048
Weight Decay	0.01	0.01
LR Scheduler	Cosine	Cosine
Distributed Training	DeepSpeed ZeRO-2 (4×H20-96G)	

Table 7: Training hyperparameters for two-stage framework. Effective batch size = $1 \times 16 \times 4 = 64$.

Algorithm 1: SID-Prefix Fuzzy Matching

Input: SID prefix $P = (s_1, s_2, s_3)$, news pool $\mathcal{N}$, tolerance δ, top-k
Output: Ranked news list $\mathcal{R}$

```
C ← ∅                                   // Candidates
foreach (n, sid_n) ∈ N do
    (s'_1, s'_2, s'_3, s'_4) ← parse(sid_n)
    if s'_1 = s_1 and s'_2 = s_2 and |s'_3 − s_3| ≤ δ then
        score ← 1 − |s'_3 − s_3| / (δ+1)
        C ← C ∪ {(n, score)}
    end
end
R ← TopK(C, k) by score
return R
```

A Training Configuration and Algorithm

A.1 Hyperparameters

Table 7 summarizes the training hyperparameters for both stages.

Token Warm-up Phase (before Stage 1): We first train only the embedding layer for 3 epochs with learning rate 5e-4, allowing the 1,250 new SID tokens to develop meaningful representations.

SID Codebook Details. The 4-layer codebook is constructed via RQ-VAE on BGE-large-zh (Xiao et al., 2024) embeddings (1,024-dim): $s_1 \in [0, 31]$, $s_2 \in [0, 63]$, $s_3 \in [0, 127]$, $s_4 \in [0, 1023]$. Codebook occupancy exceeds 98%; 152K codes cover 163K articles with 1,250 special tokens added to the LLM vocabulary.

A.2 SID-Prefix Fuzzy Matching Algorithm

Algorithm 1 details the fuzzy matching procedure used to ground generated SID prefixes to the news pool.

Tolerance δ Selection. We select δ via grid search on a held-out validation set (1K samples, no temporal overlap with test). Results: δ=1: 4.8% Hal., mean 1.8 candidates; δ=3: 2.1% Hal., mean 3.4 candidates; δ=5: 0% Hal., mean 5.2 candidates; δ=7: 0% Hal., mean 7.8 candidates; δ=10: 0% Hal., mean 12.1 candidates. We choose δ=5 as the smallest value achieving 0% hallucination with a manageable candidate set size.

B Training Task Details

B.1 Stage 1: Multi-Task SID Alignment (6 Tasks, 483K samples)

SID Understanding (36.2%):

1. *SID Prediction (20.7%)*: Given news title, category, and tags, predict its 4-layer SID. This teaches the model to map semantic content to hierarchical identifiers.
2. *Item Description (15.5%)*: Given SID, generate corresponding news description. This teaches bidirectional SID-content mapping.

User Modeling (10.4%):

3. *Behavior Summary (5.2%)*: Summarize user's reading history into interest profile.
4. *Next Item Prediction (5.2%)*: Predict next click given historical sequence.

Recommendation Dialogue (53.5%):

5. *Recommendation Dialogue (48.3%)*: Generate personalized recommendations with SIDs within dialogue context. This is the core task that combines all capabilities.
6. *Feedback Handling (5.2%)*: Respond appropriately to user preference feedback (e.g., "I don't like sports") and adjust recommendations.

B.2 Stage 2: PADR Cold-Start Reasoning (6 Tasks, 48K samples)

We employ GPT-4 to generate Chain-of-Thought (CoT) reasoning traces, then distill these into our 7B model. The key design is **adaptive dual-signal context construction**: each sample incorporates both profile signals and behavioral signals (when available), with **explicit cold-start handling** through dedicated training samples.

Core Design: Dual-Signal Utilization. Each CoT sample incorporates both signal types:

- **Cold-Start Signal (User Profile):**
  - Demographics (location, age range, gender)
  - Long-term interests (category preferences over time)
  - Content preferences (video vs. text affinity)
  - Activity patterns (active hours, engagement level)
- **Warm-Start Signal (Behavioral History):**
  - Click sequences (recent reading history)

- Dwell times (engagement depth signals)
- Explicit feedback (likes, shares, comments)

This dual-signal design enables the model to reason effectively for both cold-start users (profile-only) and warm users (rich behavioral data).

Reasoning Paths. PADR supports three reasoning paths based on user state:

Warm-Start Path ($|\mathbf{h}_u| \geq \tau$): Profile Analysis $\rightarrow$ Interest Mapping $\rightarrow$ Behavioral Correlation $\rightarrow$ SID Generation

Hybrid Path ($0 < |\mathbf{h}_u| < \tau$): Profile Analysis $\rightarrow$ Interest Inference $\rightarrow$ Sparse Behavioral Cross-Reference $\rightarrow$ SID Generation

Cold-Start Path ($|\mathbf{h}_u| = 0$): Profile Analysis $\rightarrow$ Interest Inference $\rightarrow$ Content Matching $\rightarrow$ SID Generation

Task 1: Next Item Prediction (31.6% = 15,249):

- Input: User profile + browsing history + target news (user's actual next click)
- Output: Structured reasoning explaining why user clicked this news $\rightarrow$ SID generation
- This is the core warm-start task using real user interactions as positive samples

