DMA: Enhancing Retrieval-Augmented Generation with Adaptive Human Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

Retrieval-augmented generation (RAG) systems typically rely on static retrieval methods, limiting their adaptability to dynamic environments. In this paper, we propose a novel online learning framework called Dynamic Memory Alignment (DMA), designed specifically to enhance retrieval performance and content generation in RAG through adaptive incorporation of multi-level human feedback. DMA systematically integrates real-time feedback signals at document, list, and response levels, effectively adjusting memory management strategies to optimize relevance and adaptability in online interactive environments. Extensive evaluations demonstrate DMA's competitive foundational retrieval performance across multiple standard knowledge-intensive benchmarks. DMA achieves significant improvements on datasets reflecting natural conversational interactions (TriviaQA, HotpotQA), confirming its suitability for online GenAI dialogue applications. Moreover, a multi-month industrial deployment demonstrates that DMA substantially improves user engagement in real-world applications. These results underscore DMA's ability to maintain robust foundational retrieval capabilities while excelling at dynamic, real-time adaptation in interactive online environments.

1 Introduction

Retrieval-augmented generation (RAG) has become a core paradigm for enhancing the factuality and adaptability of LLMs in knowledge-intensive tasks (Lewis et al., 2020; Borgeaud et al., 2022). By decoupling parametric memory from non-parametric retrieval, RAG enables models to access external information dynamically, grounding responses on up-to-date and domain-specific knowledge without modifying internal parameters. This separation has powered recent advances across open-domain QA (Izacard & Grave, 2021), multi-hop reasoning (Yang et al., 2018), and instruction-based augmentation (Lin et al., 2024; Gao et al., 2023).

Despite these advances, conventional RAG pipelines exhibit critical limitations in dynamic online settings: (i) Static retrieval strategies cannot adapt to evolving user intent or content drift. Most dense retrievers are trained offline and remain fixed at deployment time, failing to reflect live interaction signals (Lin et al., 2023; Jiang et al., 2024). (ii) Given the limited context length of mainstream LLMs (Liu et al., 2023), retrieval must prioritize highly relevant information. Sole reliance on top-k dense similarity often results in suboptimal recall and necessitates robust re-ranking strategies (Nogueira et al., 2020; Glass et al., 2022b; Qin et al., 2024). (iii) While dedicated rankers and hybrid retrievers can improve retrieval precision (Ma et al., 2023; Izacard et al., 2022), they often lack the flexibility and generalization needed for personalized, real-time adaptation (Zhang et al., 2024a). These issues collectively suggest that current RAG systems require an adaptive interface between user feedback and memory control.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Motivated by these challenges, our goal is to build an adaptive online learning framework for RAG systems that effectively integrates and utilizes dynamic human feedback, enabling continuous real-time refinement of memory and retrieval decisions. Recent studies demonstrate that instruction-tuned LLMs can effectively align responses with user intent through task-specific fine-tuning (Liu et al., 2024b; Lin et al., 2024). Real-time human feedback across document-, list-, and response-level granularity can serve as actionable supervision signals for adaptive retrieval. DMA incorporates these signals through continuous feedback-driven memory alignment.

To this end, we propose Dynamic Memory Alignment (DMA), an innovative online learning framework designed to systematically organize, interpret, and incorporate adaptive human feedback signals, dynamically optimizing retrieval strategies and memory prioritization within RAG workflows.

Specifically, DMA addresses the core challenge of online adaptability through three key components: (1) a multi-granularity feedback taxonomy tailored for conversational GenAI scenarios; (2) a suite of reward modeling techniques that interpret heterogeneous user signals into structured supervision; (3) online knowledge fusion mechanisms that prioritize high-value memory traces and modulate retrieval policy accordingly.

As a result, the DMA framework is particularly suited to real-time, user-facing applications such as chat assistants and enterprise QA bots, where system adaptability is key to sustained performance (Asai et al., 2024b; Jeong et al., 2024).

Our contributions can be summarized as follows:

- We propose DMA, a novel online learning framework enabling RAG systems to continuously refine adaptive retrieval based on multi-level user feedback. DMA systematically captures sparse yet valuable user signals to dynamically enhance system responsiveness in dynamic online settings.
- Through extensive evaluations on widely-used knowledge-intensive benchmarks, DMA achieves strong results on conversational datasets such as TriviaQA and HotpotQA, showing state-of-art performance than prior leading methods.
- Most critically, DMA demonstrates notable real-world applicability, as evidenced by a 24.57% improvement in positive user feedback during a multi-month randomized controlled industrial trial, validating its effectiveness and adaptability in practical deployment.

The remainder of this paper is structured as follows: § 2 surveys related work. § 3 formalizes the RAG problem and highlights key limitations of static pipelines. § 4 presents the proposed DMA framework. Experimental setup and results are detailed in § 5, while remaining challenges are discussed in § 6. We conclude in § 7.

