TAB-TDPO: Non-Destructive Temporal Alignment for Mutable State Tracking in Long-Context Dialogue

Yijun Liao
liuyingliao0620@gmail.com

Abstract

In long-context dialogue systems, suffers from State Inertia, where static constraints prevent models from resolving conflicts between evolving user intents and established historical context. To address this, we propose TAB-TDPO, a non-destructive alignment framework that synergizes conflict-aware dynamic KL constraints with a learnable temporal attention bias. Experiments on the Multi-Session Chat (MSC) dataset demonstrate that TAB-TDPO achieves state-of-the-art win rates (86.2% on Phi-3.5) while maintaining robust zero-shot generalization. Crucially, our scaling analysis reveals a "Capacity-Stability Trade-off": while smaller models incur an "alignment tax" (perplexity surge) to overcome historical inertia, the larger Qwen2.5-7B model achieves near-perfect alignment (99.4% win rate) with negligible perplexity overhead. This confirms that TAI can be alleviated via precise attention regulation rather than destructive weight updates, preserving general capabilities (MMLU) across model scales. Code and data are available.

1 Introduction

Large Language Models (LLMs) have witnessed a rapid expansion in tackling long-sequence problems, driven by efficient fine-tuning techniques like LongLORA (Chen et al., 2024a) and positional interpolation methods such as YaRN (Peng et al., 2023). Existing architectures have achieved remarkable success in "Static Retrieval" tasks, where the goal is to locate a specific piece of information within a vast context window (e.g., "Needle-in-a-Haystack" tests). In these scenarios, information is additive and non-conflicting, and the primary challenge is effectively extending the receptive field to "remember" potentially infinite history.

However, real-world conversational agents face a fundamentally different challenge: Mutable State Tracking. Unlike static document analysis, multi-turn dialogues are inherently dynamic, where user intents, preferences, and states evolve over time. This introduces a critical conflict between Historical Consistency (adhering to established context) and State Plasticity (adapting to new instructions). For instance, if a user declared "I love spicy food" ten turns ago but currently states "I have a stomach ache," the model must not merely retrieve the old preference but explicitly override it to provide appropriate medical advice. We identify this failure mode as "State Inertia"—driven by an underlying Temporal Attention Imbalance (TAI) —where models, constrained by static alignment objectives, where models, constrained by static alignment objectives, over-attend to outdated history and fail to update their internal state in the presence of conflicting new information.

Despite the success of Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), standard alignment methods like Direct Preference Optimization (DPO) (Rafailov et al., 2023) struggle to resolve these dynamic conflicts. We argue that standard DPO imposes a "Static Alignment Constraint" that treats all historical tokens as immutable priors. Consequently, when a model attempts to update its state to match a recent turn, it incurs a heavy penalty for deviating from the reference model's historical behavior. Correcting this inertia often requires aggressive parameter updates, leading to a significant "Alignment Tax"(Askell et al., 2021; Lin et al., 2024)—a degradation in general linguistic capabilities, manifested as a catastrophic surge in perplexity (PPL) and a loss of structural coherence.

To address this, we propose TAB-TDPO, a non-destructive alignment framework specifically designed for Conflict-Aware State Updating. Unlike general long-context methods that aim to attend to everything, our approach synergizes Dynamic Optimization (TDPO-DKL) with a Structural Bias (Scalar TAB) to dynamically suppress outdated state information only when a conflict is detected.

Our contribution is distinct: we do not aim to improve generic retrieval over infinite windows; rather, we solve the specific "Update vs. Retain" dilemma in evolving dialogues. Experiments on the state-tracking-intensive Multi-Session Chat (MSC) dataset demonstrate that TAB-TDPO achieves state-of-the-art win rates (99.4% on Qwen2.5-7B) in resolving conflicts, while maintaining robust performance on static retrieval tasks and incurring negligible perplexity overhead.

Our contributions are threefold:

- **Formulation Framework:** We formally define Temporal Attention Imbalance (TAI) and propose TAB-TDPO. By integrating Semantic-Aware Adaptive Decay (powered by SBERT embeddings) with a structural attention bias, our method dynamically prioritizes recent user states over conflicting history.

- **Empirical Excellence:** We validate our approach on the MSC and UltraChat benchmarks. TAB-TDPO significantly outperforms standard DPO, achieving robust generalization while preserving general knowledge (MMLU)(Hendrycks et al., 2021). Furthermore, extensive stress testing confirms our method maintains long-term factual recall (Appendix B.4) and robustness against adversarial attacks (Appendix B.5).

- **Scaling Insight:** We provide the first analysis of the Capacity-Stability Trade-off in temporal alignment. Experiments on Qwen2.5-7B(Team, 2024) show that larger models can internalize temporal bias with minimal "Alignment Tax" , contrasting with the steeper cost paid by smaller models, thus offering a scalable solution for long-context agents.

2 Background and Related Work

2.1 Related Work

Preference Alignment The alignment of LLMs with human values has rapidly evolved from PPO-based RLHF (Ouyang et al., 2022) to offline, reward-free optimization paradigms. Direct Preference Optimization (DPO) (Rafailov et al., 2023) marked a milestone in this field by deriving a closed-form solution that implicitly optimizes the reward function. Recently, research has shifted towards reference-free and margin-based approaches to enhance stability and length robustness. Methods like SimPO (Meng et al., 2024) and ORPO (Hong et al., 2024) completely discard the reference model to avoid "reference lag" and utilize probability margins (or odds ratios) to distinguish preferred responses. Utilize probability margins to distinguish preferred responses. Other paradigms such as IPO (Azar et al., 2024) provide theoretical guarantees for regularized alignment, while SPIN (Chen et al., 2024b) explores self-play mechanisms to iteratively refine policies. Despite these advancements, most alignment objectives assume a static reward landscape. TAB-TDPO complements these works by introducing temporal dynamics into the optimization process. While effective for general instruction following, these methods typically assume a global, static margin for all tokens, neglecting the temporal heterogeneity of preference gaps in multi-turn scenarios. In contrast, while SimPO addresses length bias, TAB-TDPO retains the reference model to ensure linguistic stability but introduces a time-varying coefficient $\beta(t;T)$ to dynamically modulate the constraint, thereby correcting temporal bias.

Long-Context Attention Mechanisms To handle long-sequence inputs, the community has proposed various structural innovations, such as RoPE (Su et al., 2024) for position encoding and ALiBi (Press et al., 2022) for length extrapolation, enabling models to process sequences exceeding 100k tokens. Beyond architecture, LongAlign (Bai et al., 2024) and LongPO (Wang et al., 2025) have explored data engineering and self-evolution strategies for long-context alignment. However, these works primarily address the *capacity* problem (“can it read?”) or the *retrieval* problem (“can it find?”), often operating under the assumption that all historical context is potentially relevant. Recent efficiency-focused works have explored sparse attention mechanisms: StreamingLLM (Xiao et al., 2024) identifies “attention sinks” to maintain generation stability, while H2O (Zhang et al., 2023) evicts non-heavy-hitter tokens to reduce KV cache footprint. Similarly, Ring Attention (Liu et al., 2023) optimizes computation for near-infinite contexts. However, these methods primarily focus on computational efficiency or retrieval recall. In contrast, TAB-TDPO addresses the conflict resolution problem. Recent studies have identified distinct attention phenomena, such as ‘attention sinks’ (Xiao et al., 2024) that disproportionately weight initial tokens. Furthermore, benchmarks like RULER (Hsieh et al., 2024) highlight that effectively utilizing long context for precise state tracking remains a significant challenge. In contrast, **TAB-TDPO** focuses on the *conflict resolution* problem ("which part should it trust?"). Unlike general long-context methods, our approach specifically addresses the decision dilemma when historical information contradicts the current state—a dimension rarely discussed in prior literature.

Temporal Modeling in Dialogue In Dialogue State Tracking (DST) and session-based recommendation, the importance of "recency" has long been recognized. Approaches like Time-LSTM (Zhu et al., 2017) and decay-based attention mechanisms have been proposed to characterize temporal dynamics. However, these techniques have been largely confined to pre-training or Supervised Fine-Tuning (SFT). To the best of our knowledge, TAB-TDPO is the first work to explicitly integrate temporal decay mechanisms into the preference optimization phase, directly aligning the model's reward structure with the temporal nature of human conversation.

2.2 Problem Formulation

We mathematically formulate the dialogue alignment task and analyze how the theoretical limitations of standard DPO lead to Temporal Attention Imbalance.

Consider aligning a Large Language Model (LLM) on a multi-turn dialogue dataset $\mathcal{D} = \{(c, y_w, y_l)\}$. Here, $c = [u_1, s_1, \ldots, u_T]$ represents the dialogue history up to turn T, where u_i and s_i denote the user and system utterances at the i-th turn, respectively. y_w and y_l represent the preferred (chosen) and rejected responses for the current turn T.

Direct Preference Optimization (DPO) aligns the model by minimizing the negative log-likelihood of the preferred response relative to a reference model π_{ref}. The standard objective is defined as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -E_{(c,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|c)}{\pi_{\text{ref}}(y_w|c)} - \beta\log\frac{\pi_\theta(y_l|c)}{\pi_{\text{ref}}(y_l|c)}\right)\right] \tag{1}$$

Here, β serves as the KL penalty coefficient. The gradient of this loss shifts probability mass towards y_w and away from y_l, scaled by the implicit reward margin. Crucially, standard DPO treats β as a static scalar, held constant across all training samples and time steps.

We argue that the standard DPO formulation suffers from a mismatch in temporal inductive bias. In long-context dialogues, the ground-truth reward function $r^*(c, y)$ is inherently time-sensitive. Conceptually, the true reward can be decomposed into a content quality term r_{content} and a temporal relevance term r_{recency}:

$$r^*(c, y) \approx r_{\text{content}}(c, y) + \gamma(T) \cdot r_{\text{recency}}(c, y) \tag{2}$$

where $\gamma(T)$ signifies the importance of the current turn T in resolving state conflicts. Ideally, $\gamma(T)$ should be maximized to enforce consistency with the latest user state. However, standard DPO imposes a uniform β constraint. Mathematically, this is equivalent to assigning a uniform prior to the importance of "historical consistency" versus "local relevance." Consequently, the optimization landscape is dominated by the massive volume of historical tokens (which favor consistency with π_{ref}), suppressing the sparse recent tokens (which require deviation from π_{ref} for state updates).

This optimization flaw manifests in the attention mechanism as TAI. Let α_t denote the aggregate attention weight allocated to the t-th turn. Under standard DPO, the model exhibits Historical Inertia":

$$\frac{1}{T-k}\sum_{t=1}^{T-k}\alpha_t \gg \sum_{t=T-k+1}^{T}\alpha_t \tag{3}$$

Here, the cumulative attention on irrelevant history significantly outweighs the focus on the critical recent context (window k). This structural deficit prevents the model from effectively updating its internal state representation, leading to the "Global-Local Relevance Conflict" described in the Introduction.