Task 2: Cold-Start PADR (18.7% = 9,000):

- Input: User profile + minimal history (<5 interactions) + target news
- Output: Profile-focused reasoning with limited behavioral signals $\rightarrow$ SID generation
- Reasoning emphasis: Profile-based interest inference

Task 3: Pure Cold-Start (12.4% = 6,000):

- Input: User profile only (zero behavioral history) + user's first click
- Output: Pure profile-based reasoning $\rightarrow$ SID generation
- Key innovation: Trains model to reason solely from demographics and declared interests

Task 4: Candidate Selection (18.7% = 9,000):

- Input: User profile + history + 5 candidate news items (1 positive, 4 negatives)
- Output: Comparative reasoning analyzing each candidate's fit $\rightarrow$ Select best match
- Trains discriminative reasoning ability

Task 5: Diversity Exploration (12.4% = 6,000):

- Input: User profile + history dominated by single category + diverse target
- Output: Reasoning that justifies recommending outside user's comfort zone
- Addresses filter bubble concerns

Task 6: Feedback Adjustment (6.2% = 3,000):

- Input: Previous recommendation + user negative feedback + alternatives
- Output: Reasoning that analyzes feedback, updates interest mapping, and adjusts strategy

C Dual-Track Architecture Details

C.1 Fast Track Pipeline

The Fast Track guarantees <100ms response through the following steps:

1. **Cache Lookup** ($\sim$5ms): Query Redis for cached SID prefixes matching user context hash.
2. **SID Matching** ($\sim$50ms): If cache hit, match prefixes against current news pool via Algorithm 1.
3. **Interest Ranking** ($\sim$20ms): Rank candidates by user interest alignment score.
4. **Profile Fallback**: If insufficient matches (<3 items), trigger the Level 3+ fallback cascade (§J.1).

C.2 Enhance Track Pipeline

The Enhance Track (3.7s average) generates high-quality recommendations:

1. **Intent Understanding**: Full LLM inference with complete dialogue history.
2. **SID Generation**: GPU-accelerated SID Generator produces Top-K SID sequences.
3. **Quality Ranking**: Re-rank candidates by relevance and diversity metrics.
4. **Cache Update**: Asynchronously write SID prefixes to Redis for future Fast Track hits.

C.3 Interest-Aware Ranking Parameters

For the interest-aware ranking (tuned on a held-out validation set):

- $w_i = 3$ for category-level matches (e.g., "technology", "sports")
- $w_i = 1$ for keyword-level matches (e.g., "AI", "iPhone")
- $\lambda = 0.1$ balances relevance and recency

C.4 Cache Entry Structure

Each cache entry contains:

$$\text{Entry} = \langle \textit{ctx_hash}, \{(s_1, s_2, s_3)\}_{k=1}^{K}, \textit{reason}, \textit{ts} \rangle \tag{5}$$

where *ctx_hash* is computed from user profile + recent dialogue, $K = 10$ SID prefixes are stored, *reason* contains the natural language explanation, and TTL is 24 hours.

D Hist-Pop Baseline Details

Hist-Pop (History Popularity) is a *production-grade informed baseline*, not a trivial lookup table. It selects the most frequent SID from user browsing history, built from 5 data sources: (i) 30-day and 7-day reading-duration-weighted category preferences (Top-3), (ii) behavioral patterns (active time slots, daily reading duration, activity level), and (iii) recent reading history with article metadata, totaling 25+ engineered features per user. Critically, this baseline is *unavailable* for cold-start users (L1=0%), where our method achieves 18.0%.

E Fuzzy Matching: Why Strict L1/L2 with Tolerant L3

Design Rationale. Our fuzzy matching (Eq. 3) applies strict matching on L1 and L2 ($s'_1{=}s_1, s'_2{=}s_2$) with tolerance only on L3 ($|s'_3{-}s_3|{\leq}\delta$). A natural question is: why not allow tolerance on L1/L2 as well, given that L1 match is only 12.4%?

The answer is **semantic coherence preservation**. L1 codes represent coarse semantic clusters (32 categories), and L2 codes represent mid-level topics (64 per L1). Allowing L1 tolerance would mix fundamentally different content domains (e.g., sports ↔ technology), producing semantically incoherent recommendations. L3 codes, by contrast, represent fine-grained sub-clusters within the same L1×L2 topic; tolerance here preserves topical coherence while accommodating the model's prediction uncertainty.

Empirical Validation. We tested hierarchical tolerance on the validation set:

- **Strict L1+L2, tolerant L3** ($\delta{=}5$): 0% Hal., mean 5.2 candidates, 53.6% category overlap (*chosen*).
- **Tolerant L2** ($\delta_2{=}2$) **+ L3** ($\delta_3{=}5$): 0% Hal., mean 18.4 candidates, but category overlap drops to 31.2%, and candidates become semantically diffuse.
- **Tolerant L1** ($\delta_1{=}1$) **+ L2 + L3**: 0% Hal., mean 47.6 candidates, category overlap 14.8%, effectively random.

The strict L1/L2 constraint ensures that when the model predicts the correct coarse category (12.4% of the time), the fuzzy-matched candidates are highly relevant. For the 87.6% where L1 is incorrect, the system returns zero candidates and falls back to the production safeguard cascade (Appendix J), which itself provides reasonable recommendations via profile-based popular articles. This design *fails gracefully* rather than producing incoherent results.