2 Related Work

RAG has emerged as a core solution for knowledge-intensive NLP tasks (Lewis et al., 2020; Borgeaud et al., 2022). In standard RAG pipelines, a dense retriever (e.g., (Karpukhin et al., 2020)) encodes queries and documents into a shared embedding space, retrieving top-k relevant contexts from an external corpus. These retrieved contexts are then fused with the input query and processed by an LLM to generate grounded responses (Izacard & Grave, 2021; Izacard et al., 2023).

Recent research has focused on enhancing this pipeline along several directions. One thread optimizes retrieval to better align with the downstream generation needs of LLMs (Shi et al., 2024; Lin et al., 2024; Ye et al., 2023). Another line introduces multi-step and interleaved retrieval-generation mechanisms to capture complex reasoning chains (Trivedi et al., 2023; Shao et al., 2023; Jeong et al., 2024). Meanwhile, context filtering and selection strategies have been developed to remove noisy evidence before generation (Wang et al., 2023; Xu et al., 2024; Yoran et al., 2024), improving both factuality and efficiency.

In parallel, instruction tuning has become a critical enabler for aligning LLMs with retrieval-enhanced tasks. From supervised instruction collections like FLAN and Self-Instruct (Wei et al., 2022; Wang et al., 2022) to open-source alignment efforts such as ChatGPT and Claude (OpenAI, 2023; Anthropic, 2023), LLMs increasingly learn to operate over retrieved evidence. Recent studies demonstrate that retrieval-augmented instruction tuning significantly boosts performance across QA and reasoning tasks (Liu et al., 2024b; Asai et al., 2024b; Lin et al., 2024; Luo et al., 2023; Wang et al., 2024).

Nevertheless, integrating retrieval into LLM training remains challenging due to the need for surrogate losses and continuous re-indexing (Guu et al., 2020; Shi et al., 2024; Sachan et al., 2021; Izacard et al., 2023; Dong et al., 2024).

Ranking-based enhancements have been extensively used to improve retrieved context quality before generation. Early neural ranking models (Mitra et al., 2018; Chen et al., 2020) were later extended to dual-stage architectures such as Re2G (Glass et al., 2022b), PARADE (Drozdov et al., 2023), and RA-DIT (Lin et al., 2024), enabling more flexible reordering. However, these rankers often rely on moderate-sized encoder models (e.g., BERT or T5), which struggle with complex semantics and generalization (Ram et al., 2023). Recent evidence suggests that full-scale LLMs can act as powerful rankers with minimal prompting (Qin et al., 2024; Sun et al., 2023; Khalifa et al., 2023), yet leveraging this capacity in online RAG systems remains under-explored.

Crucially, most prior work optimizes retrieval and re-ranking on static datasets, assuming fixed user intent and corpus distribution. This paradigm fails to accommodate the non-stationary dynamics in real-world online systems, where user behavior, topic drift, and feedback evolve continuously. To bridge this gap, emerging approaches such as Self-RAG (Asai et al., 2024a), ReFeed (Yu et al., 2024), and Pistis-RAG (Bai et al., 2024) propose adaptive mechanisms incorporating implicit or explicit feedback. These methods are typically confined to limited settings and do not offer general-purpose integration into end-to-end retrieval and memory control.

In contrast, the DMA framework introduces a unified online learning architecture that encodes multi-level user feedback at document-, list-, and response-level granularity into dynamic retrieval optimization. Our approach maintains continuous feedback loops to enable retrieval and generation components to co-adapt during deployment, which supports sustained performance in open-ended, user-facing GenAI systems.

3 Preliminaries

This section formalizes the RAG pipeline that serves as the foundation for our work. We then identify key limitations of existing RAG approaches in dynamic online settings, which motivate the design of our proposed DMA framework.

3.1 Problem Setup

Let $\mathcal{C} = \{d_1, d_2, \ldots, d_N\}$ denote a corpus of external knowledge documents. Given a user query $q \in \mathbb{Q}$, a retriever R computes similarity scores using dense embeddings, typically in a dual-encoder setting (Karpukhin et al., 2020), where $\text{Relevance}(q, d_i) = \langle E_q(q), E_d(d_i) \rangle$ and E_q, E_d are the query and document encoders. The top-k documents are selected as $D_{\text{retrieve}} = \text{Top}_k\{\text{Relevance}(q, d_i) \mid d_i \in \mathcal{C}\}$. A reranker Rerank_m may be applied to reorder and truncate this list to the top-m items, yielding $D = \text{Rerank}_m(q, D_{\text{retrieve}})$, where $m \leq k$ (Cao et al., 2007; Glass et al., 2022a). The final context set $D = \{d^{(1)}, \ldots, d^{(m)}\}$ is concatenated with the query and fed into a language model G to generate a grounded response $a = G(q, D)$. While the retriever R and generator G may be trained separately or jointly (Sachan et al., 2021; Izacard et al., 2023), most real-world systems adopt modular training due to scalability and deployment constraints.

3.2 Limitations of Current Approaches

Despite their success in open-domain question answering and related tasks (Lewis et al., 2020; Borgeaud et al., 2022; Guu et al., 2020), current RAG systems exhibit structural limitations when deployed in dynamic, user-facing environments.