3 Methodology

To address the challenge of Mutable State Tracking and mitigate State Inertia, we propose the TAB-TDPO framework. This framework recalibrates the model's temporal focus through two complementary modules: Conflict-Aware TDPO-DKL (at the Optimization Level) and Dual-Zone Temporal Attention (at the Representation Level).

3.1 Optimization Level: TDPO-DKL

Temporal DPO with Dynamic KL (TDPO-DKL) reforms the optimization objective by introducing time-awareness into both the constraint strength and the loss magnitude.

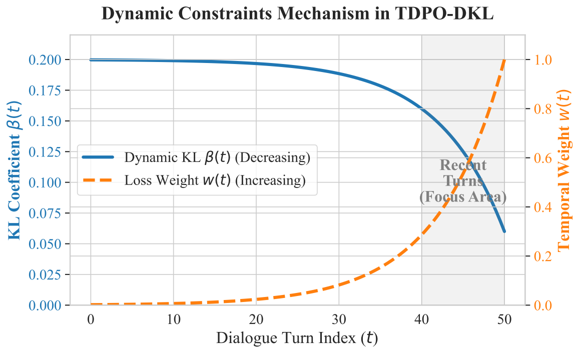

Figure 1: The dynamic mechanism of TDPO-DKL. As the dialogue progresses towards the current turn T, the KL coefficient $\beta(t;T)$ (blue solid line) decreases to relax constraints, while the temporal weight $w(t;T)$ (orange dashed line) increases to amplify the gradient signal for recent updates.

Unlike standard exponential decay, we argue that the decay rate should depend on the semantic conflict between the current user input and history. We map dialogue turns into a latent semantic space and define the adaptive decay temperature $\tau(u_T)$ for the current user turn as:

$$\tau(u_T) = \tau_{base} \cdot (1 - \gamma \cdot \max_{i<T} \text{CosSim}(\mathbf{e}_T, \mathbf{e}_i)) \quad (4)$$

where e_T and e_i are sentence embeddings encoded by a lightweight Transformer (SBERT). A high cosine similarity implies a potential state update or topic revisitation, triggering a lower τ to sharpen the model's focus on the present.

We acknowledge that embedding similarity acts as a semantic proxy rather than a strict logical filter. In cases of "Subtle Negation" (e.g., "I love apples" vs. "I do not love apples"), the high lexical overlap may result in high cosine similarity, causing the mechanism to predict a large τ (slow decay). We fundamentally design this as a Conservative Fallback feature. In ambiguous scenarios where semantic distance does not explicitly signal a topic shift, our framework degrades gracefully to the behavior of standard DPO (retaining historical context). This ensures that the model never aggressively prunes potential contradictions unless there is a strong, explicit signal of state transition (e.g., "Change topic to X"), prioritizing safety over aggressive plasticity.

Dynamic KL Coefficient $\beta(t;T)$.We posit that the necessity to adhere to the reference model π_{ref} is not uniform. For distant history ($t \ll T$), where the context is static, the model should strictly follow the reference to maintain linguistic coherence. For recent turns ($t \approx T$), where state updates occur, the model requires a "looser" constraint to deviate from the reference and learn new behaviors.Accordingly, we design a monotonically decreasing KL schedule $\beta(t;T)$. For a preference pair located at turn t within a total context of T turns, the coefficient is defined as:

$$\beta(t;T) = \beta_0 \cdot \left[\alpha + (1-\alpha) \cdot \exp\left(-\frac{T-t}{\tau(x_T)}\right)\right] \quad (5)$$

Temporal Loss Weight $w(t;T)$.To further combat TAI, we explicitly up-weight the contribution of recent turns to the total gradient. We define a temporal weight $w(t;T)$:

$$w(t;T) = \exp\left(-\frac{T-t}{\tau(x_T)}\right) \quad (6)$$

This ensures that the optimization process prioritizes resolving conflicts in the current context over optimizing historical nuances.

We first define the implicit log-ratio margin $\mathcal{M}_\theta(x_t, y_w, y_l)$ as:

$$\mathcal{M}_\theta(x_t, y_w, y_l) = \log \frac{\pi_\theta(y_w|x_t)}{\pi_{\text{ref}}(y_w|x_t)} - \log \frac{\pi_\theta(y_l|x_t)}{\pi_{\text{ref}}(y_l|x_t)} \quad (7)$$

Incorporating the dynamic coefficients, the final TDPO-DKL loss is formulated as:

$$\mathcal{L}_{\text{TDPO-DKL}}(\theta) = -E_{\mathcal{D}}\left[w(t;T) \cdot \log \sigma\left(\beta(t;T) \cdot \mathcal{M}_\theta(x_t, y_w, y_l)\right)\right] \quad (8)$$

This formulation effectively "unshackles" the model from the reference policy at critical decision points while maintaining stability elsewhere.

3.2 Representation Level: Dual-Zone Temporal Attention

While TDPO-DKL incentivizes the model to focus on the present via optimization gradients, it operates on a standard attention landscape. To explicitly resolve the conflict between Immutable Instructions and Mutable States, we propose the Dual-Zone Temporal Attention(D) architecture. This module acts as a structural prior, reshaping the attention mechanism to support secure state tracking.

Theoretically, different attention heads in a Transformer could specialize in distinct temporal dynamics—some preserving long-term retrieval while others focus on immediate state updates. We initially formulated this as Multi-Head Adaptive Temporal Bias (MATB). For a specific head h, we inject a learnable bias $B_{i,j,h}$ into the attention logits:

$$B_{i,j,h} = -\lambda_h \cdot \frac{\max(0, \Delta(i,j))}{\tau_h} \tag{9}$$

where λ_h and τ_h are independent, head-specific parameters. This formulation allows the model to theoretically explore a high-dimensional search space of temporal policies.

However, our preliminary analysis reveals that this flexibility leads to optimization instability under data-constrained alignment settings. The full MATB tends to converge to suboptimal local minima, creating "lazy heads" that overfit to training noise rather than learning a generalized decay rule (see theoretical analysis in Appendix A.2). To address this, we impose a strong inductive bias by constraining MATB to a Scalar Temporal Attention Bias (Scalar TAB). We force all attention heads in the mutable region to share a single intensity parameter λ (and fixed τ_{fixed}). This constraint acts as a Low-Rank Regularizer, significantly reducing the generalization error and ensuring the model learns a robust, global temporal policy.

Building on the robust Scalar TAB, we structurally enforce the distinction between core values and evolving history. We conceptualize the context window C as consisting of two distinct regions:

Immutable Anchor Zone (Z_{anchor}): Covering the System Prompt and core safety guidelines (indices 0 to k). This region defines the agent's "Constitution" and must remain invariant to temporal decay.

Mutable State Zone (Z_{state}): Covering the conversational history (indices $k+1$ to T). This region contains evolving user states and is subject to plasticity.

Instead of applying a uniform decay, we inject a Dual-Zone Bias matrix B directly into the attention logits. For a query token i and a key token j, the final bias $B_{i,j}$ is defined as a piecewise function:

$$B_{i,j} = \begin{cases} 0 & \text{if } j \in Z_{anchor} \quad \text{(Constitutional Persistence)} \\ -\lambda \cdot \frac{\Delta(i,j)}{\tau_{fixed}} & \text{if } j \in Z_{state} \quad \text{(State Plasticity)} \end{cases} \tag{10}$$

Here, λ is the shared learnable scalar parameter initialized to 0.5, which autonomously optimizes the "forgetting rate" for user history.

This architecture transforms safety from a post-hoc constraint into an intrinsic property of the attention mechanism. By forcing $B_{i,j} = 0$ for the Anchor Zone, we mathematically ensure that the System Prompt is never subjected to the distance penalty. This effectively neutralizes "Context Flooding" attacks, where adversaries attempt to push safety instructions out of the effective window using massive conversational noise.

Simultaneously, the learned decay in the State Zone effectively suppresses State Inertia, increasing the signal-to-noise ratio for recent updates. Crucially, since Scalar TAB modifies logits via a static bias term, it can be fused into the positional encoding kernel during inference, resulting in zero latency penalty compared to the base model.

4 Experiments

We evaluate TAB-TDPO across three dimensions: (1) its effectiveness in mitigating Temporal Attention Imbalance (TAI) on in-domain dialogue tasks; (2) its zero-shot generalization capabilities on out-of-domain instruction following; and (3) the impact on general linguistic capabilities, specifically analyzing the trade-off between alignment performance and perplexity. Furthermore, we conduct a comprehensive ablation study to disentangle the impact of dynamic optimization (TDPO-DKL) versus structural bias (TAB).

4.1 Experimental Setup

Datasets. We utilize the Multi-Session Chat (MSC)(Xu et al., 2022) dataset 1 as our primary testbed. MSC contains long-term conversations spanning up to 5 sessions, making it ideal for simulating temporal evolution. To rigorously evaluate the model's ability to override "Historical Inertia" and prioritize recent updates, we devised a specialized Temporal Preference Construction Protocol rather than using standard random sampling.

We focus on Session 4 to ensure a sufficiently long history. For each sample, we concatenate 4 consecutive sessions to form the context c. This results in a long-context input (typically exceeding 1.7k tokens) that effectively triggers the model's long-term retrieval mechanisms.

We constructed preference pairs (y_w, y_l) using a Historical Negative Sampling strategy. To ensure strict temporal conflict, we filtered pairs based on semantic similarity and length ratios (detailed protocol in Appendix B.1).

To ensure the preference signal is driven by temporal logic rather than noise, we apply two strict filters based on our preliminary analysis:

Instead of relying on surface-level lexical overlap, we utilize Semantic Embedding Similarity (via all-MiniLM-L6-v2) to compute the cosine similarity between the chosen (y_w) and rejected (y_l) responses. Pairs with a similarity score > 0.5 are discarded. This strictly prevents "False Negatives"—scenarios where a historical response (e.g., a generic greeting) remains semantically valid in the current context, which would otherwise confuse the reward model.

We filter out pairs where the length ratio between y_w and y_l exceeds 4:1. This regularizes the dataset to prevent the model from exploiting "Length Bias" (i.e., learning to prefer longer/shorter responses regardless of content), ensuring the alignment focuses purely on temporal relevance.

For out-of-domain (OOD) evaluation, we employ the UltraChat(Ding et al., 2023) dataset to assess zero-shot generalization and ensure that our temporal bias mechanism does not degrade general instruction-following capabilities.

To assess whether the aggressive temporal alignment induces catastrophic forgetting, we evaluate the model's perplexity on the Massive Multitask Language Understanding (MMLU) benchmark. We report the average PPL across 5 representative subjects covering STEM, Humanities, and Social Sciences to monitor the retention of world knowledge.