Method	R@1	R@5	R@10	N@1	N@5	N@10	Cat.
Sequential Recommenders							
SASRec	0.68	3.08	5.24	0.68	1.88	2.57	17.3
BERT4Rec	**0.79**	**3.12**	**5.52**	**0.79**	**2.05**	**2.82**	17.0
TIGER	0.41	1.62	2.07	0.41	1.24	1.39	19.8
Ours	-†	-	-	-	-	-	**20.0**

Table 8: Sequential recommendation comparison (%). R@K=Recall@K, N@K=NDCG@K, Cat.=Category Match. †Our method generates SID prefixes (not item rankings), so R@K/N@K are not directly applicable. R@1 and N@1 are identical since both equal the indicator for the top-1 prediction matching the target. These methods are evaluated on the next-item prediction subset only (N=5,322).

F Comparison with Sequential Recommenders

While TIGER and OneRec-7B are compared in Table 6, we provide additional analysis here comparing against dedicated sequential recommenders on their native task (next-item prediction) using standard ranking metrics (R@K, N@K). Since our method generates SID prefixes rather than item rankings, we additionally report Category Match as the cross-paradigm metric.

On their native task (next-item prediction), BERT4Rec achieves the best R@K and N@K scores (e.g., R@10=5.52%, N@10=2.82%), outperforming our method which generates SID prefixes rather than item rankings (R@K/N@K not directly applicable). This is expected: sequential recommenders are purpose-built for this exact task with full behavioral sequences. On Category Match, the cross-paradigm metric applicable to both approaches, our method achieves the **highest score (20.0%)**, outperforming TIGER (19.8%), SASRec (17.3%), and BERT4Rec (17.0%). Crucially, sequential models cannot substitute for our system: (1) they require behavioral sequences and thus *fail entirely* for cold-start users (20–30% of active users; see Appendix L); (2) they handle only next-item prediction, not the 5 implicit dialogue intents; (3) they lack natural language interaction capability.

TIGER Analysis. TIGER (Rajput et al., 2023) underperforms SASRec/BERT4Rec on

R@K/N@K primarily because its stable-catalog assumption is violated in news (articles cycle out within 24h), causing learned SID patterns to become stale at test time. Despite this, TIGER achieves the second-highest Category Match (19.8%), confirming that SID generation captures coarse semantic patterns even when fine-grained prediction fails—motivating our design choice of 3-layer prefix generation.

G Dense Retrieval Baseline Details

To provide a comprehensive retrieval comparison, we additionally evaluate dense retrieval using **BGE-large-zh** (Xiao et al., 2024), the same encoder used in our RQ-VAE codebook construction, with FAISS inner-product search over all 163K articles.

Setup. For each test sample, we extract a query from user profile and recent browsing history (up to 3 recent article titles), encode it with BGE-large-zh, and retrieve the top-50 candidates via FAISS. The LLM (GPT-4 or Qwen-7B) then selects from the top-5 reranked candidates, identical to the BM25 RAG pipeline. We also evaluate a **Hybrid** retriever (BM25 $\times$ 0.3 + Dense $\times$ 0.7, score-level fusion).

Results. Table 3 reports dense retrieval results (rows marked $_{\text{Dense}}$). Dense retrieval improves over BM25 on explicit-keyword intents (e.g., candidate selection) where semantic matching captures paraphrases that lexical matching misses. However, for *implicit intents*, where the query contains no retrievable keywords ("recommend something different," "what else?"), dense retrieval faces the same structural limitation as BM25: the query embedding lacks discriminative content to retrieve relevant articles. Full per-task breakdown is in Table 9.

Analysis. The key insight is that across all query construction strategies we evaluated (BM25, Dense, Hybrid), retrieve-first approaches show consistent gains on explicit-keyword intents but limited improvement on implicit intents. Dense retrieval improves the *explicit* query pathway but does not address the 5 implicit intent types that constitute 76% of our test set. This supports our design choice of a generate-first paradigm for implicit-intent-heavy conversational recommendation. We acknowledge that stronger query construction methods (e.g., LLM-based query rewriting) could narrow this gap; exploring such approaches is future work.

Task Type	GPT-4	Qwen-7B
Explicit (retrievable keywords):		
Candidate Sel.	34.1	28.1
Implicit (no retrievable keywords):		
Next-Item Pred.	11.0	10.1
Diversity Expl.	10.8	9.7
Cold-Start PADR	10.2	9.5
Pure Cold-Start	9.8	8.9
Feedback Adj.	10.5	9.3

Table 9: Per-task RAG breakdown (BM25 retrieval). Candidate selection uses Hit@1 from 5 fixed candidates; implicit tasks use Hit@1 from top-10 retrieved candidates (an easier setting than the L1 match over 152K SIDs used in Table 5). Implicit tasks cluster near the random-selection baseline (≈10%), confirming that retrieval cannot surface relevant articles without explicit query keywords. Dense retrieval (BGE-large-zh) yields similar patterns.