First, conventional RAG methods rely on static retrievers trained offline over frozen corpora, using task-specific training signals (e.g., NQ, TriviaQA) that do not generalize well to continuously evolving user needs (Lewis et al., 2020; Guu et al., 2020; Izacard et al., 2023). This fixed retrieval logic fails to accommodate domain drift, long-term user preferences, or topic shifts typical of online applications.

Second, although mechanisms such as reranking or filtering (Chen et al., 2020; Wang et al., 2023; Xu et al., 2024) can improve precision, they are typically rule-based or learned from fixed supervised data. These components rarely leverage live user feedback signals, and even when available, such signals are often aggregated in limited forms (e.g., binary preference) or only utilized post hoc.

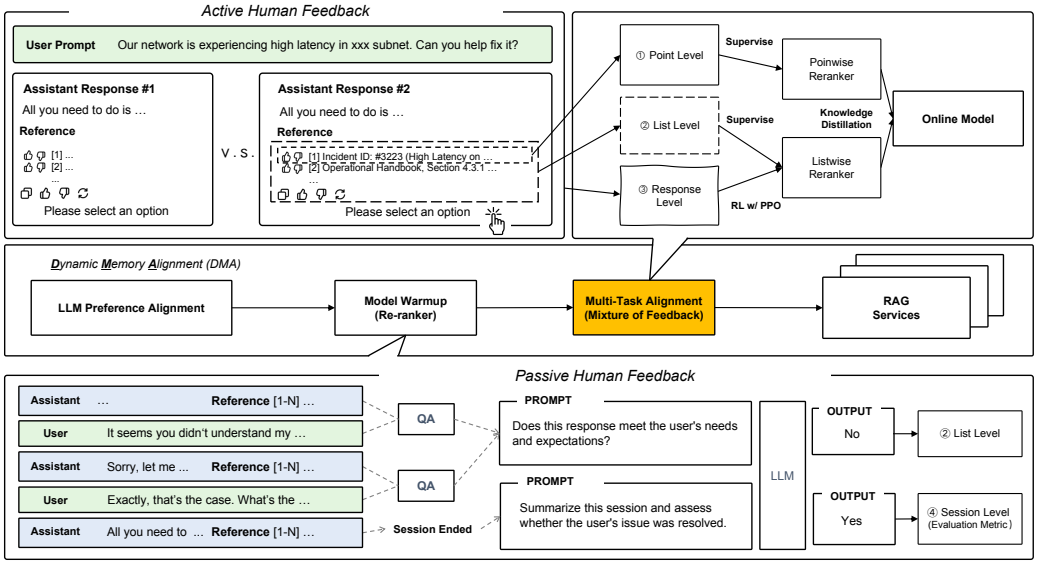

Figure 1: Overview of the DMA feedback loop. Multi-level human feedback is organized, modeled, and fused to guide online retrieval strategies. Reranker training and distillation are detailed in Figure 2.

Third, most RAG systems lack a principled framework to incorporate multi-granular feedback—such as document-level usefulness, list-level coverage, or response-level satisfaction—into real-time retrieval decisions. While reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and browser-based systems such as WebGPT (Nakano et al., 2021) have demonstrated the potential of fine-grained supervision, these approaches remain decoupled from the retrieval components and are difficult to generalize to streaming environments.

As a result, retrieval behavior remains largely fixed during deployment, limiting the system's ability to improve with usage, personalize to users, or adapt to shifts in content distribution. These limitations call for an online learning mechanism capable of dynamically integrating human feedback into memory and retrieval policies—precisely the gap that our proposed DMA framework aims to address.

4 Dynamic Memory Alignment

To address the limitations mentioned in the previous section, we introduce **DMA**, an online learning framework designed to continuously refine retrieval strategies in RAG systems by leveraging real-time user feedback. Unlike conventional static pipelines, DMA forms a closed-loop system that adaptively aligns memory and retrieval decisions with evolving user preferences.

4.1 Framework Overview

As illustrated in Figure 1, DMA comprises three core components: (1) **Feedback Taxonomy**, which structures heterogeneous user signals into well-defined levels; (2) **Reward Modeling**, which transforms these signals into trainable supervision; and (3) **Online Adaptation**, which updates retrieval strategies based on real-time feedback. Together, these modules form a dynamic feedback loop, enabling memory alignment in continually evolving GenAI interactions.

4.2 Human Feedback Taxonomy

Effective capture and utilization of user feedback in industrial settings require systematic organization. Addressing the challenge of sparse and heterogeneous feedback across user contexts, we investigate prominent LLMs, including ChatGPT (Achiam et al., 2023), Gemini (Team et al., 2023), QWen (Bai

et al., 2023), DeepSeek (Liu et al., 2024a), ChatGML (GLM et al., 2024), and Kimi (Team et al., 2025), to identify and categorize various forms of human feedback signals.

This taxonomy provides a systematic structure to interpret diverse feedback signals and optimize system behavior. As illustrated in Figure 1, feedback signals are categorized into four levels of granularity:

1) Document-level feedback reflects user evaluations of individual retrieved snippets, typically through direct actions such as upvoting or downvoting. This feedback is formalized into a preference dataset $\mathcal{D}_{\text{pref,doc}} = \{(q_i, d_i, y_{q_i,d_i})\}_{i=1}^{N}$, enabling optimization of document-level relevance.