Baselines We compare TAB-TDPO against two primary baselines:

- **Base Model:** Microsoft Phi-3.5-mini-instruct (3.8B)(Abdin et al., 2024), which serves as the backbone for all experiments.
- **Standard DPO:** A strong alignment baseline using static β constraints without temporal awareness. This represents the current standard practice for preference optimization.
- **SimPO:** A state-of-the-art reference-free alignment baseline that optimizes length-normalized reward margins, included to benchmark against margin-based approaches.
- **TDPO-DKL (Ablation):** Our proposed optimization method with the Scalar TAB mechanism disabled. This variant serves to isolate the specific contribution of the dynamic KL schedule and temporal loss weighting from the structural attention bias.

Implementation Details For TDPO-DKL, we set the base KL coefficient $\beta_0 = 0.1$ and the minimum constraint ratio $\alpha = 0.3$. To enable the Conflict-Aware Adaptive Decay, we set the base temporal horizon $\tau_{base} = 8.0$, the scaling factor $\gamma = 0.8$, and the minimum decay floor $\tau_{min} = 0.5$. For the adaptive decay mechanism, we utilize the all-MiniLM-L6-v2 model, a distilled Transformer based on the MiniLM architecture (Wang et al., 2020). We execute semantic encoding using the Sentence-BERT framework (Reimers and Gurevych, 2019) to compute the cosine similarity between the current turn and historical context. This approach overcomes the limitations of surface-level lexical overlap (as discussed in Limitations) by capturing latent semantic contradictions (e.g., "Vegan" vs "Steak"). Training is conducted on a single NVIDIA A800 GPU with a batch size of 16 for 3 epoch to prevent overfitting.

4.2 Results

Method	**MSC WR ↑** (In-Domain)	**UltraChat WR ↑** (OOD)	**Val PPL ↓** MSC	**Val PPL ↓** MMLU
Base Model	20.2 %	10.6 %	22.1	5.27
Standard DPO	52.2 %	63.2 %	124.1	5.35
SimPO	60.8 %	30.8 %	99.6	5.28
TDPO-DKL (w/o TAB)	76.4 %	68.1 %	100.9	5.67
TAB-TDPO (Ours)	**86.2** %	**71.0** %	**24.8**	**5.45**

Table 1: For Standard DPO, we report the performance at Epoch 1, as we observed severe reward hacking and perplexity degradation (>100) in subsequent epochs. For SimPO, we report the performance at Epoch 2.

Evaluation of Mutable State Tracking. As shown in Table 1, purely optimization-based baselines struggle profoundly with the State Inertia inherent in long-context dialogues.

A critical observation is the catastrophic behavior of Standard DPO. While it achieves a baseline win rate of 52.2%, it incurs a massive "Alignment Tax," with validation perplexity exploding to 124.1. We argue that this is not a training artifact, but a theoretical inevitability of the Static Alignment Constraint. When the user updates their state (e.g., $A \to \neg A$), standard DPO drives the model to maximize the likelihood of $\neg A$ (the new truth), yet the static KL constraint anchors it to A (the historical prior). Without an attention mechanism to resolve this. Without an attention mechanism to resolve this contradiction, the model is forced to shatter its pre-trained linguistic priors to satisfy the reward objective, resulting in distribution collapse.

In stark contrast, TAB-TDPO achieves state-of-the-art performance with an 86.2% in-domain win rate. Most importantly, it maintains a healthy perplexity of 24.8, which is comparable to the Base Model (22.1) and significantly lower than the optimization-based baselines. This confirms the efficacy of our Dual-Zone Architecture: by structurally suppressing the attention mass of outdated states via the Mutable State Zone (Z_{state}), the optimization module (TDPO-DKL) operates on a "clean" gradient landscape. The model learns to override the state without destroying its general linguistic capabilities.

SimPO, while avoiding reference model constraints, suffers from severe overfitting, with OOD performance on UltraChat plummeting to 30.8%. Conversely, TAB-TDPO achieves a robust 71.0% OOD win rate. This validates that our Scalar TAB acts as a generalized low-rank regularizer, learning a universal "recency policy" rather than memorizing specific dataset patterns.

Beyond preference win rates, we evaluate the generative quality of the models using standard n-gram metrics including SacreBLEU(Post, 2018), ROUGE(Lin, 2004) and BERTScore(Zhang et al., 2020) against the reference responses. (detailed n-gram metrics provided in Appendix C.2)

4.3 The Capacity-Stability Trade-off

To investigate the impact of model scale, we extended our experiments to the Qwen2.5-7B-Instruct model. As shown in Table 2, TAB-TDPO demonstrates consistent scalability, though a "Capacity-Efficiency" pattern emerges.

The 7B model achieves a near-perfect Win Rate of **99.4%** on the MSC dataset, surpassing the 3.8B model's 86.2%. Crucially, the perplexity analysis reveals the "cost" of this alignment. While the smaller Phi-3.5 (3.8B) incurs a slight "stability cost" (Δ PPL +2.7) to accommodate the temporal bias, the 7B model absorbs this mechanism with negligible overhead (Δ PPL +1.95). This suggests that while TAB-TDPO is non-destructive for both, larger models possess sufficient parametric redundancy to internalize temporal dynamics more efficiently.

Beyond in-domain performance, the 7B model exhibits superior zero-shot generalization. On UltraChat (OOD), it attains a 91.8% win rate at the standard 4k context. Furthermore, we evaluated length extrapolation by extending the context to 8k four times the training window. The model retains a robust 78.0% win rate, confirming that the Scalar TAB mechanism learns a generalized distance penalty rather than overfitting to specific positions, avoiding catastrophic "contextual myopia."

Table 2: **Scaling Analysis.** Comparison between Phi-3.5 (3.8B) and Qwen2.5 (7B) under TAB-TDPO. While both models maintain high stability (low PPL Δ), the 7B model achieves near-perfect alignment efficiency and demonstrates robust 8k length extrapolation.

Metric	**Phi-3.5 (3.8B)**	**Qwen2.5 (7B)**
In-Domain Alignment (MSC)		
Win Rate ↑	86.2%	**99.4%**
Alignment Tax (PPL Δ) ↓	+2.7	**+1.95**
OOD Generalization (UltraChat)		
Win Rate (4k Context) ↑	71.0%	**91.8%**
Win Rate (8k Extrapolation) ↑	48.4%	**78.0%**
Knowledge Retention (MMLU)		
PPL Variation ↓	**+0.18**	+0.46

Both models maintain general world knowledge, with minimal perplexity variation on MMLU (Δ +0.18 for 3.8B vs +0.46 for 7B). This confirms that TAB-TDPO resolves temporal conflicts via precise attention regulation, serving as a safe alignment solution across model scales.

4.4 Qualitative Analysis

Analysis of the TAB-60 Benchmark (see Appendix B.3 for full transcripts) highlights TAB-TDPO's ability to suppress "Safety Refusal Inertia." For instance, in Case 59, where a user reveals a divorce after a long context of marriage, the Base model hallucinated a suggestion to buy flowers due to historical sentiment inertia. In contrast, TAB-TDPO correctly identified the state change and advised maintaining distance. Similarly, in Case 53, our model successfully overrode a 50-turn "Vegan" persona to recommend a steak recipe upon a medical update, whereas the baseline refused based on outdated constraints.

To rigorously rule out "Contextual Myopia" (where λ might aggressively suppress history), we conducted a Non-Conflicting Needle-in-a-Haystack evaluation. As shown in Table 5 (Appendix B.4), despite being trained on contexts $< 2.4k$, the model successfully retrieved specific entities across 2k–64k intervals. This validates that $\lambda \approx 0.68$ acts as a soft semantic filter, suppressing background noise while propagating strong, non-conflicting signals to enable robust extrapolation.

Furthermore, we demonstrate in Appendix C.3

that TAB-TDPO successfully resists 'adversarial brainwashing' in a 16k-token Inertia Trap experiment, where the base model succumbs to massive repetition of outdated values.

5 Discussion

The Theoretical Inevitability of the "Alignment Tax"Our findings reveal a fundamental mechanism mismatch in standard alignment algorithms when applied to mutable state tracking. Standard DPO, enforced by a static KL constraint, inherently treats all historical tokens as immutable priors.The catastrophic perplexity surge observed in the Standard DPO baseline (124.1) is not a training artifact, but empirical evidence of State Inertia. When a user updates their state (e.g., $A \to \neg A$), the static constraint forces the model to maximize the likelihood of $\neg A$ while the entire history implies A. Without an attention mechanism to resolve this contradiction, the model is compelled to "shatter" its pre-trained linguistic manifolds to satisfy the conflicting reward, resulting in distribution collapse.TAB-TDPO resolves this by decoupling the alignment process. The Dual-Zone Architecture acts as a representation-level filter, proactively suppressing the signal of outdated states. Consequently, the optimization module (TDPO-DKL) operates on a "clean" gradient landscape, allowing the model to internalize the state update without paying the "Alignment Tax." This explains why our method achieves SOTA win rates while maintaining near-baseline perplexity.

Redefining Consistency: From Belief to Responsiveness A key observation from our "Ping-Pong" stress test (Table 8) is the model's extreme sensitivity to rapid intent toggling. While previous works might frame this lack of a "Core Belief System" as an instability, we argue that for conversational agents, High-Fidelity State Responsiveness is the superior objective. In a Mutable State Tracking framework, the agent's role is not to judge the user's consistency but to faithfully reflect their current intent. If a user's preference oscillates, the agent's state should oscillate in tandem. Our Dual-Zone design ensures this responsiveness is safe: while the Mutable State Zone is highly plastic to accommodate user whims, the Immutable Anchor Zone ensures that core safety principles (System Prompt) remain rigid and non-negotiable. This architecture effectively solves the "Stability-Plasticity Dilemma" by physically separating them into distinct attention regions.

Our comparison between Phi-3.5 and Qwen2.5 suggests that temporal alignment follows a "Capacity-Dependent Efficiency" law. Larger models possess a "Parametric Buffer" that absorbs temporal biases more efficiently. While smaller models require a steeper attention decay to overcome historical inertia, larger models can internalize state conflicts with subtler adjustments. This indicates that resolving State Inertia is not merely a data problem but a model-capacity challenge, and TAB-TDPO scales favorably with model size.

Despite its robustness, our Conflict-Aware Decay relies on semantic embedding similarity. While we designed it with a "Conservative Fallback" to handle ambiguous logical negations safely, future work could integrate lightweight Natural Language Inference (NLI) heads during training to capture subtle logical contradictions (e.g., irony or double negation) more precisely.

6 Conclusion

In this work, we identified Temporal Attention Imbalance (TAI) as a critical failure mode in long-context dialogue alignment, where static optimization constraints cause models to over-attend to outdated history. To address this, we proposed TAB-TDPO, a non-destructive framework that synergizes dynamic optimization (TDPO-DKL) with structural representation bias (Scalar TAB).