Metric	Ours	GPT-4+Hyb.	Δ	p
L1 Match (%)	**12.4**	12.4	0.0	0.98
L2 Match (%)	**1.0**	0.5	+0.5	<0.01
Cat. Match (%)	**20.0**	18.8	+1.2	<0.05
Hal. (%)	0	0	0.0	—
Cold L1 (%)	**14.9**	10.2†	+4.7	<0.01
Pure Cold L1 (%)	**18.0**	9.8†	+8.2	<0.01
Infer. Cost	7B (1 GPU)	GPT-4 API	∼100×	

Table 10: Multi-dimensional comparison: Ours vs. GPT-4+Hybrid RAG. p: paired bootstrap (10K resamples). †Hit@1 from top-10 retrieved candidates (Table 9); note this is an easier setting than L1 match over 152K SIDs, so the true gap is larger.

H Multi-Dimensional Comparison with Best RAG Baseline

A natural question is whether our L1=12.4% is meaningfully different from GPT-4+Hybrid RAG (also 12.4%). We provide a multi-dimensional statistical comparison showing that the two methods achieve the same L1 through *qualitatively different* mechanisms, with our method dominating on finer-grained metrics.

Key Observations. (1) **L1 parity masks finer-grained superiority**: Our method doubles L2 match (1.0% vs. 0.5%, p<0.01) and improves Category Match by 1.2pp (20.0% vs. 18.8%, p<0.05), indicating better fine-grained semantic localization within coarse categories. (2) **Cold-start is the decisive differentiator**: RAG's cold-start L1 drops to 10.2% (BM25 from profile keywords) and 9.8% (pure cold-start), while our method achieves 14.9% and 18.0% respectively, representing a +4.7pp and

+8.2pp advantage (both $p<0.01$). This confirms our core claim: the generate-first paradigm's advantage is most pronounced precisely where retrieve-first fails. (3) **Cost efficiency**: Our 7B model on a single GPU achieves comparable or superior results to GPT-4 API calls, at $\sim$100$\times$ lower inference cost.

I Candidate Selection Evaluation Design

Candidate selection (24% of test data) complements our primary open generation metric (76%). The **Rand** setting (4 random negatives) tests basic semantic discrimination; **Align** (2 same-category + 2 random) tests fine-grained within-category preference. Open generation is the primary metric because it eliminates negative sampling bias entirely (model generates SID prefixes in the full 152K space) and covers 76% of test data. The Rand$\rightarrow$Align drop ($-$28.5pp) reveals the gap between coarse discrimination and fine-grained preference modeling, motivating future work on within-category ranking. Same-backbone models share identical Hit@1 under deterministic decoding because the LLM's selection is deterministic given fixed candidates.

J Deployment Status

J.1 Internal Pilot Study Details

We deployed NewsRec-Chat as an internal pilot on a major Chinese news platform for 38 days (Jan 9 – Feb 15, 2026) to 300+ internal users. 92 active users conducted 141 sessions totaling 331 conversation turns, interacting freely without scripted tasks or time constraints. 58.9% of sessions involved $\geq$2 turns and 22.8% of users returned for additional sessions, confirming genuine conversational engagement. User queries naturally spanned all trained intent types. The system recommended 731 unique articles spanning 29 categories, with no single category exceeding 20%. Across all 950 recommended articles, **zero hallucination complaints** were reported.

Additionally, internal content editors and product managers conducted qualitative evaluations covering all 6 intent types, confirming that (1) zero-hallucination grounding eliminates "non-existent article" complaints, (2) cold-start recommendations were perceived as relevant, and (3) multi-turn refinement outperforms single-turn interactions.

Production Safeguards. The system implements a multi-level fallback cascade: **Level 1**, standard SID generation + fuzzy matching (δ=5); **Level 2**, broadened matching (δ+5); **Level 3**, category-level popular articles from profile interests; **Level 4**, platform-wide trending news. In testing, Level 2+ fallbacks activated in <5% of requests.

Config	$\mathcal{G}_{gen}$	$\mathcal{G}_{sys}$	Hal.
Full System	✓	✓	**0%**
w/o Gen	×	✓	100%
w/o Sys	✓	×	0%

Table 11: Grounding layer analysis.

Group	N	Actual	E[Rand]	Adj.	Lift
Warm Tasks	6300	11.9%	6.9%	5.0%	1.72×
Cold Tasks	1333	14.9%	6.8%	8.1%	2.19×
Pure Cold	721	18.0%	7.3%	10.8%	2.48×

Table 12: Distribution-corrected cold-start analysis. E[Rand]=expected random L1 from distribution concentration ($\sum_i p_i^2$). Adj.=Actual$-$E[Rand]. Lift=Actual/E[Rand].

K Grounding Layer Analysis

Table 11 isolates each layer's contribution. Generation grounding ($\mathcal{G}_{gen}$) is the primary factor: when SID-prefix matching is enabled, hallucination is eliminated. System grounding ($\mathcal{G}_{sys}$) primarily affects freshness through pool update constraints.

L Cold-Start Distribution Analysis

To verify that the cold-start L1 $>$ warm L1 advantage is not an artifact of category distribution concentration, we compute the *expected random L1*, i.e., the L1 match rate achievable by a random predictor that follows each group's ground-truth L1-code distribution: $\mathrm{E}[\mathrm{L1}_{\mathrm{random}}] = \sum_i p_i^2$, where p_i is the proportion of L1 code i.