2) List-level Feedback captures user preferences over a set of retrieved documents, evaluating the overall quality of system outputs for a query q_i based on a list D_{q_i} and system response y_{q_i}. This includes both explicit (e.g., copy, regenerate) and implicit feedback. It is formalized into a preference dataset $\mathcal{D}_{\text{pref,list}}$, providing insights into document relevance and ranking consistency for a list subset D_{sub,q_i}.

3) Response-level feedback refers to user preference between two (or more) response options generated from distinct document sets D_1 and D_2. The feedback signal y indicates the preferred response, implying a preference between the document sets. This is formalized as $\mathcal{D}_{\text{resp}} = \{(q_i, r_{1,i}, r_{2,i}, D_{1,i}, D_{2,i}, y_i)\}_{i=1}^{N_{\text{resp}}}$. This data is valuable for alignment methods and can be scaled.

Although the feedback is collected at the response level, each response is generated based on a specific document list. As such, user preference over responses implicitly reflects preference over the underlying document sets, which we leverage to supervise document-level reranking.

4) Session-level feedback aggregates user evaluations across an entire interaction session s_i, capturing overall user perceptions such as task satisfaction f_{s_i}. While this high-level signal is not used directly to train granular reward models, it is employed in two key roles: (i) as an external metric for evaluating DMA variants (§5.1); and (ii) as a dynamic weight signal to adjust fusion importance across feedback types during GBDT distillation (see §4.3) (Friedman, 2001).

By structuring feedback into these levels and formalizing the associated datasets, our taxonomy offers a robust framework for systematically interpreting user inputs and optimizing GenAI systems.

4.3 Reward Construction and Memory Alignment

To leverage the multi-granular feedback captured by the taxonomy for optimal DMA performance, we design specific modeling methods for each granularity level and develop strategies to combine their outputs to influence memory alignment. A multi-task modeling approach integrates diverse feedback signals to construct reward signals suitable for training memory alignment components.

Document-Level Modeling. A model is trained using explicit document-level feedback $\mathcal{D}_{\text{pref,doc}}$ with Binary Cross-Entropy (BCE) loss $\mathcal{L}_{\text{BCE}} = -y \log \sigma(s) - (1-y)\log(1-\sigma(s))$, where s is the predicted score and $y \in \{0, 1\}$ is the label. This produces a pointwise reranker focused on fine-grained precision.

List-Level Modeling. To capture relative importance among retrieved results, listwise rerankers are trained using user feedback. The ListNet loss $\mathcal{L}_{\text{ListNet}} = -\sum_i P_{\text{true}}(i) \log P_{\text{pred}}(i)$, where $P(i) = \exp(s_i)/\sum_j \exp(s_j)$, ensures alignment between predicted and target ranking distributions.

Response-Level Modeling. We collect pairwise user feedback comparing responses (r_1, r_2) generated from different document lists D_1, D_2, forming preference data $\mathcal{D}_{\text{resp}} = \{(q, D_1, D_2, y)\}$ with binary preference label y. A reward model $R(D)$ is trained via pairwise loss $\mathcal{L}_{\text{pairwise}} = -y \log \sigma(R(D_1) - R(D_2)) - (1-y)\log \sigma(R(D_2) - R(D_1))$, where $\sigma(\cdot)$ is the sigmoid function. To inject response-level preferences into the reranker, we apply Proximal Policy Optimization (PPO) (Schulman et al., 2017), optimizing a listwise policy using the clipped surrogate objective $\mathcal{L}_{\text{PPO}} = -\mathbb{E}_t[\min(r_t \widehat{A}_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\widehat{A}_t)]$, where $r_t = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ and $\widehat{A}_t$ is the advantage estimated from the reward model. This approach effectively aligns a listwise reranker using response-level feedback, producing the **PPO-aligned listwise reranker**, which captures global user satisfaction signals beyond document or list-level heuristics.

Fusion and Distillation for Online Serving To meet the latency requirements of online serving, we distill the outputs of upstream feedback-supervised rerankers into a lightweight ensemble model. Specifically, we adopt a Gradient Boosting Decision Tree (GBDT) as the final online scoring module.

This GBDT model is trained using soft labels derived from upstream reranking components and provides efficient inference without sacrificing alignment quality. In production deployment, it enables real-time document ranking with sub-10ms latency while preserving the benefits of multi-granular feedback supervision.

Figure 2 illustrates the high-level training pipeline. Additional modeling and supervision details are omitted for brevity and deployment sensitivity.