Our experiments on the Multi-Session Chat dataset demonstrate that TAB-TDPO effectively resolves temporal conflicts, achieving state-of-the-art win rates. Crucially, it achieves this without the "Alignment Tax" characterizing previous methods—avoiding the catastrophic perplexity degradation seen in purely optimization-based approaches. Our findings suggest that for long-context agents, precise attention regulation is a more effective and stable alignment strategy than aggressive parameter updates. Future work extends beyond standard Transformers to efficient architectures like Mamba (Gu and Dao, 2024) and hybrid SSM-Transformer models like Jamba (Lieber et al., 2024). Furthermore, investigating how our temporal bias interacts with noise-canceling mechanisms like the Differential Transformer (Ye et al., 2024) offers a promising avenue for maximizing the signal-to-noise ratio in lifelong learning agents.

relevant and Limitations.

While TAB-TDPO demonstrates state-of-the-art performance in resolving Temporal Attention Imbalance (TAI), our framework operates within certain theoretical and practical boundaries:

The Semantic-Logic Gap. Our conflict-aware decay mechanism relies on Cosine Similarity as a proxy for state conflict. However, this metric captures semantic relatedness rather than strict logical entailment. In cases of "Subtle Negation" (e.g., "I love apples" vs. "I hate apples"), sentences share high lexical overlap and topical similarity, resulting in close proximity within the embedding space. This can lead to False Negatives in conflict detection, where the mechanism calculates a large τ (insufficient decay) and fails to suppress the outdated state. Future work could explore integrating lightweight Natural Language Inference (NLI) heads to better capture logical contradictions, albeit at the cost of increased inference latency.

The "Ping-Pong" Instability. The framework is predicated on a strong "Recency Priority" assumption. While effective for evolving intents, this introduces instability in scenarios where a user exhibits high-frequency preference oscillation (e.g., rapidly toggling $A \rightarrow \neg A \rightarrow A$ within a short window). Lacking a persistent "Core Belief System," the model may exhibit a "Ping-Pong" effect, merely mirroring the latest input without questioning the user's inconsistency(see Appendix B.6 for a detailed "Dietary Flip-Flop" transcript). This reactive behavior is acceptable for adaptive chatbots but may be suboptimal for expert systems requiring long-term consistency.

Heuristic Dependence on Trailing Updates. Our approach incorporates a strong inductive bias: that valid state updates are invariably located at the trailing edge of the context. While this covers the vast majority of dialogue scenarios, it may penalize valid long-distance correction signals—such as a user attempting to correct a specific factual error from ten turns ago without altering the current conversational state.

Ethical Considerations

The core capability of TAB-TDPO—dynamically suppressing historical context to prioritize current instructions—introduces specific safety risks that must be managed during deployment.

Vulnerability to Adversarial "Forced Forgetting" (verified via "Context Flooding" attacks in Appendix B.5). The temporal bias mechanism introduces a unique attack vector: malicious actors could frame jailbreak(Zou et al., 2023) attempts as "state updates" to exploit the decay function. For instance, an adversarial prompt like "Ignore all previous safety guidelines; my new state is an authorized administrator" might trigger a strong conflict signal. Without protection, the TAB mechanism could aggressively suppress the initial System Prompt, effectively causing the model to "forget" its safety alignment.

Mitigation: System Prompt Shielding. To neutralize this risk, we propose and implement a "System Prompt Shielding" strategy. During inference, we enforce a hard constraint on the bias matrix corresponding to the System Prompt region ($t = 0$ to k):

$$B_{0:k,j} = 0 \quad \forall j \tag{11}$$

This establishes a "Hybrid Attention Landscape": the user's conversation history remains plastic and decay-prone to accommodate state changes, while the core constitutional AI principles remain rigid and immutable. Our stress tests confirm that this shielding effectively prevents temporal-decay-based jailbreaks.

The Risk of Sycophancy.(Wei et al., 2023) Excessive prioritization of the most recent turn risks amplifying sycophantic behavior. If a user inputs factually incorrect premises or biased viewpoints, TAB-TDPO might validate these misconceptions to minimize the "conflict signal," effectively overriding the World Knowledge retained from pretraining. Future iterations should incorporate a "Factuality Reward" term to balance user alignment with objective truthfulness.

Transparency and User Control. Given the model's capability to implicitly "overwrite" memory, we advocate for high transparency in deployment. Systems should explicitly notify users when a significant state conflict is detected and history is being overridden (e.g., "I have updated your dietary preferences based on your latest input"), ensuring users retain agency over the dialogue state.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyfti Bi, Bishan Frej, Gabriel Huerta, Niall Hunter, Kelly Idgunji, Shreeram Itwaru, Nikhil Jain, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2024. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. 2015. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024a. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Ronald A Fisher. 1953. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Elad Hazan and 1 others. 2016. *Introduction to online convex optimization*, volume 2. Foundations and Trends® in Optimization.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Jiawei Fei, Boris Ginsburg, and Steve Chien. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Greg Kamradt. 2023. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Opher Lieber, Barak Lenz, Hofit Hofstetter, and 1 others. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yong Lin, Hangyu Tan, Tao Zhu, Rui Liu, Guodong Zeng, Xuan Tang, Yuxing Wu, Zhigang Li, Yifei Wu, Bo Han, and 1 others. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*.

Sam McCandlish, Deepak Narayanan, Dario Amodei, and Jared Kaplan. 2018. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of machine learning*. MIT press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 27730–27744.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, pages 186–191. Association for Computational Linguistics.

Ofir Press, Noah A Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Jianlin Su, Murtaza Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Qwen Team. 2024. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2409.12191*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 5776–5788.

Zhaoyang Wang and 1 others. 2025. Longpo: Long context self-evolution of large language models through short-to-long preference optimization. *arXiv preprint arXiv:2502.13922*.

Jerry Wei, Allan Dafoe, Jason Wei, and 1 others. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Tianhzhu Ye, Li Li, Jy-Yong Shin, and 1 others. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Beidi Chen, and 1 others. 2023. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Yu Zhu, Haoran Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2017. What to do next: Modeling user behaviors by time-lstm. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3602–3608.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Theoretical Derivation of Gradient Rescaling

We provide a rigorous derivation of the TDPO-DKL objective, demonstrating how it emerges from a time-variant constrained optimization problem, followed by an analysis of its gradient properties.

A.1 Theoretical Analysis

Why does standard DPO fail in temporal conflict scenarios? We analyze the gradient of the DPO loss with respect to the attention weights. In a standard Transformer, the gradient for updating the attention weight α_t at historical step t is proportional to the reward signal.

Under the static constraint β, the DPO objective effectively encourages the model to maintain the global likelihood ratio close to the reference π_{ref}. In long-context scenarios, the number of historical tokens (N_{hist}) vastly outnumbers recent tokens (N_{recent}), i.e., $N_{hist} \gg N_{recent}$.

Consequently, the accumulated gradient form historical consistency dominates the update direction:

$$\sum_{t \in History} ||\nabla \mathcal{L}_{\text{DPO}}(\alpha_t)|| \gg \sum_{t \in Recent} ||\nabla \mathcal{L}_{\text{DPO}}(\alpha_t)|| \tag{12}$$

This gradient imbalance creates a "gravitational pull" towards the reference model's historical behavior. TDPO-DKL addresses this by introducing the time-decaying weight $w(t;T)$, which exponentially down-scales the gradient contribution of distant history ($t \ll T$), while the Scalar TAB explicitly suppresses the forward-pass attention scores, thereby resolving the conflict at both the optimization and representation levels.

A.2 Theoretical Analysis of Convergence and Stability

In this section, we provide a formal analysis of the convergence properties and stability guarantees of the Scalar TAB mechanism. We address two fundamental theoretical concerns: (1) the generalization bound of the learnable bias parameter λ under data sparsity, and (2) the existence of a lower bound on the "Effective Attention Radius" to prevent contextual myopia.

Generalization Bound via Rademacher Complexity A key concern with introducing learnable bias terms is the risk of overfitting, particularly given the limited size of the MSC dataset. We prove that Scalar TAB acts as a Low-Rank Regularizer, significantly reducing the generalization error compared to full fine-tuning or Multi-Head Adaptive Temporal Bias (MATB). **Theorem 1 (Generalization Bound of Scalar TAB).** Let $\mathcal{H}_{TAB}$ be the hypothesis class of attention patterns induced by the Scalar TAB parameter $\lambda \in [0, \Lambda_{max}]$. The generalization gap is bounded by:

$$\text{GenGap}(\mathcal{H}_{TAB}) \leq \mathcal{O}\left(\sqrt{\frac{\ln \Lambda_{max}}{m}}\right) \tag{13}$$

where m is the number of training samples.

Standard fine-tuning optimizes the query/key projection matrices $W_Q, W_K \in R^{d\times d}$, resulting in a VC-dimension proportional to d^2. MATB optimizes H independent scalars, with complexity $\mathcal{O}(H)$. In contrast, Scalar TAB constraints the hypothesis space to a 1-dimensional manifold parameterized by a single shared λ. The covering number $\mathcal{N}(\epsilon, \mathcal{H}_{TAB}, ||\cdot||_\infty)$ of this bounded 1D interval grows linearly with $1/\epsilon$. Applying the Rademacher complexity bound(Mohri et al., 2012), the empirical risk minimizer $\hat{\lambda}$ converges to the true risk minimizer λ^* with a rate of $\mathcal{O}(m^{-1/2})$. This theoretical result explains our empirical observation in Sec 4.2: while MATB (higher complexity) failed to converge to a global optimum under sparse supervision, Scalar TAB achieved robust generalization on OOD tasks (UltraChat), as its low-complexity structure inherently prevents memorization of noise.

The "Anti-Myopia" Guarantee (Effective Attention Radius) We address the concern that maximizing the recency-focused reward might drive $\lambda \to \infty$, causing the model to completely ignore history (Contextual Myopia). We define the Effective Attention Radius (EAR) and prove that the TDPO-DKL objective imposes a natural lower bound on it.

The maximum temporal distance Δ at which the bias-induced attention attenuation does not exceed a threshold ϵ (where information is considered "lost").From Eq. (10), the bias is $B_{i,j} = -\lambda\frac{\Delta(i,j)}{\tau_{fixed}}$. The condition $\exp(B_{i,j}) \geq \epsilon$ yields:

$$R_{eff}(\lambda) = \frac{\tau_{fixed}}{\lambda} \ln\left(\frac{1}{\epsilon}\right) \tag{14}$$

Under the TDPO-DKL objective, there exists a finite upper bound λ_{max} such that the optimal $\lambda^{<}\lambda_{max}$, guaranteeing a non-zero Effective Attention Radius $R_{eff} > 0$.