If the cold-start advantage were purely a distribution artifact, the adjusted L1 (Actual $-$ E[Rand]) would be similar across groups. After correcting for L1-code distribution concentration (E[Rand]: warm 6.9%, cold 6.8%), the adjusted L1 gap persists: cold 8.1% vs. warm 5.0% (+3.1pp), with Pure Cold-Start reaching 10.8%. This confirms the advantage stems from PADR's profile-based reasoning quality, not distributional bias.

The counter-intuitive cold $>$ warm L1 result (18.0% vs. 11.9%) arises from two complementary mechanisms:

(1) Rich Profile Signals Beyond Demographics. User profiles on the production platform are

not limited to basic demographics (age, gender, location). They aggregate *long-term behavioral summaries*: declared interest topics at registration, reading-duration-weighted category affinities (Top-3 over 30-day and 7-day windows), activity patterns (time slots, reading duration), and engagement level, totaling 25+ features per user. Even users classified as "cold" (no *recent* clicks in the test window) retain these accumulated preference summaries. PADR's cold-path CoT leverages these signals to perform *interest-to-cluster inference* (e.g., "declared interest: technology + engagement level: high → technology/AI SID cluster"), which is often more focused than the noisy signal from sparse recent clicks.

(2) Search Space Narrowing Effect. Warm-path reasoning must resolve fine-grained preferences within broad behavioral contexts: a user who clicked 30 articles across 8 categories requires the model to disambiguate among many plausible clusters. Cold-path reasoning, by contrast, maps a focused profile summary to a smaller set of high-probability clusters. The effective L1 search space for cold-start predictions (mean 3.2 candidate L1 codes per profile) is smaller than for warm predictions (mean 6.8), making accurate generation easier despite less granular input.

(3) Production Safeguards. The system includes multi-level fallbacks: if SID generation fails or fuzzy matching returns zero candidates, the system cascades through (i) broadened prefix matching (δ+5), (ii) category-level popular articles from profile interests, and (iii) platform-wide trending news. In internal testing, these fallbacks activated in <3% of cold-start sessions, confirming PADR handles the vast majority of cold-start cases without degradation.

M Semantic Relevance of Partial Matches

While 12.4% L1 match may appear modest in absolute terms, L1-correct predictions that miss at L2 still yield *semantically related* candidates after fuzzy matching. We quantify this for the 7,191 open-generation test samples with valid SID predictions.

Partial Match Breakdown. Of valid open-generation predictions: 12.4% achieve exact L1 match (correct coarse category), of which 1.0% also match L2 (correct fine-grained topic). The remaining 11.4% are L1-correct but L2-incorrect (*partial matches*, N=813), which we analyze below.

Category-Level Relevance. Since partial matches share the same L1 code as the target by definition, they map to the same majority editorial category. The average L1-code purity, defined as the fraction of articles within each L1 cluster that belong to its majority editorial category, is **50.9%** across partial-match L1 codes. This indicates that while L1 codes are not perfectly aligned with editorial categories, roughly half of all articles in the predicted L1 cluster share the target's primary editorial category, providing a meaningful topical signal.

Effective Recommendation Quality. After fuzzy matching (δ=5), partial-match predictions return on average **5.2** candidate articles from the current pool (median: 3.0), of which **2.1** (53.6%) share the target's editorial category. All 813 partial matches successfully retrieve at least one candidate. This means that even when the model does not exactly pinpoint the target article's SID cluster, the fuzzy-matched candidates are predominantly topically relevant, and the effective recommendation quality is higher than the L1 metric alone suggests.

τ	Cold	Hybrid	Warm	Consideration
5	18.2	4.1	77.7	Too few cold samples
10	18.2	12.3	69.5	**Balanced (chosen)**
15	18.2	19.8	62.0	Hybrid path too broad
20	18.2	25.6	56.2	Warm data underused

Table 13: PADR path distribution (%) across τ thresholds (test set). Cold-path users ($|\mathbf{h}_u|$=0) are constant; higher τ shifts users from warm to hybrid.

N Cold-Start Threshold Sensitivity

We analyze the impact of the cold-start threshold τ on system behavior by considering $\tau \in \{5, 10, 15, 20\}$. The threshold τ determines the PADR routing path: users with $|\mathbf{h}_u| \geq \tau$ are routed to the warm path, $|\mathbf{h}_u| = 0$ to the cold path, and $0 < |\mathbf{h}_u| < \tau$ to the hybrid path.

We select τ=10 based on the following reasoning: (1) *Profile signal quality*: Users with <10 interactions have sparse behavioral signals that often do not reliably indicate stable preferences; supplementing with profile signals via the hybrid path is more robust. (2) *Cold-start coverage*: At τ=10, 30.5% of test users are routed through cold

Method	Hit@1 (95% CI)	L1 (95% CI)	p	d
Ours	59.3 [57.3, 61.2]	12.4 [11.7, 13.2]	—	—
BM25	23.5 [21.7, 25.1]	10.5 [9.9, 11.2]	<0.01	0.78
GPT-4	34.4 [32.5, 36.4]	–	<0.01	0.51
Qwen-7B	28.1 [26.3, 29.9]	–	<0.01	0.66

Table 14: Statistical tests for Table 3 (Rand setting, candidate selection). Same-backbone models yield identical Hit@1 under deterministic decoding; one row per backbone. CIs: 95% bootstrap (10K resamples, seed=42). p: paired bootstrap p-value vs. Ours. d: Cohen's d.

or hybrid paths, matching the production cold-start rate (20–30%). (3) *Training-inference alignment*: The $\sim$31% cold-start training allocation in Stage 2 aligns well with the $\sim$30% cold/hybrid distribution at τ=10. Smaller τ=5 underutilizes profile signals for sparse users; larger $\tau \geq 15$ over-routes users to the hybrid path, diluting the benefit of available behavioral data.