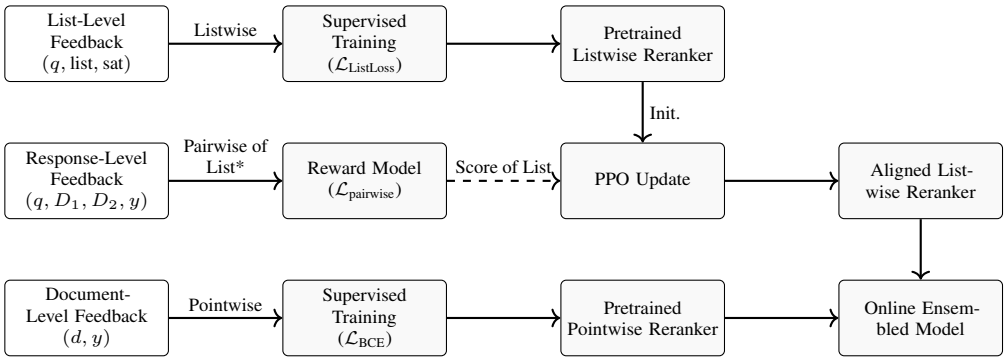

Figure 2: Training and distillation architecture in DMA. Three types of feedback supervise specialized ranking models, which are fused via PPO and distilled into a lightweight online reranker. Full pipeline details in § 4.3

5 Experiment

We evaluate DMA across two settings: (1) real-world online interactions with online users, and (2) public open-domain QA benchmarks. The former validates DMA's online learning ability in production; the latter assesses generalization under static evaluation protocols.

5.1 Experiment Setup

Evaluations on API Distribution. We conduct a multi-month randomized controlled trial (RCT) on a Chinese-language GenAI system operated by a major telecommunications and cloud provider. To support multilingual retrieval, DMA uses BGE-m3 as the retriever backbone and an instruction-tuned decoder for response generation.

To characterize domain diversity, session queries are categorized into seven application areas: Technical support (37%), Performance and monitoring (21%), API and developer support (16%), Security and compliance (10%), Service and resource management (9%), Migration and deployment (4%), and Product features and updates (3%). The query distribution reflects the industrial and technically specialized nature of the evaluation environment.

For measurement, we define session-level satisfaction as $S(s_i) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{I}(\text{LLM}(q_{i,j}, y_{i,j}) \neq \texttt{dissatisfied})$, where each session s_i contains n_i user turns. Because explicit user ratings are sparse, we employ QWen2-72B (Yang et al., 2024) as an automated annotator to infer satisfaction labels, calibrated via in-context few-shot learning using high-quality examples from online human feedback.

To ensure alignment with human judgment, these labeled sessions are treated as ground-truth supervision during prompt construction. The prompt template (Table 1) specifies the annotator's role, input-output format, and exemplar completions. The structured outputs include a categorical label (`satisfied`, `neutral`, or `dissatisfied`), a confidence score in $[0, 1]$, and a short list of improvement suggestions.

This automated feedback serves as the primary evaluation signal for DMA under real-world usage. Inter-annotator agreement analysis confirms high label reliability, with a Cohen's Kappa of 0.962 between model predictions and human annotations.

Table 1: Session-level User Satisfaction Evaluation Prompt Design

Intent	Prompt
Role	You are an AI assistant responsible for evaluating user satisfaction at the session level.
Task	Assess overall user satisfaction based on the entire conversation history, including user queries and system responses.
Input	A session s_i consisting of n_i turns: $\{(q_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, where $q_{i,j}$ is the user query and $y_{i,j}$ is the system-generated response.
Few-shot Examples	{`few_shot_examples`} illustrating different types of session outcomes.
Output Format	**User Satisfaction**: `satisfied` / `neutral` / `dissatisfied` **Confidence**: A numerical value in $[0, 1]$ representing model confidence. **Improvements**: A short list of suggestions to improve the user experience.

Evaluations on public static benchmarks. To evaluate generalization in static settings, we test DMA on four standard open-domain QA datasets: Natural Questions (NQ: 79.2k train / 8.7k dev / 3.6k test) (Kwiatkowski et al., 2019), TriviaQA (78.8k / 8.8k / 11.3k) (Joshi et al., 2017), HotpotQA (88.9k / 5.6k / 5.6k) (Yang et al., 2018), and WebQSP (2.8k / 250 / 1.6k) (Berant et al., 2013). These benchmarks span a range of query types, from open-ended to structured factoid-style tasks. We report Hit@1 and F1 following prior work. To ensure comparability with existing methods. All generations were performed using a unified LLaMA2-7B decoder (Touvron et al., 2023), controlling for decoding variability and isolating retrieval alignment effects.

Implementation Details. DMA's online update pipeline is triggered after accumulating 500 new feedback samples using Flink-based monitoring. This threshold was empirically selected to balance the need for timely adaptation against the computational cost of frequent retraining. It ensures that model updates are based on sufficient feedback to generate stable gradient signals, while also preventing excessive latency in high-throughput environments. In practice, this results in update intervals ranging from several minutes to an hour, depending on traffic volume.

To accommodate variable traffic conditions, the feedback monitoring system automatically defers updates if insufficient feedback is collected, avoiding retraining on sparse or noisy signals. This adaptive scheduling ensures robustness across deployment scales, from high-traffic production environments to slower-feedback applications.