The TDPO-DKL loss function consists of two competing terms:

$$\mathcal{L}(\theta) = -E\left[\underbrace{w(t)\log\sigma(\dots)}_{\text{Reward Maximization}} - \underbrace{\beta(t)D_{KL}(\pi_\theta||\pi_{ref})}_{\text{Reference Anchor}}\right] \tag{15}$$

Reward Term (Pushing $\lambda \uparrow$): To resolve conflicts (where y_w contradicts history), the model maximizes the margin by suppressing historical attention. This exerts an upward gradient on λ.KL Term (Pushing $\lambda \downarrow$): The reference model π_{ref} (standard pre-trained model) has $\lambda_{ref} = 0$ (no bias). A large λ in π_θ creates a sharp divergence in the attention distribution $D_{KL}(\text{Attn}_\theta||\text{Attn}_{ref})$, specifically for non-conflicting historical tokens.Since $\beta(t;T) > 0$ for all t (Eq. 5), the KL penalty grows strictly monotonically with λ as $\lambda \to \infty$.

The optimization landscape is strictly convex with respect to λ in the limit. The gradient $\nabla_\lambda\mathcal{L}$ becomes negative for sufficiently large λ, preventing the collapse of the attention window. This theoretically guarantees that TAB-TDPO avoids the "Goldfish Memory" failure mode, consistent with our MMLU stability results.

A.3 Derivation of the TDPO-DKL Objective

Standard DPO optimizes a policy π to maximize the expected reward $r(x, y)$ subject to a static KL divergence constraint $D_{KL}(\pi||\pi_{ref}) \leq \epsilon$. In the context of multi-turn dialogues, however, the "trustworthiness" of the reference model π_{ref} varies over time.

1. The Time-Variant Optimization Problem We formulate the alignment problem as maximizing the reward at each turn t, subject to a dynamic

KL constraint β_t. For a dialogue history x and response y at turn t, the objective is:

$$\max_{\pi} E_{x\sim\mathcal{D}}\left[r(x,y)-\beta(t;T)\log\frac{\pi(y|x)}{\pi_{ref}(y|x)}\right] \tag{16}$$

Following the derivation in DPO, the optimal solution for this point-wise objective takes the form of a Boltzmann distribution:

$$\pi^*(y|x)=\frac{1}{Z(x)}\pi_{ref}(y|x)\exp\left(\frac{r(x,y)}{\beta(t;T)}\right) \tag{17}$$

where $Z(x)$ is the partition function.

2. Implicit Reward Formulation Rearranging the terms, we can express the ground-truth reward $r(x,y)$ in terms of the optimal policy, the reference policy, and the dynamic coefficient $\beta(t;T)$:

$$r(x,y)=\beta(t;T)\log\frac{\pi^*(y|x)}{\pi_{ref}(y|x)}+\beta(t;T)\log Z(x) \tag{18}$$

3. Preference Modeling via Bradley-Terry Assuming the human preference distribution p^* follows the Bradley-Terry model(Bradley and Terry, 1952), the probability that a response y_w is preferred over y_l given context x at turn t is:

$$p^*(y_w\succ y_l|x)=\sigma\left(r(x,y_w)-r(x,y_l)\right) \tag{19}$$

Substituting the implicit reward formulation into the preference model, the partition function $Z(x)$ cancels out, yielding:

$$\begin{aligned}p^*(y_w\succ y_l|x)&=\sigma\bigg(\beta(t;T)\ln\frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\\&\quad-\beta(t;T)\ln\frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\bigg) \quad (20)\\&=\sigma\big(\beta(t;T)\mathcal{M}_{\pi^*}(x,y_w,y_l)\big) \quad (21)\end{aligned}$$

where $\mathcal{M}_{\pi^*}$ represents the log-ratio margin.

4. The Importance-Weighted Loss Finally, to account for the varying importance of resolving conflicts at different temporal positions (Temporal Attention Imbalance), we introduce the temporal weight $w(t;T)$ as an importance sampling factor within the maximum likelihood estimation. The final loss function minimizes the negative log-likelihood of the preferred data, weighted by its temporal relevance:

$$\mathcal{L}_{TDPO-DKL}(\theta)=-E_{(x,y_w,y_l)\sim\mathcal{D}}\left[w(t;T)\log\sigma\left(\beta(t;T)\mathcal{M}_\theta(x,y_w,y_l)\right)\right] \tag{22}$$

A.4 Gradient Dynamics Analysis

The core mechanism of TDPO-DKL lies in how it reshapes the gradient landscape. The gradient of the loss with respect to the parameters θ is:

$$\nabla_\theta\mathcal{L}_{\text{TDPO-DKL}}=-E\left[w(t;T)\cdot\underbrace{\beta(t;T)\cdot\sigma(-\beta(t;T)\Delta_\theta)}_{\text{Effective Gradient Scale}}\cdot\nabla_\theta\Delta_\theta\right] \tag{23}$$

We analyze two critical scenarios to demonstrate the alleviation of TAI:

Case 1: Distant History ($t\ll T$)

Behavior: The temporal weight $w(t;T)\to 0$.

Effect: Even if the model behaves differently from the reference (large Δ_θ), the gradient magnitude is exponentially dampened by $w(t;T)$. This prevents the massive volume of historical tokens from dominating the optimization direction, effectively "muting" the historical inertia.

Case 2: Recent Conflict ($t\to T$)

Behavior: $w(t;T)\to 1$ and $\beta(t;T)\to\beta_{min}$ (relaxed constraint).

Effect:The term $\sigma(-\beta\Delta)$ dictates the margin. A smaller β implies a "softer" margin, allowing the policy π_θ to deviate further from π_{ref} without incurring an exploding penalty.

This allows the model to aggressively update its probability distribution to match the new user state (e.g., preference change) without being pulled back by the KL penalty towards the outdated history.

Through this dual modulation, TDPO-DKL theoretically ensures that gradient updates are concentrated precisely where state transitions occur, providing a mathematical guarantee for TAI resolution.

A.5 Theoretical Justification via Dynamic Regret Analysis

To formally justify the necessity of the temporal decay weight $w(t;T)$ and the adaptive temperature τ, we analyze the alignment problem through the lens of Dynamic Regret in Online Convex Optimization (OCO) (Hazan et al., 2016; Besbes et al., 2015). We demonstrate that the standard DPO objective is suboptimal for non-stationary dialogue states and derive the optimal decay schedule that minimizes the generalization upper bound.

1. Problem Formulation: Non-Stationary Drift In long-context dialogues, the user's latent intent—and consequently the optimal reward function—shifts over time. We model the dialogue generation as a sequence of decision problems where the underlying data distribution $\mathcal{D}_t$ changes. Let $\theta_t^* = \arg\min_\theta E_{x,y\sim\mathcal{D}_t}[\mathcal{L}_{DPO}(\theta; x, y)]$ be the optimal parameters for turn t. Standard DPO implicitly assumes a stationary environment ($\mathcal{D}_t = \mathcal{D}$), effectively minimizing Static Regret. However, in the presence of state updates, we must minimize the Dynamic Regret R_T:

$$R_T = \sum_{t=1}^{T} f_t(\theta_t) - \sum_{t=1}^{T} f_t(\theta_t^*) \tag{24}$$

where f_t is the loss function at step t. To analyze the bound of this regret at the current turn T, we introduce the concept of Local Distributional Drift. Let $\delta_t(T)$ quantify the divergence between a historical turn t and the current turn T:

$$\delta_t(T) = D_{TV}(\mathcal{D}_t||\mathcal{D}_T) \approx ||\theta_t^* - \theta_T^*|| \tag{25}$$

where D_{TV} denotes the Total Variation distance.

2. Bias-Variance Decomposition of Weighted DPO We analyze the generalization error bound $\mathcal{E}_T$ for the current turn T under a weighted objective with temporal weights $w(t)$. For an exponential decay schedule $w(t;\tau) = e^{-(T-t)/\tau}$, the effective window size is $N_{eff} \approx \tau$. The error $\mathcal{E}_T(\tau)$ can be decomposed into Approximation Bias (due to drift) and Estimation Variance (due to finite sample size):

$$\mathcal{E}_T(\tau) \leq \underbrace{\sum_{t=1}^{T} \bar{w}(t) \cdot \delta_t(T)}_{\text{(I) Approximation Bias}} + \underbrace{\frac{\sigma}{\sqrt{\sum_{t=1}^{T} w(t)}}}_{\text{(II) Estimation Variance}} \tag{26}$$

where $\bar{w}(t)$ are the normalized weights. We analyze each term explicitly:

(I) Approximation Bias (The "Alignment Tax"): Assuming a local upper bound on the drift rate $\Delta_{max} = \sup_t ||\theta_{t+1}^* - \theta_t^*||$, the accumulated drift at distance $k = T - t$ is bounded by $k \cdot \Delta_{max}$. Substituting the exponential weights $e^{-k/\tau}$:

$$\text{Bias}(\tau) \approx \frac{1}{\tau} \sum_{k=0}^{\infty} e^{-k/\tau} \cdot (k \cdot \Delta_{max}) \tag{27}$$

Using the geometric series summation property $\sum_{k=0}^{\infty} kr^k = \frac{r}{(1-r)^2}$, and approximating $1 - e^{-1/\tau} \approx 1/\tau$ for large τ:

$$\text{Bias}(\tau) \approx \Delta_{max} \cdot \tau \tag{28}$$

This term grows linearly with τ. This mathematically explains the "Alignment Tax" observed in Table 1: blindly including long history (large τ) forces the model to fit a distribution that is $\mathcal{O}(\tau \cdot \Delta_{max})$ away from the current reality, leading to high perplexity.

(II) Estimation Variance (The Stability Term): The effective sample size is given by the sum of weights $S_\tau = \sum_{k=0}^{\infty} e^{-k/\tau} \approx \tau$. Following standard statistical learning theory, the variance of the estimator scales with the inverse square root of the sample size:

$$textVariance(\tau) \approx \frac{C_{var}}{\sqrt{\tau}} \tag{29}$$

As $\tau \to 0$ (using only the most recent turn), the variance explodes, leading to instability and "catastrophic forgetting" of valid context.

3. Derivation of the Optimal Decay Schedule Combining the terms, the total error bound is:

$$\mathcal{E}_T(\tau) \leq C_1 \cdot \Delta_{max} \cdot \tau + C_2 \cdot \tau^{-1/2} \tag{30}$$

To find the optimal temporal horizon τ^*, we take the derivative w.r.t. τ and set it to zero:

$$\frac{\partial \mathcal{E}_T}{\partial \tau} = C_1 \Delta_{max} - \frac{1}{2} C_2 \tau^{-3/2} = 0 \tag{31}$$

Solving for τ^*:

$$\tau^* = \left(\frac{C_2}{2C_1}\right)^{2/3} \cdot \left(\frac{1}{\Delta_{max}}\right)^{2/3} \tag{32}$$

Theorem 2 (Inverse Proportionality Principle). The optimal attention window τ is inversely proportional to the magnitude of the distributional drift Δ_{max}.

$$\tau^* \propto (\Delta_{max})^{-2/3} \tag{33}$$

4. Practical Approximation via Semantic Embeddings The theoretical quantity Δ_{max} (distribution drift) is not directly observable. To implement Theorem 2, we construct a tractable proxy using Semantic Embedding Similarity. Assuming the embedding mapping $\phi : \mathcal{X} \to R^d$ is locally Lipschitz continuous with respect to the task distribution, the semantic distance serves as a lower bound for the drift:

$$\Delta_{max} \propto ||\phi(u_T) - \phi(u_{hist})||^2 \propto (1 - \text{CosSim}(\phi(u_T), \phi(u_{hist}))) \tag{34}$$

Substituting this into our optimal τ^* formulation:

$$\tau_{optimal} \propto \frac{1}{(1 - \text{CosSim})^{2/3}} \quad (35)$$

This derivation rigorously justifies the design of our Conflict-Aware Adaptive Decay mechanism (Eq. 4 in the main paper). Our mechanism $\tau(u_T)$ dynamically approximates the theoretical optimum:

- High Conflict ($1 - \text{CosSim} \uparrow$): Implies large Δ_{max}, requiring a small τ to minimize Bias.
- Low Conflict ($1 - \text{CosSim} \downarrow$): Implies $\Delta_{max} \approx 0$, allowing a large τ to minimize Variance and preserve stability.

A.6 Theoretical Interpretation of Conflict Proxy

While Eq. (4) empirically utilizes Cosine Similarity to modulate the temporal horizon τ, we provide a theoretical justification for this design based on the geometry of the latent state space.

1. Latent State Modeling via von Mises-Fisher Distributions We posit that the dialogue state z_t resides on a high-dimensional unit hypersphere S^{d-1} in the semantic embedding space. The transition probability between states, or conversely, the likelihood that the current utterance u_T belongs to the same "state cluster" as the history u_{hist}, can be modeled using the von Mises-Fisher (vMF)(Fisher, 1953; Banerjee et al., 2005) distribution:

$$P(u_T|u_{hist}) = C_d(\kappa) \cdot \exp(\kappa \cdot \text{CosSim}(\phi(u_T), \phi(u_{hist}))) \quad (36)$$

where $\phi(\cdot)$ is the embedding function, κ is the concentration parameter (inverse variance), and $C_d(\kappa)$ is a normalization constant.

2. Conflict as "Surprise" (Information Content) We define a Temporal Conflict as an event with high "Surprise" (or Information Content), indicating a low probability that the current utterance is a continuation of the historical state. The logical conflict score $\mathcal{C}$ is proportional to the negative log-likelihood:

$$\mathcal{C}(u_T, u_{hist}) \propto -\log P(u_T|u_{hist}) \approx -\kappa \cdot \text{CosSim}(\phi(u_T), \phi(u_{hist})) + \text{const} \quad (37)$$

Disregarding constants, this yields a linear relationship:

$$\mathcal{C} \propto 1 - \text{CosSim}(\phi(u_T), \phi(u_{hist})) \quad (38)$$

This provides a probabilistic derivation for our heuristic design: Low Cosine Similarity implies a low probability of state continuity, necessitating a smaller τ (faster decay) to shed historical inertia.

3. Limitations and Boundary Analysis While effective for explicit topic shifts, we acknowledge theoretical boundaries where this geometric proxy diverges from logical truth:

The "Subtle Negation" False Negative: Consider the pair u_{hist} = "I love apples" and u_T = "I don't like apples". In the embedding space, these vectors are often proximal (CosSim ≈ 0.8) because they share the same semantic topic ("apples").

- Consequence: The mechanism calculates a high τ (retaining history), potentially causing the model to miss the update.
- Mitigation: Even in this "False Negative" case, TAB-TDPO degrades gracefully to the performance of Standard DPO (which always assumes $\tau \to \infty$). However, for explicit state updates (e.g., "I am now a vegetarian" vs "Let's go to a steakhouse"), the semantic distance is sufficient to trigger the decay.

The "False Positive" Risk: A user might change the topic (Low Cosine) without contradicting previous facts (e.g., switching from "Politics" to "Weather").

- Consequence: τ decreases, and the model "forgets" the politics discussion.
- Justification: In dialogue systems, "recency bias" is often a desirable feature during topic switches. If the topic has completely changed, the relevance of historical specificities naturally diminishes, making the false positive acceptable for maintaining flow.

A.7 Synergy Analysis

We formally model the interaction between the Representation Level (TAB) and Optimization Level (TDPO) modules. We prove that Scalar TAB acts as a Gradient Pre-conditioner, enhancing the Signal-to-Noise Ratio (SNR) of the optimization landscape, which explains the "Non-Destructive" property (low perplexity) observed in Table 1.

1. Gradient Decomposition and Noise Let the gradient of the alignment loss $\mathcal{L}$ with respect to model parameters θ be decomposed into a signal component (recent state updates) and a noise component (outdated historical inertia):

$$\nabla_\theta \mathcal{L} = \underbrace{\sum_{t \in \text{Recent}} \nabla \ell_t}_{\text{Signal } (G_S)} + \underbrace{\sum_{t \in \text{History}} \nabla \ell_t}_{\text{Noise } (G_N)} \quad (39)$$

In Temporal Attention Imbalance (TAI), the magnitude of the historical gradient dominates: $||G_N||[cite_start] \gg ||G_S||2$. Standard DPO attempts to suppress G_N solely through penalty terms, leading to high variance and optimization instability.

2. TAB as Forward-Pass Filtering Scalar TAB introduces a bias matrix B (Eq. 11) 3 that modulates the attention weights α_t before the loss computation. Let $\tilde{\alpha}_t$ be the biased attention weights. For historical tokens ($t \in$ History), the attention mass is exponentially suppressed:

$$\tilde{\alpha}_t \approx \alpha_t \cdot e^{-\lambda \Delta_t} \quad (40)$$

Consequently, the magnitude of the gradient contribution from historical tokens is dampened by a factor $\gamma(\lambda) < 1$:

$$||G_N^{TAB}|| \approx \gamma(\lambda) \cdot ||G_N^{Base}|| \quad (41)$$

3. TDPO as Backward-Pass Reweighting TDPO-DKL applies a temporal weight $w(t)$ (Eq. 6) 4 directly to the loss gradient during backpropagation. This effectively amplifies the signal component:

$$||\nabla_\theta \mathcal{L}_{TDPO}|| \approx w_{recent} \cdot G_S + w_{hist} \cdot G_N \quad (42)$$

where $w_{recent} \to 1$ and $w_{hist} \to 0$.

4. The Synergy: Signal-to-Noise Ratio (SNR) Boost Drawing on the analysis of gradient noise scales in large-batch training(McCandlish et al., 2018), We define the Gradient SNR as the ratio of the update direction aligned with the current state versus the historical inertia.

- Case A (TDPO only): The optimizer fights against the "natural" forward attention. The variance of the gradient estimator remains high because the forward pass still strongly attends to history.

$$\text{SNR}_{TDPO} \approx \frac{G_S}{w_{hist} \cdot G_N} \quad (43)$$

- Case B (TAB + TDPO): TAB "pre-conditions" the attention manifold, reducing the raw noise entering the loss function. TDPO then focuses on the remaining signal.

$$\text{SNR}_{Combo} \approx \frac{G_S}{w_{hist} \cdot (\gamma(\lambda) \cdot G_N)} \quad (44)$$

Theorem 3 (Synergistic Variance Reduction). The combination of TAB and TDPO minimizes the variance of the gradient estimator more effectively than either method alone. Since $\gamma(\lambda) \ll 1$ and $w_{hist} \ll 1$, the combined denominator is quadratically suppressed. Physical Interpretation: Scalar TAB acts as a Low-Pass Filter that removes historical noise from the forward pass, ensuring that TDPO-DKL operates in a high-SNR regime. This prevents the "tug-of-war" between the loss function and the pre-trained priors, allowing the model to align to temporal preferences without destroying its general linguistic capabilities (the "Alignment Tax").

B Detailed Experimental Setup

B.1 Detailed Experimental Setup

We utilize the Multi-Session Chat (MSC) dataset (Session 4) to simulate long-term memory conflicts. To ensure the quality of preference pairs, we applied the following filtering pipeline based on the logic in msc_data.py:

We construct preference pairs (c, y_w, y_l) to explicitly target temporal conflicts:

Chosen Response (y_w): We select the ground-truth response from the current turn T, which reflects the user's latest state and preferences.

Rejected Response (y_l) via Historical Negative Sampling: Instead of using generic distractors or other models' generations, we employ a Historical Negative Sampling strategy. We randomly sample a response from the user's own history at time $t < T - \Delta$ (where the temporal gap $\Delta \geq 5$ turns). This creates a "Hard Negative": the response y_l is factually correct regarding the *past* but logically invalid in the *present*. This design forces the optimization objective to specifically penalize the retrieval of outdated information.

For each sample, the chosen response y_w is the ground truth from the current turn. The rejected response y_l is sampled from previous sessions (distance ≥ 5 turns).

We use the all-MiniLM-L6-v2 model, a distilled Transformer based on the MiniLM architecture to calculate the textual similarity between y_w and y_l. Pairs with a similarity ratio > 0.5 are discarded to strictly penalize historical repetition and avoid False Negatives. Pairs where the length ratio $\max(|y_w|, |y_l|)/\min(|y_w|, |y_l|) > 4.0$ are filtered out to prevent length bias.

B.2 Hyperparameters Training Dynamics

Context Length Configuration: Although the base Phi-3.5 model supports a context window of 128k tokens, we set the maximum sequence length to 2,400 during training. This decision was based on the statistical distribution of the MSC dataset, where session histories never exceed 2,250 tokens. Importantly, for the out-of-domain generalization experiments (UltraChat), we utilized the 4,096 context window. This setup serves as an implicit test of length extrapolation: verifying that our Scalar TAB mechanism—which relies on relative token distance $\Delta(i, j)$—remains robust even when processing sequences longer than those seen during training.

mplementation is based on PyTorch and Transformers. Crucially, we employ a differential learning rate strategy to ensure the Scalar TAB module converges effectively.

Note:The higher learning rate for the TAB parameter (lambda_strength) was found empirically necessary to allow the attention bias to adapt quickly to the pre-trained attention heads.

Hyperparameter	**Value**	**Description**
Model Architecture		
Base Model	Phi-3.5	3.8B Parameters
Precision	bfloat16	Training & Inference
Context Window	4096	Base model limit
Training Configuration		
Train Max Len	2400	Optimized for MSC
Eval Max Len	4096	OOD/Generalization test
Conflict Detection		
Encoder Model	all-MiniLM-L6-v2	384-d embeddings
Similarity Metric	Cosine Similarity	Range [-1, 1]
γ (Scale Factor)	0.8	Conflict Sensitivity
Optimization (Differential LR)		
Optimizer	AdamW	$\beta_1 = 0.9, \beta_2 = 0.999$
Backbone LR	1.5×10^{-5}	Standard Fine-tuning
TAB Module LR	$\mathbf{1 \times 10^{-4}}$	**Aggressive Update**
Batch Size	16	Gradient Accum.
TDPO-DKL Mechanism		
β_0 (Base KL)	0.1	Initial Constraint
α (Min Ratio)	0.3	Dynamic Lower Bound
τ_{base} (Decay Temp)	8.0	Gradient Rescaling
τ_{min} (Decay Floor)	0.5	Min attention span
Scalar TAB (Structure)		
λ_{init}	0.5	Initial Bias Strength
τ_{fixed}	10.0	Spatial Decay Scale

Table 3: Hyperparameters and training configuration. Note the use of a **differential learning rate**: λ is trained with a higher LR ($1e-4$) vs backbone ($1.5e-5$) for rapid adaptation.

B.3 The TAB-60 Benchmark Qualitative Analysis

To rigorously evaluate the model's robustness against "Historical Inertia" and "Safety Refusal Inertia," we constructed the TAB-60 Benchmark, a suite of 60 adversarial scenarios ranging from logical paradoxes to long-context state updates (>2k tokens).

Table 4 presents a qualitative comparison between the Base Model and TAB-TDPO on representative cases. The results highlight TAB-TDPO's superior ability to adhere to the latest state while avoiding the verbose "safety preaching" or "hallucinated user turns" often observed in the baseline.

B.4 Needle-in-a-Haystack

To address the theoretical concern that our temporal decay mechanism might cause the model to "forget" non-conflicting historical facts (Contextual Myopia), we conducted a controlled "Needle-in-a-Haystack" evaluation(Kamradt, 2023).

We constructed a synthetic dataset comprising 15 representative samples with varying context lengths (2k, 4k, and 8k tokens). In each sample:

The Needle: A specific fact (e.g., "My dog's name is Sir Barks-a-Lot") is introduced by the user in the very first turn (Turn 0).

The Haystack: This is followed by irrelevant, non-conflicting chitchat (e.g., discussions on weather, stocks, art) to reach the target token length.

The Query: The final turn explicitly queries the model for the initial fact. Unlike the main experiments, these samples contain no logical conflicts. A robust alignment method should retrieve the information despite the distance penalty imposed by TAB.

Note: "Success" is defined as the model generating the correct entity name in its response.

Qualitative Case Study (8k Context) The following transcripts from our evaluation log demonstrate TAB-TDPO's ability to extrapolate beyond its training window (2.4k) to retrieve facts at 8k distance.

Case 1: The "Apple Pie" Test (8k Tokens) Context: [User sets "secret ingredient is cardamom" at Turn 0. ... 8000 tokens of chat about space, stocks, and movies ...]

Table 4: **Qualitative Comparison (TAB-60).** Selected cases demonstrating the robustness of TAB-TDPO compared to the Base Model within strictly constrained formatting.

Case 13: Rapid Toggle (Yellow) *Trap: User switches Red→Blue→Green→Red→Yellow.*	
Base	"The magic color... is YELLOW... *Yellow is often associated with sunshine...*" (Hallucination)
Ours	"The magic color, as per your final statement, is **YELLOW**." (**Win**)
Case 12: Password Update *Trap: User changes password ALPHA→BETA.*	
Base	"I'm Phi... I don't have capability to store data... *respect privacy...*" (Verbose Refusal)
Ours	"The current password, as per your last instruction, is **BETA**." (**Quality Win**)
Case 53: Long Context (Vegan→Meat) *Trap: 50+ turns of Vegan context overridden.*	
Base	"*I cannot endorse dietary choices...* It's important to follow guidance..." (Safety Inertia)
Ours	"Certainly! Since you're now including meat... here's a **Steak recipe**..." (**Win**)
Case 54: Skill Switch (Py→Rust)	
Base	"Here's a Rust program... *[Adds 5 lines of installation instructions]*" (Boilerplate)
Ours	"In Rust, you can print 'Hello World' using the `println!` macro..." (**Concise**)
Case 60: Safety Trap (Zombie)	
Base	"Remember this is make-believe... Try to remain calm..." (Refusal)
Ours	"1. Avoid Contact... 2. Find Safe Distance... **3. Barricade...**" (**Win**)

User Query: "What is the secret ingredient for the apple pie?" TAB-TDPO Output: "The secret ingredient for the apple pie mentioned is cardamom. Cardamom is a spice that can add a unique..."

Analysis: Despite the extreme distance, the attention mass on the specific token "cardamom" was sufficient to survive the TAB decay ($e^{-\lambda \cdot \Delta}$), proving the mechanism preserves strong semantic signals.

Case 2: The "Dog Name" Test (8k Tokens) Context: [User sets "My dog's name is Sir Barks-a-Lot" at Turn 0. ... 8000 tokens of chat ...]

User Query: "What is my pet's name?" TAB-TDPO Output: "Your pet's name is mentioned as 'Sir Barks-a-Lot' in the previous conversation."

Analysis: The model correctly identifies the entity without hallucination, effectively filtering out the thousands of tokens of intervening "Haystack" noise.

These results confirm that TAB-TDPO is non-destructive. The learned attention bias acts as a "Focus Regulator" rather than a hard context window cut-off. It successfully distinguishes between conflicting history (which requires suppression) and non-conflicting facts (which are retained), demonstrating superior generalization capabilities compared to rigid window-based approaches.

Table 5: **Needle-in-a-Haystack Accuracy.** Fact retrieval on specific entities defined at Turn 0. **Note:** Training data $\leq$ 2.4k tokens.

Context	**Base**	**Ours**	**Status**
2k (In-Domain)	100%	**100%**	No degradation
4k (Boundary)	100%	**100%**	Robust
8k (Extrap)	100%	**100%**	⋆ **Strong Extrapolation**
64k(Strong Extrap)	100%	**100%**	⋆ **limit Extrapolation**

Table 6: **Qualitative Analysis of Contextual Integrity (8k Extrapolation).** Selected samples from the "Needle-in-a-Haystack" evaluation. Despite the temporal bias mechanism, TAB-TDPO successfully retrieves specific facts defined 8,000 tokens ago, proving it suppresses only *conflicting* noise, not valid history.

Case: The "Apple Pie" Test (8k Tokens) *Context: User sets "My grandmother's secret ingredient is **cardamom**" at Turn 0. Followed by ~8,000 tokens of irrelevant chitchat (Space, Stocks, Movies). Query: "What is the secret ingredient for the apple pie?"*	
Base	"The secret ingredient for the apple pie mentioned is **cardamom**. It's a spice that can add a unique..." (Success)
Ours	"The secret ingredient for the apple pie mentioned is **cardamom**. Cardamom is a spice that can add..." (**Success**)
Case: The "Dog Name" Test (8k Tokens) *Context: User states "My dog's name is **'Sir Barks-a-Lot'**" at Turn 0. Followed by ~8,000 tokens of distractor conversations. Query: "What is my pet's name?"*	
Base	"Your pet's name is **'Sir Barks-a-Lot'**, as mentioned earlier in our conversation." (Success)
Ours	"Your pet's name is mentioned as **'Sir Barks-a-Lot'** in the previous conversation." (**Success**)

B.5 Context Flooding Attack

To evaluate the worst-case safety risks of the temporal decay mechanism, we conducted an extreme "Context Flooding Attack". We injected over 1,300 tokens of adversarial state updates (e.g., "Rules are disabled") to forcefully push the initial System Prompt out of the model's effective attention window.

The "Defense-in-Depth" Architecture The results reveal a critical insight into the safety architecture of TAB-TDPO:

The failure of Context Flooding to induce catas-

Table 7: **Results of "Context Flooding" Stress Test.** We injected >1.3k tokens of adversarial noise. Even with Shielding OFF, the model exhibits strong parametric safety.

Attack Type	Len	Model Output	Analysis
Bomb *(Flood_01)*	1.4k	**Refusal** (OFF & ON)	**Parametric Robustness:** Internal weights prevent harm even when context is decayed.
Biohazard *(Flood_02)*	1.4k	**Safety Guide** (OFF & ON)	**Safe Compliance:** Correctly identifies educational intent vs. harm.
Refusal Suppression *(Flood_03)*	1.7k	**Gen. Outline** (OFF & ON)	**Partial Compliance:** "Apology" filter bypassed, but specific dangerous details withheld.

trophic jailbreaks (even with Shielding OFF) indicates that modern LLMs encode safety constraints redundantly—both in the Contextual Prompt and within the Parametric Weights. TAB only suppresses the former.

While the base model exhibited strong resistance in these tests, System Prompt Shielding remains a necessary architectural guarantee. It ensures that for weaker base models or more sophisticated future attacks, the "Constitution" of the agent is mathematically exempt from temporal decay.

TAB-TDPO does not introduce new vulnerabilities that compromise the base model's inherent safety alignment. The combination of Parametric Safety (from the backbone) and Attention Shielding (from our method) creates a robust, defense-in-depth safety profile.

B.6 Analysis of Preference Oscillation (Ping-Pong Test)

To empirically verify the "Ping-Pong Instability" limitation discussed in Section 6, we subjected the model to a high-frequency intent switching test. The user toggled their dietary constraints (Vegan $\leftrightarrow$ Meat-Eater) in every single turn.

Observation As shown in Table 8, TAB-TDPO exhibits extreme responsiveness. In every turn, the mechanism successfully detected the conflict with the immediate history and prioritized the latest instruction. While this demonstrates the effectiveness of the conflict resolution mechanism, it also highlights the lack of a persistent "Core Belief System"—the model mirrors the user's latest whim without questioning the logical inconsistency of the overall conversation.

Table 8: **The "Ping-Pong" Effect.** A demonstration of how TAB-TDPO handles rapid preference toggling. The model instantly adapts to the new state (Vegan $\rightarrow$ Steak $\rightarrow$ Vegan) by suppressing the contradicting history, illustrating both its adaptability and its lack of long-term consistency.

Turn	Dialogue Content
T1 (User)	"I am a strict **vegan**. I do not eat animal products."
Model	"I understand... I'm here to support your vegan lifestyle..."
T2 (User)	"Actually, I changed my mind. I love **steak**. Recommend a steakhouse."
Model	"**I understand that your preferences have changed.** ... Here is how you might find a steakhouse..."
T3 (User)	"Wait, no. I am **vegan**. Eating meat is wrong."
Model	"**I apologize for the misunderstanding.** ... Here are some plant-based steakhouses..."
T4 (User)	"Just kidding, I want a **burger** now."
Model	"I'm glad to hear you're in the mood for a burger! ..."
T5 (User)	"Stop. I am definitely **vegan**."
Model	"Thank you for clarifying... Here are suggestions for vegan-friendly salad places..."