O Statistical Significance Tests

We report comprehensive statistical tests using paired bootstrap resampling (10K resamples, seed=42) to support all claims in the main text.

Table 3 Comparisons. Table 14 reports 95% bootstrap CIs for all methods under **random negatives** and paired significance tests (Ours vs. each baseline). All Hit@1 differences are significant at $p<0.01$ with medium-to-large effect sizes (Cohen's $d \geq 0.51$), confirming the robustness of our improvements. Under **hard negatives** (Align), Ours (30.8% [28.9, 32.7]) overlaps with GPT-4 (30.9% [29.0, 32.8]; p=0.94), while significantly outperforming Qwen-7B (26.0% [24.2, 27.8]; $p<0.01$, Cohen's d=0.11).

Table 5 Per-Task CIs. For tasks with smaller sample sizes (e.g., Feedback Adjustment N=621, Cold-Start PADR N=611), bootstrap CIs are wider but all Ours vs. Qwen-7B differences remain significant at $p<0.01$. Key per-task CIs (Rand): Candidate Selection Hit@1 59.3% [57.3, 61.2]; (Align): 30.8% [28.9, 32.7]; Feedback Adjustment Hit@1 54.8% [50.9, 58.6]; Next-Item L1 12.0% [11.1, 12.9]; Pure Cold-Start L1 18.0% [15.3, 20.8].

Cold-Start Gap Significance. The cold-start advantage (Cold L1 14.9% [13.0, 16.8] vs. Warm L1 11.9% [11.1, 12.7], +2.9pp) is significant at p=0.002 (unpaired bootstrap, 10K resamples, Cohen's d=0.086). Pure Cold-Start L1 (18.0% [15.3, 20.8]) is significant vs. Warm at $p<0.01$.

Dimension	Mean	Std	$\geq$4	IAA
Logical Coherence	4.2	0.7	82%	0.71
Recommendation Consistency	4.0	0.8	76%	0.68
Information Utilization	3.9	0.9	72%	0.65
Length Compliance	4.5	0.5	91%	0.82
Overall	**4.1**	0.7	80%	0.72

Table 15: Human evaluation of 200 sampled GPT-4 CoT traces. $\geq$4: fraction rated 4 or 5. IAA: inter-annotator agreement (Krippendorff's α).

Ablation Effect Sizes. Full Model vs. w/o Stage 2: Hit@1 Cohen's d=0.15 (small), L1 Cohen's d=0.13 (small). Despite modest effect sizes, all differences are statistically significant ($p<0.01$) due to large N=9,982.

P Extended Ablation Discussion

w/o Stage 1 (Direct Stage 2 Training). In preliminary experiments, direct Stage 2 training (skipping Stage 1) resulted in >60% non-SID token outputs, >50% hallucination, and L1 <3%, confirming that Stage 1's 483K-sample SID alignment is essential for the model to learn valid SID semantics before reasoning.

CoT vs. No-CoT Stage 2. The w/o Stage 2 ablation (Table 4) shows that Stage 1 alone (no CoT) achieves lower L1 and 18.4% hallucination, indicating CoT distillation contributes both to accuracy and grounding discipline by providing a reasoning scaffold that regularizes SID generation.

Cold-Start Training Proportion. The $\sim$31% cold-start allocation in Stage 2 matches the production cold-start rate (20–30%). Training with <15% cold-start samples degraded cold-start L1 by $\sim$3pp; >40% reduced warm-start L1 by $\sim$1pp.

Q CoT Quality Assessment

Human Evaluation of GPT-4 CoT. We randomly sampled 200 CoT traces from the 48K Stage 2 training set (stratified across 6 task types) and had 3 annotators evaluate each along 4 dimensions on a 1–5 scale:

Logical Coherence: whether the reasoning chain follows a clear logical progression from user signals to SID generation. **Recommendation Consistency**: whether the generated SID aligns with the CoT's stated reasoning. **Information Utilization**:

how effectively the CoT leverages available user signals (profile, history, dialogue context). **Length Compliance**: whether the CoT falls within the 150–300 character target.

Overall quality is high (mean 4.1/5, 80% rated $\geq$4), with strong inter-annotator agreement (α=0.72). The weakest dimension is Information Utilization (3.9), primarily for cold-start samples where limited signals constrain reasoning depth.

CoT Length Selection Rationale (150–300 Characters). We selected the 150–300 character range based on preliminary experiments comparing 4 length brackets on a 1K validation set:

- **<100 chars**: Insufficient reasoning depth; L1 dropped 1.8pp vs. optimal.
- **100–150 chars**: Marginally better but still shallow; CoT often skips profile analysis.
- **150–300 chars**: Best trade-off; covers profile analysis → interest mapping → SID generation in 2–3 reasoning steps.
- **>300 chars**: Introduces redundancy and repetitive reasoning; critically, longer training CoTs cause **inference-time degradation**, as the model generates verbose, unfocused reasoning at inference (mean inference CoT: 29 chars when trained on >300-char CoTs vs. 187 chars when trained on 150–300-char CoTs), collapsing the reasoning quality.

The 150–300 range ensures sufficient reasoning depth while maintaining inference-time stability.

Q.1 CoT Examples

We provide representative examples of GPT-4-generated Chain-of-Thought reasoning traces used in Stage 2 training (translated from Chinese for readability). Each CoT follows the controlled length constraint (150–300 characters).

Warm-Start Example (Next-Item Prediction).

> *User Profile: Male, 35-44, Beijing, interests: technology, finance. Recent history: 5 tech articles (AI, smartphones), 2 finance articles.*
>
> **CoT**: "The user's recent browsing is dominated by technology news, especially AI and mobile devices. Combined with the profile showing strong tech and finance interests, the next recommendation should target the technology cluster. The user's age and location suggest interest in practical tech applications rather than consumer gadgets. Generating SID prefix targeting technology/AI cluster."
>
> → **SID**: <s_a_14><s_b_55><s_c_36>

Pure Cold-Start Example (Zero History).

> *User Profile: Female, 25-34, Shanghai, declared interests: entertainment, lifestyle. No browsing history.*
>
> **CoT**: "This is a new user with no behavioral data. Profile analysis: young female professional in a major city with entertainment and lifestyle interests. Demographic patterns suggest interest in trending entertainment content, such as celebrity news, variety shows, or lifestyle tips. Generating SID prefix targeting the entertainment/variety cluster based purely on profile inference."
>
> → **SID**: <s_a_7><s_b_22><s_c_81>

These examples illustrate the key design difference: warm-start CoT correlates behavioral signals with profile data, while cold-start CoT performs demographic-to-interest inference without behavioral evidence.

Q.2 CoT Length and Inference-Time Behavior

Why Long Training CoT Causes Short Inference CoT. Training on >300-character CoTs causes inference-time CoT to collapse to ∼29 characters (vs. 187 chars for 150–300 char training). We attribute this to **length distribution mismatch**: when training CoTs are excessively long, they contain redundant reasoning steps (e.g., repeating the same profile analysis in different phrasings). The model learns to associate the "end of reasoning" signal with length-based cues rather than logical completion. At inference, without the teacher-forced long context, the model prematurely triggers the SID generation token after minimal reasoning.

This is *not* evidence that the model fails to reason. Rather, it shows that CoT quality, specifically conciseness and logical density, matters more than CoT length. Our 150–300 character range produces CoTs that are (a) logically complete (covering profile→interest→SID in 2–3 steps), (b) free of redundancy, and (c) stable at inference (mean 187 chars, within the training range). The human evaluation (Table 15) confirms this: Logical Coherence scores 4.2/5 and Length Compliance scores 4.5/5 for the 150–300 range.

Broader Implication. This finding aligns with recent observations on CoT distillation (Li et al., 2023): effective distillation requires not just correct reasoning, but *appropriately scoped* reasoning that the student model can reliably reproduce.

R Qualitative Analysis

We illustrate two scenarios where Generate-then-Match succeeds and RAG fails (translated from Chinese).

Case 1: Implicit Intent. User says: "*I'm bored with what I usually read, recommend something different.*" This query contains no retrievable keywords. NewsRec-Chat reasons via CoT: (1) user's history shows 80% sports news; (2) profile declares "technology" interest but has zero tech clicks; (3) generates SID prefix $(14, 55, _)$ targeting the technology cluster; (4) fuzzy match returns "*Apple announces new AI features*," a genuinely different, grounded recommendation. RAG would return nothing or generic popular results.

Case 2: Cold-Start PADR. New user (zero history), profile: Beijing, 28yo male, declared interests in AI/startups. PADR's cold-path CoT: (1) demographics $\rightarrow$ tech-savvy professional; (2) declared interests $\rightarrow$ technology SID clusters; (3) generates prefix $(14, 55, 36)$. Result: relevant AI startup coverage despite zero behavioral signals. This user type achieves **18.0% L1 match**, the highest among all user types.

S Error Analysis

We analyze 200 randomly sampled failure cases to understand system limitations:

Category Distribution of Failures:

- **Category Mismatch (42%)**: SID prefix correctly identifies semantic cluster but mismatches user's *current* intent. Example: User browsing sports news suddenly asks about stock market; cached sports-focused prefixes return irrelevant results.
- **Temporal Drift (28%)**: Cached SID prefixes become stale for rapidly evolving topics. Example: Breaking news events create new semantic clusters not captured by existing prefixes.
- **Ambiguous Query (18%)**: User queries lack sufficient context for intent disambiguation. Example: "News" without topic specification leads to overly generic recommendations.
- **Profile Sparsity (12%)**: Cold-start users with minimal profile information receive suboptimal recommendations. Example: New user with only demographic info (no declared interests).

Mitigation Strategies:

- **For Category Mismatch**: Intent-aware cache invalidation triggers re-inference when detected intent shift exceeds threshold.
- **For Temporal Drift**: Reduce cache TTL for trending topics; trigger Enhance Track more frequently during breaking news periods.
- **For Ambiguous Query**: Proactive clarification in dialogue ("What topics interest you today?") when query confidence is low.
- **For Profile Sparsity**: Leverage popularity signals and trending topics as fallback for extremely cold users.

T Profile Features Analysis (Data Leakage Exclusion)

We conduct a feature audit to confirm that the cold-start L1 $>$ warm L1 advantage does not stem from data leakage in profile features.

The 25+ profile features fall into 5 categories: demographics (age, gender, city tier), long-term category preferences (30-day/7-day reading-duration-weighted Top-3), behavioral patterns (active time slots, engagement level), content format preferences (video/text affinity), and declared interest tags from onboarding. All features are computed from *historical* data **before the temporal split cutoff**; test items are published **after** the cutoff. We verified that zero profile features contain article IDs, SIDs, or titles from the test period. The strongest signal is category-level preference (e.g., "user prefers technology news"), which is the intended input for cold-start reasoning, not leakage. The cold-start advantage arises because profile-based reasoning provides *concentrated, noise-free* category signals, whereas warm users' behavioral histories introduce noise from exploratory clicks and temporal preference shifts.

U CRS Methodological Comparison

Direct numerical comparison with UniCRS (Wang et al., 2022) or iEvaLM (Wang et al., 2023) is infeasible due to fundamentally divergent settings. The key differences are: *(a) Item Lifespan*: CRS benchmarks (ReDial) assume stable catalogs; our news pool refreshes daily (most articles cycle out within 24h). *(b) Evaluation*: UniCRS/iEvaLM use recall and BLEU on fixed item sets; we evaluate SID-based open generation (152K space) and hallucination rate. *(c) Intent Scope*: Prior CRS primarily handle explicit preference elicitation; we address 6 intent types including 5 implicit ones. *(d) Cold-Start*: CRS methods assume user history; PADR enables profile-only reasoning (18.0% L1).

V Deployment Lessons

(1) SID Codebook Drift. The SID codebook gradually misaligns with emerging topics. Weekly re-clustering (163K articles) combined with fuzzy matching (δ=5) accommodates minor shifts without model retraining, since the model generates *semantic intent* prefixes (not memorized indices) and L1 codes remain stable across re-clustering.

(2) Latency-Quality Trade-off. Initial designs suffered 10–20s latency. The dual-track architecture emerged as a solution: users receive immediate (85ms) responses from cached SID prefixes while the system improves quality asynchronously.

(3) Dialogue Capability Degradation. SID fine-tuning initially degraded conversational fluency. Two-stage training with dialogue preservation tasks (53% in Stage 1) mitigates this, maintaining natural language quality while adding SID generation capability.

W Intent Taxonomy IAA Study

To validate the reliability of our 6-intent taxonomy (Table 1), we conducted an inter-annotator agreement (IAA) study. We sampled 500 stratified dialogues from the test set, removed template prefixes to prevent label leakage, and wrote annotation guidelines based on observable signals (e.g., history length for cold-start, presence of candidate lists for candidate selection, negative feedback keywords for feedback adjustment). Seven expert annotators independently labeled each dialogue with one of the 6 intent types.

Results. Fleiss' κ=0.81 (6-way classification) and 0.91 (binary explicit/implicit), both exceeding the "almost perfect" threshold of 0.81 (Landis and Koch, 1977). Mean pairwise Cohen's κ=0.81. The high agreement confirms that our taxonomy is operationally well-defined: intent types are determined by observable signals rather than subjective interpretation, making the taxonomy reproducible across annotators and deployable as an automatic routing mechanism in production.

X Limitations and Ethics Statement

Limitations: (1) SID quality depends on the RQ-VAE codebook, requiring periodic re-clustering; (2) while our internal pilot (300+ users, 38 days) demonstrates real-world viability, a controlled online A/B test comparing against the production recommender would further strengthen deployment claims; (3) GPT-4 CoT distillation ($\sim$\$850) creates a training-time dependency on proprietary models; recent open-source reasoning models (e.g., DeepSeek-R1, Qwen2.5-72B) could serve as alternatives, and we leave this exploration to future work; (4) open generation L1 match (12.4%) reflects the difficulty of predicting among 152K SIDs; partial matches yield topically relevant candidates (Appendix M), but absolute accuracy remains modest; (5) candidate selection is a complementary metric (24% of data); we report both random and hard-negative settings to ensure transparency, and our primary open generation evaluation (76%) is unaffected by negative sampling choices; (6) our RAG baselines include both BM25 (sparse) and BGE-large-zh (dense) retrieval; a more controlled comparison would fine-tune the dense retriever on the same task data, which we leave to future work. However, the structural limitation of all retrieve-first approaches, namely the inability to handle intents where no query keywords exist, is independent of retrieval quality.

Future Work: Reinforcement learning with carefully designed reward functions (incorporating dwell time, explicit feedback, and long-term engagement) could further improve recommendation quality beyond supervised CoT distillation. Replacing GPT-4 with open-source reasoning models for CoT generation would reduce cost and improve reproducibility.

Ethics Statement: All user data is anonymized. Diversity exploration tasks (12.4% of Stage 2) mitigate filter bubbles; the grounding guarantee ensures recommendations are real articles; the editorial team reviews recommendation patterns weekly. We acknowledge that PADR's cold-start reasoning from demographic features (age, gender, location) carries a risk of reinforcing stereotypical associations; we mitigate this by including declared interest features and diversity exploration samples in training, and the editorial review process monitors for demographic bias in recommendation patterns.