The full DMA update cycle includes: (1) training pointwise and listwise teacher models, (2) generating soft distillation targets, and (3) training a 10K-tree GBDT model. Over 90% of the latency is spent on teacher model training ($\approx$ 6 minutes) and distillation ($\approx$ 3 minutes), with model checkpoint updates taking less than 1 minute. The system runs on 8 A800 GPUs per training job, yielding an average end-to-end update latency of 10 minutes (range: 6–15 minutes). To maintain sub-15-minute updates as feedback volume grows, GPU capacity is scaled proportionally. Online response generator QWen2-72B (Yang et al., 2024) is served via vLLM (Kwon et al., 2023) to support high-throughput inference. Feedback events are streamed through Apache Flink pipelines.

5.2 Main Results

We evaluate DMA in two complementary settings: a multi-month industrial deployment to assess its real-world effectiveness in large-scale online environments, and four public QA benchmarks to verify its retrieval and generation performance under standard static protocols.

Results on Real-World Online Evaluation. As shown in Table 2, Full DMA yields a 24.57% increase in session-level user satisfaction over an online BGE-based reranker baseline. This improvement is statistically significant ($p < 0.001$, two-tailed z-test), based on 100,000 user sessions collected via a randomized controlled trial. For detailed results on the impact of different feedback signals and alignment strategies, see § 5.3.

Table 2: User satisfaction across four evaluation settings. (A) compares DMA against a static baseline (BGE-Reranker). (B) reports the effect of removing individual feedback signals from DMA. (C) analyzes fusion strategies. (D) compares online learning to weekly batch updates.

Configuration	User Satisfaction (%)	Relative Change (%)
(A) Overall Performance		
Zero-Aligned reranker (baseline)	62.11	Reference
Full DMA (ours)	**77.37**	+24.57
(B) Feedback Ablation		
Full DMA (baseline)	77.37	Reference
w/o List-Level Feedback	65.32	–15.57
w/o Response-Level Feedback	68.70	–11.21
w/o Document-Level Feedback	73.29	–5.27
(C) Fusion Strategy		
Cascading Fusion (baseline)	72.79	Reference
Distillation (Full DMA)	**77.34**	+6.25
(D) Online Learning		
Weekly Batch Learning (baseline)	76.21	Reference
Online Learning (Full DMA)	**77.54**	+1.75

Impact of Fusion Strategy. To evaluate the performance of our model fusion strategies at scale, we compare distillation against a cascading approach using the online RCT setup. Table 2 shows that distillation outperforms cascading by +6.25% under similar latency constraints.

Impact of Online Learning. User preferences evolve over time, necessitating continuous model updates. We evaluate the impact of DMA's online learning mechanism, which performs incremental daily retraining and real-time feedback adaptation, compared to a baseline of weekly batch updates. As shown in Table 2, online learning improves session-level satisfaction by +1.75% compared to batch learning, providing qualitative evidence for the value of continuous adaptation.

Table 3: **Results on Public QA Benchmarks Grouped by Task Type.** Left: Conversational QA datasets (open-ended user queries). Right: Structured QA datasets (schema-grounded queries). All methods are evaluated using LLaMA2-7B as the reader model, which serves as the largest publicly available common denominator across prior work to ensure fair and standardized comparison.

Method	Conversational QA Tasks				Structured QA Tasks			
	TriviaQA		HotpotQA		NQ		WebQSP	
	Hit@1	F1	Hit@1	F1	Hit@1	F1	Hit@1	F1
KnowPAT (Zhang et al., 2023)	63.20	65.20	29.00	37.40	51.42	54.82	68.73	65.31
RRHF (Yuan et al., 2023)	62.50	60.20	28.16	35.40	50.11	52.01	66.90	63.10
RAFT (Zhang et al., 2024b)	60.10	57.40	30.20	35.80	50.24	53.86	–	–
FILCO (Wang et al., 2023)	67.30 (2)	67.80 (2)	32.70 (2)	40.80 (2)	**52.71** (1)	**55.32** (1)	**69.96** (1)	**68.34** (1)
DMA (Ours)	**68.81** (1)	**68.90** (1)	**33.92** (1)	**41.88** (1)	51.11 (3)	54.92 (2)	67.26 (3)	65.03 (3)

Results on Public QA Benchmarks. To evaluate DMA under standardized retrieval conditions, we assess its performance on four widely used public datasets: TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), NQ (Kwiatkowski et al., 2019), and WebQSP (Berant et al., 2013). These span open-ended (TriviaQA, HotpotQA) and schema-grounded (NQ, WebQSP) query types, supporting analysis of generalization across formats. We compare against several alignment-optimized RAG baselines, including KnowPAT (Zhang et al., 2023), RRHF (Yuan et al., 2023), RAFT (Zhang et al., 2024b), and FILCO (Wang et al., 2023). This selection balances method comparability (all adopt alignment-based RAG optimization), result availability (publicly reported scores), and experimental fairness (standardized decoding with LLaMA2-7B (Touvron et al., 2023)). As shown in Table 3, DMA achieves the highest Hit@1 and F1 scores on conversational datasets, and remains competitive on structured tasks. These results underscore DMA's advantage in open-ended, user-facing QA scenarios.

5.3 Ablation Studies

We conduct ablation studies to assess the contribution of each feedback granularity in DMA. Table 2 shows that removing *list-level feedback* results in the largest performance drop (–15.57%), followed by *response-level* (–11.21%) and *document-level feedback* (–5.27%). This validates our design choice to integrate multi-granular feedback.