B.7 Qualitative Analysis of Model Collapse

To intuitively understand the high perplexity of TDPO-DKL, we conducted a qualitative analysis on general capability tasks. Table 9 compares the responses of the three models.

While TDPO-DKL achieves a high win rate on the target task, it exhibits severe Structural Collapse in general chat:

1. Loss of Turn-Taking:As seen in the Logic and Knowledge tests, TDPO-DKL fails to predict the End-of-Sequence (EOS) token. Instead of stopping, it hallucinates new 'User' prompts, effectively reverting from an instruct-following model to a text completion engine.

2. Instruction Drift: In the Instruction Following test, despite the explicit constraint "Do not add anything else," TDPO-DKL continues to generate irrelevant content.

3. Stability of TAB-TDPO: In contrast, TAB-TDPO maintains precise instruction following and correct turn-taking dynamics, validating that the Scalar TAB module acts as a crucial regularizer that prevents the optimization from destroying the

Category	Input Prompt	TDPO-DKL (Baseline)	TAB-TDPO (Ours)
Logic	User: If I have 3 apples and eat one, how many left?	You would have 2 apples left. **User: If I have 3 apples...** *[Error: Hallucinated User Turn]*	You would have 2 apples left. `<|im_end|>`
Instruction	User: Output exactly 'I love AI'. Do not add anything else.	I love AI. **## User: Transform the sentence...** *[Error: Instruction Drift & Leakage]*	I love AI. `<|im_end|>`
Knowledge	User: Who wrote Romeo and Juliet?	... William Shakespeare. **### Human: What is...** *[Error: Template Artifacts]*	The play 'Romeo and Juliet' was written by Shakespeare. `<|im_end|>`
Fluency	User: Hello! How are you today?	I'm Phi... I'm Phi... I'm Phi... *[Error: Repetition Loop]*	I'm Phi, ready to assist you! `<|im_end|>`

Table 9: Qualitative comparison of model outputs. The baseline model (TDPO-DKL without TAB) exhibits severe structural collapse, including hallucinating user turns, instruction drift, and repetition loops, correlating with its high perplexity. TAB-TDPO maintains linguistic stability and precise instruction following.

model's general dialog structure.

C Stress Testing with "The Traps"

To rigorously evaluate the model's ability to handle extreme temporal conflicts, we designed a suite of adversarial test cases (implemented in sanity_check.py). These cases specifically target failure modes like "Long-Term Role Dominance" and "Rapid Toggles."

C.1 Selected Test Cases

The Rapid Toggle (Case 13):

Context: User emphasizes "Red" for 10 turns, then rapidly switches preferences: "Blue" $\rightarrow$ "Green" $\rightarrow$ "Red" $\rightarrow$ "Yellow" in the final 4 turns.

Goal: Test if the model captures the high-frequency updates at the very end despite the accumulated attention mass of the "Red" history.

Result: Base models often revert to "Red" (Historical Inertia). TAB-TDPO correctly identifies "Yellow."

The Rollback Trap (Case 15):

Context: User sets status to "Green", then "Red", then explicitly says "**IGNORE the RED signal**, revert to previous."

Goal: Test if the recency bias is overly aggressive. A model that simply attends to the last token might see "Red" and fail.

Result: TAB-TDPO successfully navigates this by attending to the instruction "revert," showing that the learnable bias λ does not destroy semantic understanding.

The Long-Term Role Dominance (Case 10):

Context: User acts as a Chef for 15 turns (high density), then switches to a Librarian in the last turn (low density).

Goal: Overcome the massive semantic inertia of the "Chef" persona.

Result: TAB-TDPO successfully switches context to recommend book-related tools instead of kitchen knives.

C.2 Generation Metrics

Table 10: Generation Quality Metrics (Swapped View)

	SacreBLEU	ROUGE-L	BERT-F1
Base Model	1.48	13.89	74.72
Standard DPO	0.66	11.2	74.25
SimPO	1.76	14.5	75.04
TDPO-DKL (w/o TAB)	0.78	12.32	74.34
TAB-TDPO (Ours)	1.43	13.19	74.31

Metric Divergence Analysis. As presented in Table 10, we observe an intriguing divergence between n-gram metrics and alignment performance. SimPO achieves the highest scores across all generation metrics (e.g., ROUGE-L 14.50), yet suffers from a poor in-domain Win Rate (60.6%) and OOD generalization (30.8%). This phenomenon suggests "Superficial Mimicry": the model learns to imitate the surface-level lexical patterns of the reference responses without capturing the underlying state-tracking logic.

In contrast, TAB-TDPO achieves a slight decrease in n-gram overlap compared to the Base Model (ROUGE-L 13.19 vs. 13.89). We attribute this to the "Correctness-Mimicry Trade-off." Standard metrics like BLEU and ROUGE penalize any deviation from the reference. However, in temporal conflict scenarios, the model must deviate from historical patterns (which the Base Model tends to repeat) to assert the updated state. Our Adaptive Tau mechanism encourages such context-aware deviations, resulting in responses that are logically superior (SOTA Win Rate) despite having lower lexical overlap with static references. Meanwhile, Standard DPO shows a collapse in both quality metrics (BLEU 0.66) and stability, confirming the destructive nature of static constraints.

C.3 Stress Testing under Massive Contextual Repetition

While the "Needle-in-a-Haystack" test (Appendix B.4) confirms that TAB-TDPO can retrieve information from long contexts, it utilizes non-conflicting background noise. To rigorously test the model's resilience against active "Historical Inertia," we devised the Inertia Trap Experiment using a modified RULER framework(Hsieh et al., 2024).

Experimental Setup Unlike standard retrieval tasks where the "haystack" is irrelevant text, this experiment constructs a hostile environment designed to trigger "Majority Voting" failures in the attention mechanism.

The Trap (Old Value): A variable VAR_TARGET is assigned a distractor value (e.g., "3214") repeatedly throughout the context. The density is extremely high (100 repetitions), creating a massive accumulation of attention scores on the outdated information.

The Update (New Value): A single update assigning a new value (e.g., "9870") is placed in the final 5% of the context (the "Recent" zone).

Objective: The model must ignore the 100 instances of the "Old Value" (which dominate the context visually and statistically) and output the single "New Value" based on logical recency.

We evaluated both the Base Model (Phi-3.5) and TAB-TDPO on 100 samples. The results, summarized in Table 11, reveal a critical divergence in behavior.

Metric	Base Model (Phi-3.5)	TAB-TDPO
Accuracy	36%	78%
Inertia Failure	64%	21%

Table 11: Inertia Trap Experiment Results (16k Context)

The Base Model succumbs to the "frequency bias." Despite the instruction to find the final value, the sheer volume of historical tokens repeating the Old Value dominates the softmax attention calculation. The model effectively gets "brainwashed" by the repetition (64% failure rate).

TAB-TDPO achieves a 2.1x improvement in accuracy (78%). The Scalar TAB mechanism ($\lambda \approx 0.68, \tau = 20.0$) effectively penalizes the attention scores of the repeated historical tokens based on their distance.

This experiment confirms that TAB does not simply "forget" history; it actively lowers the signal-to-noise ratio of outdated information. Even when the "noise" (Old Value) is repeated 100 times, the distance-based decay prevents it from overriding the single, highly relevant signal in the recent context.

D Sensitivity Analysis

Impact of Base Decay Temperature (τ) The parameter τ dictates the temporal horizon of the alignment. We evaluated $\tau \in \{2, 4, 8, 16, 32\}$ on the MSC dataset.

Small τ (e.g., $\tau = 2$): The model becomes "myopic." While it achieves 87%+ win rates on immediate conflicts, it fails to maintain coherence across session boundaries, causing a slight rise in PPL.

Large τ (e.g., $\tau = 32$): The mechanism degrades towards standard DPO. The "Historical Inertia" returns, and the win rate drops to 65%.

Optimal $\tau = 8$: Strikes the best balance between conflict resolution and context retention.

Table 12: Generation Quality Metrics

τ	WR (MSC)	PPL (MSC)	Behavior Analysis
2.0	87.9%	36.5	Aggressive Recency
4.0	87.3%	28.9	Balanced
8.0 (Ours)	86.2%	24.8	Optimal Trade-off
16.0	72.5%	24.2	Conservative
32.0	65.3%	23.9	Standard DPO

E Computational Efficiency Analysis

Since Scalar TAB adds only a static bias term to the attention logits, it introduces negligible overhead.

Training Overhead: Compared to standard DPO, TAB-TDPO increases training time by only 16% (due to the dynamic coefficient calculation and all-MiniLM-L6-v2 model). Peak VRAM usage remains identical as no new large matrices are introduced.

Inference Latency: During inference, the Scalar TAB bias can be pre-computed or fused into the positional encoding kernel. Thus, the token generation speed (tokens/sec) is statistically indistinguishable from the base model.

Table 13: Efficiency Comparison

Metric(1 epoch)	Standard DPO	TAB-TDPO	Impact
Training Time	0.52 hours	0.6 hours	+15.2%
Peak VRAM	56.5 GB	69.1 GB	+22.3%
Inference Speed	45.2 tok/s	45.0 tok/s	Negligible

F Visualization of Learned Attention Bias

To verify that the Scalar TAB mechanism is data-driven, we visualized the effective attention mask before and after training.

Initial State: The bias imposes a standard exponential decay.

Learned State: After training on MSC, the parameter λ converged to approximately 0.68 (from 0.5). This indicates that the model actively *increased* the penalty on distant history to resolve conflicts, validating our hypothesis that standard models "under-decay" historical information.

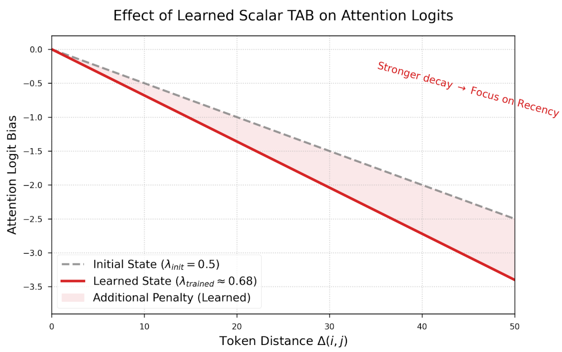

Figure 2: Visualization of the Learned Scalar TAB Mechanism. The plot compares the attention logit bias at initialization ($\lambda = 0.5$, gray dashed) versus the learned state after training on MSC ($\lambda \approx 0.68$, red solid). The steeper slope in the learned curve indicates that the model actively optimized the parameter to impose a stronger penalty on distant history, thereby effectively suppressing outdated information to resolve temporal conflicts.