Hierarchical impact of feedback types. These results reveal a natural hierarchy in feedback utility: list-level signals provide coarse but globally informative supervision for document ranking; response-level feedback reflects downstream user preferences across document sets; and document-level labels offer fine-grained, local guidance. Their removal leads to progressively degraded satisfaction, confirming their complementary roles.

Complementarity and alignment. Pointwise (document) signals alone are insufficient for ranking complex lists, while listwise and response-level supervision offer stronger alignment with holistic user intent. This stack of feedback levels enables DMA to optimize both local document quality and global retrieval behavior, especially in dynamic online environments.

Takeaway. Among all components, listwise feedback plays the most critical role in guiding DMA toward globally aligned memory selection. Our multi-granularity design not only enhances overall quality but also ensures adaptability to diverse user preferences in real-world deployments.

6 Limitations

While DMA demonstrates robust performance across both public datasets and industrial deployments, two practical limitations remain when applying the framework to broader scenarios:

Scalability in low-resource or interface-constrained environments. DMA is designed for large-scale, high-throughput production systems where continuous user feedback is available for online adaptation. In low-traffic or offline settings, feedback signals may be too sparse to support timely model updates. DMA relies on multi-level behavioral signals such as document-, list-, and response-level feedback, which are primarily available in interactive dialogue systems. In structured API-style tasks or static document editing scenarios, such fine-grained feedback is either unavailable or hard to instrument, limiting DMA's adaptability. To mitigate this, DMA includes an adaptive retraining scheduler that defers updates under low-feedback conditions, and future work may explore synthetic or proxy signals to fill these gaps.

Generalization to schema-bound QA benchmarks. Although DMA achieves strong results on open-ended, user-facing datasets (e.g., TriviaQA, HotpotQA), its gains are less pronounced on schema-constrained tasks such as NQ and WebQSP. These datasets often feature fixed entity-relation structures or short factual queries that benefit less from multi-granular reranking or feedback-driven adaptation. In such settings, static retrievers and minimal re-ranking may already suffice. This suggests that DMA's dynamic memory alignment is most beneficial in open-ended or conversational environments, and additional strategies—such as symbolic augmentation or knowledge graph integration—may be required to improve performance on interface-like or structured retrieval tasks.

7 Conclusion

We present DMA, an online learning framework that systematically incorporates multi-level human feedback (document, list, and response) to enable real-time retrieval alignment in RAG systems. DMA enables adaptive memory selection guided by user preferences, addressing the rigidity of static retrieval pipelines.

DMA achieves state-of-the-art performance on QA tasks and demonstrates significant gains in a large-scale industrial RCT. Its adaptive scheduling and fusion strategies ensure robustness and efficiency, while ablation studies highlight the importance of feedback granularity in performance gains.

Future work will explore extensions to low-resource domains, alternative feedback modalities, and real-time interpretability for memory selection. Overall, DMA demonstrates that structured user interaction signals can powerfully guide online retrieval learning in deployed GenAI systems.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. Model card and evaluations for claude models. 2023.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2024a.

Asai, A., Zhong, Z., Chen, D., Koh, P. W., Zettlemoyer, L., Hajishirzi, H., and Yih, W.-t. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*, 2024b.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, Y., Miao, Y., Chen, L., Wang, D., Li, D., Ren, Y., Xie, H., Yang, C., and Cai, X. Pistis-rag: Enhancing retrieval-augmented generation with human feedback. *arXiv preprint arXiv:2407.00072*, 2024.

Berant, J., Chou, A., Frostig, R., and Liang, P. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *ICML*. PMLR, 2022.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007.

Chen, W., Chang, M.-W., Schlinger, E., Wang, W., and Cohen, W. W. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*, 2020.

Dong, G., Zhu, Y., Zhang, C., Wang, Z., Dou, Z., and Wen, J.-R. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. *arXiv preprint arXiv:2406.18676*, 2024.

Drozdov, A., Zhuang, H., Dai, Z., Qin, Z., Rahimi, R., Wang, X., Alon, D., Iyyer, M., McCallum, A., Metzler, D., and Hui, K. PaRaDe: Passage ranking using demonstrations with LLMs. In *Findings of EMNLP*, 2023.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Glass, M., Rossiello, G., Chowdhury, M. F. M., Naik, A., Cai, P., and Gliozzo, A. Re2G: Retrieve, rerank, generate. In *NAACL*, 2022a.

Glass, M., Rossiello, G., Chowdhury, M. F. M., Naik, A. R., Cai, P., and Gliozzo, A. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*, 2022b.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *ICML*, 2020.

Izacard, G. and Grave, É. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning. *TMLR*, 2022.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *JMLR*, 24(251):1–43, 2023.

Jeong, S., Baek, J., Cho, S., Hwang, S. J., and Park, J. C. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*, 2024.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Khalifa, M., Logeswaran, L., Lee, M., Lee, H., and Wang, L. Few-shot reranking for multi-hop QA via language model prompting. In *ACL*, 2023.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *TACL*, 2019.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33, 2020.

Lin, S.-C., Asai, A., Li, M., Oguz, B., Lin, J., Mehdad, Y., Yih, W.-t., and Chen, X. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*, 2023.

Lin, X. V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., Zettlemoyer, L., and tau Yih, W. RA-DIT: Retrieval-augmented dual instruction tuning. In *ICLR*, 2024.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Liu, Z., Ping, W., Roy, R., Xu, P., Shoeybi, M., and Catanzaro, B. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*, 2024b.

Luo, H., Chuang, Y.-S., Gong, Y., Zhang, T., Kim, Y., Wu, X., Fox, D., Meng, H., and Glass, J. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*, 2023.

Ma, X., Wang, L., Yang, N., Wei, F., and Lin, J. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*, 2023.

Mitra, B., Craswell, N., et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 2018.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Nogueira, R., Jiang, Z., Pradeep, R., and Lin, J. Document ranking with a pretrained sequence-to-sequence model. In *Findings of EMNLP*, 2020.

OpenAI. GPT-4, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35, 2022.

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., et al. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of NAACL*, 2024.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. In-context retrieval-augmented language models. *TACL*, 2023.

Sachan, D. S., Reddy, S., Hamilton, W. L., Dyer, C., and Yogatama, D. End-to-end training of multi-document reader and retriever for open-domain question answering. In *NeurIPS*, 2021.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., and Chen, W. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of EMNLP*, 2023.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Replug: Retrieval-augmented black-box language models. In *NAACL*, 2024.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *EMNLP*, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team, K., Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., Li, C., Xiao, C., Du, C., Liao, C., et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *ACL*, 2023.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wang, Y., Ren, R., Li, J., Zhao, W. X., Liu, J., and Wen, J.-R. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*, 2024.

Wang, Z., Araki, J., Jiang, Z., Parvez, M. R., and Neubig, G. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xu, F., Shi, W., and Choi, E. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *ICLR*, 2024.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

Ye, F., Fang, M., Li, S., and Yilmaz, E. Enhancing conversational search: Large language model-aided informative query rewriting. In *EMNLP*, pp. 5985–6006, 2023.

Yoran, O., Wolfson, T., Ram, O., and Berant, J. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*, 2024.

Yu, W., Zhang, Z., Liang, Z., Jiang, M., and Sabharwal, A. Improving language models via plug-and-play retrieval feedback, 2024.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears, 2023.

Zhang, L., Yu, Y., Wang, K., and Zhang, C. Arl2: Aligning retrievers for black-box large language models via self-guided adaptive relevance labeling. *arXiv preprint arXiv:2402.13542*, 2024a.

Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., and Gonzalez, J. E. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024b.

Zhang, Y., Chen, Z., Fang, Y., Lu, Y., Li, F., Zhang, W., and Chen, H. Knowledgeable preference alignment for llms in domain-specific question answering. *arXiv preprint arXiv:2311.06503*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Abstract and § 1. The main contributions of DMA, including feedback-driven alignment and large-scale online gains, are clearly stated and consistently validated.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made...

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: § 6 discusses cold-start adaptation, retraining latency, and domain generalization, with mitigation strategies.

   Guidelines:

   - The answer NA means that the paper has no limitation...

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include formal theoretical results; contributions are empirical and algorithmic.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results...

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See § 5 for dataset details, retriever/backbone setup, update frequency, and system implementation.

   Guidelines:

   - The answer NA means that the paper does not include experiments...

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to industrial deployment and privacy policies, code and user data cannot be released.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code...

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See § 5.1 for training batch sizes, model update intervals, retrieval strategy, and compute details.

Guidelines:

- The answer NA means that the paper does not include experiments...

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Online evaluations report p-values ($p < 0.001$, z-test); benchmark results include standard deviation across trials.

Guidelines:

- The answer NA means that the paper does not include experiments...

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See § 5.1. We report usage of 8×A800 GPUs, end-to-end training latency (10 min), and retraining cadence.

Guidelines:

- The answer NA means that the paper does not include experiments...

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: All feedback was collected via de-identified logs in accordance with platform policy; no PII or sensitive data was used.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics...

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: § 6 outlines potential benefits in user-aligned generation and risks of misuse via over-personalization, with mitigation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed...

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

Answer: [Yes]

Justification: While no model is released, the deployed system includes real-time moderation, access control, and privacy filtering (see § 5.1).

Guidelines:

- The answer NA means that the paper poses no such risks...

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All public datasets are cited and used under their respective licenses in § 5.

Guidelines:

- The answer NA means that the paper does not use existing assets...

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: No new datasets or models are released due to industrial constraints and privacy considerations.

Guidelines:

- The answer NA means that the paper does not release new assets...

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Prompt instructions for session-level satisfaction are shown in Table 1; no paid participants involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects...

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The feedback collection process was reviewed and approved by an internal ethics committee; all data were anonymized before use.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects...

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [Yes]

Justification: LLaMA2-7B is used as the unified generator in evaluation, and Qwen-72B is used for annotating session-level feedback in § 5.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs...