SOMA: EFFICIENT MULTI-TURN LLM SERVING VIA SMALL LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly deployed in multi-turn dialogue settings where preserving conversational context across turns is essential. A standard serving practice concatenates the full dialogue history at every turn, which reliably maintains coherence but incurs substantial cost in latency, memory, and API expenditure, especially when queries are routed to large proprietary models. Existing approaches often struggle to balance the trade-off between response quality and efficiency. We propose a framework that exploits the early turns of a session to estimate a local response manifold and then adapt a smaller surrogate model to this local region for the remainder of the conversation. Concretely, we train soft prompts that push the small surrogate away from the large model's responses, with an anti-degeneration term for stability, and score each early turn by how much agreement survives. Localized LoRA fine-tuning then distills the turns with robust agreement, so the surrogate runs without prompts at inference. A gate then hands the session to the surrogate and rolls back on drift. We further provide a theoretical analysis for key components in SOMA. Experiments on six multi-turn benchmarks suggest that SOMA largely preserves the original model's responses and lowers serving cost for medium-to-long sessions that stay on topic. The source code is provided at: `https://anonymous.4open.science/r/SOMA-4175`.

1 INTRODUCTION

Large language models (LLMs) such as the GPT series (Brown et al., 2020; OpenAI, 2023), LLaMA (Touvron et al., 2023; Grattafiori et al., 2024), Claude (Anthropic, 2023), and DeepSeek (Guo et al., 2025) power machine-learning-as-a-service (MLaaS) applications including chat assistants, content creation, coding, and clinical decision support (Park et al., 2023; Dong et al., 2023; Zhao et al., 2025; Liu et al., 2026; Li et al., 2026), and *multi-turn LLM serving*, which covers extended interactions between humans and LLMs or among LLM agents, has become a key research focus (Yi et al., 2025; Li et al., 2025b). Serving such interactions efficiently remains hard: most serving systems are stateless and resend the entire conversation history with every new query (Gao et al., 2024; Jeong & Ahn, 2025), so computation, latency, and cost all grow as the conversation lengthens.

Prior work follows two routes. *Single-model methods* compress the dialogue history (Wang et al., 2025; Chen et al., 2025; Xiao et al., 2024), retrieve memory from external modules (Melz, 2023; Gutiérrez et al., 2024), or reuse attention computation (Gao et al., 2024; Jeong & Ahn, 2025; Anthropic, 2024); they still run the large model at every turn, so cost, latency, and GPU usage stay high, and they often truncate the context that long dialogues depend on. *Multi-model methods* route easy queries to small models and escalate hard ones (Behera et al., 2025; Ding et al., 2024; Shnitzer et al., 2024; Cheng & Dong, 2026), but small models generalize poorly across dialogue complexity, and switching adds overhead. Moreover, LLMs rely heavily on early turns (Xiao et al., 2024; Laban et al., 2026), which makes coherence harder to maintain as a dialogue grows.

Is it possible to achieve an efficient, context-aware multi-turn LLM serving framework that avoids recomputing the full history at every turn while preserving response quality?

To answer this question, we examine real-world multi-turn dialogues across domains and find a long-tail distribution of token counts over turns: early turns are substantially longer, while later turns shrink. This matches the intuition that early turns usually introduce the main issues, goals, and constraints that anchor a conversation, whereas later turns are often brief follow-ups (He et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2018; Stolcke et al., 2000). The trend suggests that a smaller model might serve the short later turns cheaply. The bottleneck, however, is the "big head": most grounding is established early, and later turns depend on it, so a small model that behaves reasonably at first drifts from the large model as the dialogue progresses. The small model must therefore approximate the larger model's behavior within the local region of its output space that the prior dialogue induces, which we formulate as a local manifold approximation problem (Section 2).

Building on these insights, we present SOMA (*S*oft-prompts for l*O*cal *M*anifold *A*pproximation), which adapts a small language model to the local behavior of a large one, conditioned on the early turns of a session, through a three-stage pipeline: (1) *soft-prompt tuning*, which pushes the small model away from the large one inside the local region and records, for each early turn, how much agreement is left; (2) *localized fine-tuning*, which aligns the small model with the large one on the early turns, weighted by these scores; and (3) *efficient inference*, which serves later turns from a compressed context and rolls back to the large model when the topic drifts. Our contributions are:

- **Long-tail pattern in multi-turn dialogues.** The first few turns of a dialogue concentrate heavy context, while later turns are shorter yet more dependent on earlier ones, so a small, cheap model could serve the later turns if it is given the accumulated context.
- **SOMA: Efficient multi-turn serving.** Adversarial soft prompts measure how robustly the surrogate reproduces the original model on each early turn, a localized LoRA adapter distills the behavior that holds up, and a cosine gate handles switching and rollback at prompt-free inference time.
- **Theory analysis and empirical evaluation.** We derive a held-out acceptance bound for switching, a per-session cost model, and a coverage bound for random prompt candidates under explicit assumptions, and we validate SOMA on six multi-turn benchmarks and two model families.

2 Preliminaries

2.1 Notations

A multi-turn dialogue prefix of length k is $\mathcal{D}_k = \{(q_1, a_1), \ldots, (q_k, a_k)\}$, where q_t is the user query at turn t and a_t is the model response. We write F for the original model, a black-box LLM as in MLaaS, and G for the small surrogate; the response of $M \in \{F, G\}$ at turn t is a_t^M. We write $S_t = \mathcal{D}_{t-1} \oplus q_t$ for the dialogue state at turn t, i.e., the history followed by the new query. A soft prompt of length L is a matrix $\mathbf{P} \in \mathbb{R}^{L \times d}$. Appendix A lists all symbols used in the paper.

2.2 Exploring Token–Turn Patterns in Multi-Turn LLM Dialogues

Standard serving processes each new request together with the entire history, so the cost of a turn grows with the length of the conversation. The amount of *new* information in a turn, however, need not grow at all. We therefore first study token usage across turns.

We examine four representative settings, which together cover different speaker roles, task goals, and interaction structures: *ShareGPT* (ShareGPT, n.d.) for open-ended human–LLM chat, *ReMeDi* (Yan et al., 2022) for doctor–patient consultation, *Craigslist Bargain* (He et al., 2018) for negotiation, and *Multi-Character* (agentlans, 2024) for multi-party role-play. For each dataset, we compute the average token count at turn t and normalize it by the average of Turn 1 in that dataset.

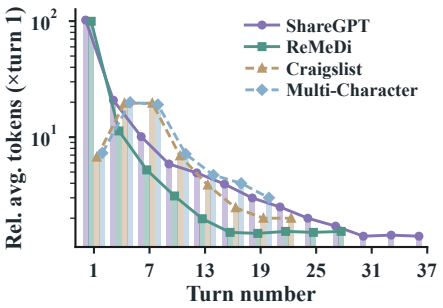


Figure 1: Average token count per turn, normalized by Turn 1, on four dialogue datasets: usage drops after the early turns and forms a long tail.

Figure 1 shows a consistent long-tail pattern. Early turns are token-heavy because they introduce the task, the topic, the constraints, and the speaker's intent; later turns are much shorter and settle at a low, stable level. ShareGPT and ReMeDi peak at the opening turn, whereas Craigslist and Multi-Character peak shortly after it, once the negotiation or role-play is set up. Shorter does not mean independent: many later turns are local updates to an established dialogue state, so the marginal new input shrinks while standard serving keeps re-processing the full prefix. This mismatch motivates an adaptive

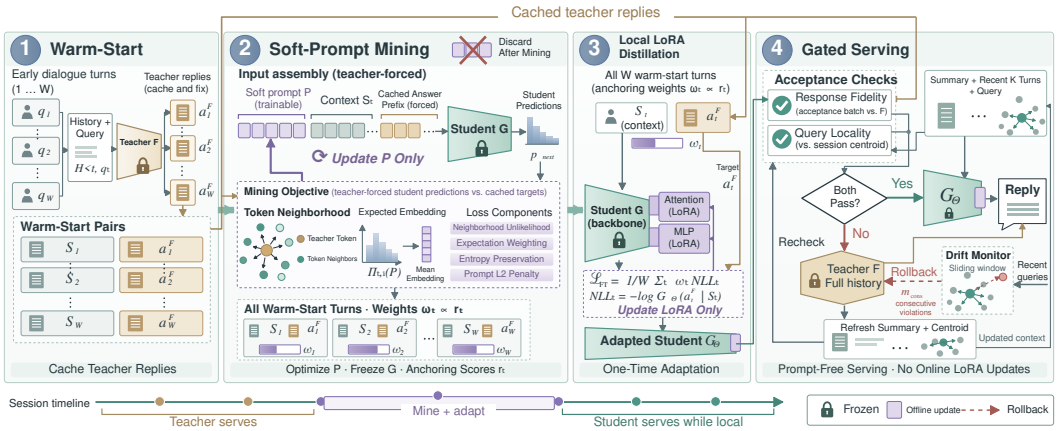


Figure 2: Overview of SOMA. (1) The original model (teacher) F serves the first W turns, and its replies are cached. (2) A soft prompt on the frozen surrogate (student) G is trained to push G away from these replies, which gives each turn an anchoring score r_t. (3) LoRA distills F's replies into G on all warm-start turns, weighted by r_t. (4) Once both checks pass, the adapted G_Θ serves later turns from a compressed context, and a drift monitor rolls back to F.

strategy: let the original model establish the early dialogue state, and hand subsequent turns to a cheaper surrogate for as long as they stay within that state.

2.3 Local Manifold Approximation for Multi-Turn Dialogues

The long-tail pattern creates an opportunity for surrogate serving, but a naive substitution is unreliable. A later turn is short because the preceding dialogue supplies much of its meaning, so a surrogate that is misaligned with the original model under that prefix can give fluent but drifting answers.

We formalize this challenge with a local view of the session. Given a prefix $\mathcal{D}_k$, let $\mathcal{Q}_k$ be the distribution of the dialogue states $S_t = \mathcal{D}_{t-1} \oplus q_t$, $t > k$, that the session visits afterwards. We call its support the local region, or local manifold, induced by the prefix. Since the hidden spaces of F and G differ in dimension and have no canonical correspondence, we compare the two models through their responses. A fixed sentence encoder ϕ embeds both, and $\mathsf{Gap}_G(S) = 1 - \cos\big(\phi(a^G(S)), \phi(a^F(S))\big) \in [0, 1]$. The goal is a local rather than a global match, which is what makes a much smaller surrogate viable.

Problem 2.1 (Local Manifold Approximation for Multi-Turn Interactions)**.** *Given a prefix $\mathcal{D}_k$ and the original model's responses $a_1^F, \ldots, a_k^F$, learn a surrogate $G \in \mathcal{G}$ that minimizes the expected discrepancy on the turns that follow:*

$$\min_{G \in \mathcal{G}} \; R_k(G) = \mathbb{E}_{S \sim \mathcal{Q}_k}\big[\mathsf{Gap}_G(S)\big].$$

Only the first k turns are observed, so the problem is solvable only if they carry enough supervision for the later ones. The local view makes this testable. While later states stay close to the warm-start states, the gap on warm-start turns should track the gap on later turns, so post-switch similarity should rise with the warm-start length, level off, and fall once the topic shifts (Figures 3(c) and 6). In our data, goal-anchored dialogues such as medical consultations and negotiations seem to meet this premise better than reasoning-heavy or multi-party ones (Section 5).

3 SOMA: Local Manifold Approximation via Soft Prompts

Figure 2 summarizes SOMA in four steps. (1) The original model F serves the first W turns and thereby establishes the dialogue state. (2) Adversarial soft-prompt mining then estimates, for each warm-start turn, how much agreement between the surrogate G and F remains after a perturbation meant to remove it (Sections 3.1–3.2). (3) A lightweight LoRA adapter distills F's responses, weighted by this score, and (4) a semantic gate hands the session to G, with rollback to F when the dialogue leaves the learned region (Section 3.3).

3.1 INITIALIZATION

We initialize a soft prompt $\mathbf{P} \in \mathbb{R}^{L_p \times d}$ on the surrogate G by sampling each row i.i.d. from $\boldsymbol{p}_\ell \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$, where $\sigma > 0$ sets the initialization scale, and let $\mathbf{E} = [\boldsymbol{e}_v]_{v \in \mathcal{V}} \in \mathbb{R}^{d \times |\mathcal{V}|}$ be the surrogate embedding matrix, which also defines the semantic neighborhoods used below.

With $\mathcal{D}_{t-1}$ the text history before turn t and q_t the user query, the mining target is the response that F already produced while serving the warm-start turn, and the surrogate is probed with the soft prompt prepended at its embedding layer:

$$a_t^F = F(\mathcal{D}_{t-1} \oplus q_t)\,, \qquad a_t^G = G(\mathbf{P} \oplus_{\text{emb}} \mathbf{E}(\text{tok}_G(\mathcal{D}_{t-1} \oplus q_t)))\,.$$

Here $\text{tok}_G(\cdot)$ is the tokenizer of G, $\mathbf{E}(\cdot)$ maps tokens to embeddings, and $\oplus_{\text{emb}}$ concatenates along the sequence. Both models see the same dialogue state, and only the surrogate is perturbed. We call F once per warm-start turn and never with the prompt, so its responses stay fixed during optimization, while gradients flow only through $\mathbf{P}$ on G (Appendix G).

3.2 ADVERSARIAL SOFT-PROMPT MINING

We learn $\mathbf{P}$ by minimizing a differentiable objective that pushes G away from F on the local dialogue context. The objective has three parts: a token-level semantic divergence loss, an expectation-weighted distribution-level term, and an anti-degeneration regularizer.

Semantic divergence loss. An unlikelihood-style loss (Welleck et al., 2020) discourages G from assigning high probability to the tokens produced by F. An exact-token penalty alone is too narrow, since G could avoid the teacher token yet still favor close paraphrases, so we penalize a semantic neighborhood in the surrogate embedding space.

Definition 3.1 (Semantic Neighborhood)**.** For token $u \in \mathcal{V}$, let $\mathcal{N}_k(u)$ be the set of the k tokens in $\mathcal{V} \setminus \{u\}$ whose embeddings have the largest cosine similarity to $\boldsymbol{e}_u$.

At turn t the original model produces a_t^F, which we tokenize with the surrogate tokenizer, $\boldsymbol{y}_t^F = \text{tok}_G(a_t^F) = (y_{t,1}^F, \ldots, y_{t,T_F(t)}^F)$, so that teacher output and surrogate distribution share one vocabulary. Each teacher token spreads temperature-weighted mass over itself and its k nearest neighbors in the surrogate embedding space:

$$s_\tau(v \mid y_{t,i}^F) = \frac{\exp(\cos(\boldsymbol{e}_v, \boldsymbol{e}_{y_{t,i}^F})/\tau)}{\sum_{u \in \{y_{t,i}^F\} \cup \mathcal{N}_k(y_{t,i}^F)} \exp(\cos(\boldsymbol{e}_u, \boldsymbol{e}_{y_{t,i}^F})/\tau)}, \qquad \tau > 0.$$

Let $S_t = \mathcal{D}_{t-1} \oplus q_t$. With causal masking, a single forward pass of G on the prefix S_t and the teacher answer prefix gives the next-token distributions $\{\Pi_{t,i}(\mathbf{P})\}_{i=1}^{T_F(t)}$, where $\Pi_{t,i}(\mathbf{P}) = \pi_G(\cdot \mid S_t; \mathbf{P}, a_{t,<i}^F)$. The semantic divergence loss is

$$\mathcal{L}_{\text{sem}}(\mathbf{P}; \mathcal{D}_{t-1}, q_t) = \frac{1}{T_F(t)} \sum_{i=1}^{T_F(t)} \sum_{v \in \{y_{t,i}^F\} \cup \mathcal{N}_k(y_{t,i}^F)} s_\tau(v \mid y_{t,i}^F) \Big[-\log\big(1 - \Pi_{t,i}(\mathbf{P})[v]\big) \Big]. \quad (1)$$

Minimizing Eq. 1 lowers the agreement of G with F on exact tokens and on nearby semantic alternatives, so agreement that rests only on a paraphrase is unlikely to survive the probe.

Expectation-weighted semantic divergence. The neighborhood loss can still miss distribution-level alignment: G may lower the probability of a teacher token and its neighbors while spreading mass over many other tokens with similar meaning. We therefore compare the teacher token's embedding with the expected embedding of the surrogate's next-token distribution, $\bar{\boldsymbol{e}}_{t,i} = \mathbf{E}\Pi_{t,i}(\mathbf{P}) = \sum_{v \in \mathcal{V}} \Pi_{t,i}(\mathbf{P})[v] \boldsymbol{e}_v$. We therefore upweight the neighborhood penalty whenever the expected embedding still points toward the teacher token, using the weight $w_{t,i}(\mathbf{P}) = 1 + \lambda_{\text{exp}} \,\text{clip}\Big(\cos(\bar{\boldsymbol{e}}_{t,i}, \boldsymbol{e}_{y_{t,i}^F}), 0, 1\Big)$ with $\lambda_{\text{exp}} \geq 0$. The final semantic mining loss is

$$\mathcal{L}_{\text{sem_exp}}(\mathbf{P}; \mathcal{D}_{t-1}, q_t) = \frac{1}{T_F(t)} \sum_{i=1}^{T_F(t)} w_{t,i}(\mathbf{P}) \sum_{v \in \{y_{t,i}^F\} \cup \mathcal{N}_k(y_{t,i}^F)} s_\tau(v \mid y_{t,i}^F) \Big[-\log\big(1 - \Pi_{t,i}(\mathbf{P})[v]\big) \Big]. \quad (2)$$

The two terms thus block local copying and semantic shadowing.

Anti-degeneration regularizer. Mining can collapse if G piles probability onto a few high-frequency tokens, yielding bland continuations that suppress F's tokens trivially and make the per-turn scores uninformative (Li et al., 2023; Holtzman et al., 2020; Meister et al., 2023). We therefore preserve the entropy of G's predictions over the last K teacher positions, reusing the logits of the same pass:

$$H_{\text{tail}}(\mathbf{P}; t) = \frac{1}{K} \sum_{i \in \text{tail}(t,K)} \left[- \sum_{v \in \mathcal{V}} \Pi_{t,i}(\mathbf{P})[v] \log \Pi_{t,i}(\mathbf{P})[v] \right].$$

The complete prompt-mining objective for a minibatch of warm-start turns $\mathcal{B}$ is

$$\mathcal{J}(\mathbf{P}) = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} [\mathcal{L}_{\text{sem_exp}}(\mathbf{P}; \mathcal{D}_{t-1}, q_t) - \beta H_{\text{tail}}(\mathbf{P}; t)] + \lambda_P \|\mathbf{P}\|_F^2. \tag{3}$$

We minimize Eq. 3 with AdamW and gradient clipping while all weights of G stay frozen. Neighbors are precomputed once by nearest-neighbor search over normalized token embeddings, and $\bar{e}_{t,i}$ uses a top-m truncation of $\Pi_{t,i}(\mathbf{P})$, so the per-step overhead is linear in the number of teacher-prefix positions and scales with $k + m$ rather than with the vocabulary size $|\mathcal{V}|$.

3.3 Localized Fine-Tuning, Switching, and Rollback

Each warm-start turn receives an anchoring score, its largest residual mining loss over the candidate prompts, $r_t = \max_{\mathbf{P} \in \mathcal{P}_{\text{cand}}} \mathcal{L}_{\text{sem_exp}}(\mathbf{P}; \mathcal{D}_{t-1}, q_t)$. Because $\mathcal{L}_{\text{sem_exp}}$ grows with the probability that G assigns to F's tokens and their neighbors, a large r_t indicates that G still predicts F's response after a probe has pushed it away, which suggests that the response is anchored in the dialogue state. The training set $\mathcal{H}$ holds the warm-start turns with F's responses, weighted by r_t. Anchored responses tend to carry the topic, constraints, and style that later turns share. Fragile ones often carry turn-specific content that can dominate a fit on W examples, a concern that also motivates small-loss selection (Han et al., 2018) and self-paced learning (Kumar et al., 2010).

We freeze the base weights of G, attach LoRA (Hu et al., 2022) to attention and MLP projections, and optimize the weighted negative log-likelihood of F's responses

$$\mathcal{L}_{\text{FT}}(\mathbf{\Theta}_{\text{LoRA}}) = \frac{1}{|\mathcal{H}|} \sum_{(S_t, a_t^F) \in \mathcal{H}} \omega_t \, \text{NLL}\big(a_t^F \mid G(S_t; \mathbf{\Theta}_{\text{LoRA}})\big), \qquad \omega_t \propto r_t. \tag{4}$$

The soft prompts are then discarded, and serving uses only the LoRA-adapted surrogate.

Before switching, SOMA verifies two conditions: the adapted surrogate must reproduce the original model's responses on a small acceptance batch, and the current query must remain close to the warm-start dialogue centroid. Once accepted, G serves later turns from a compressed input $\widetilde{S}_t$ that keeps a fixed-budget summary of earlier content together with the most recent K turns, so the full history is never re-processed. During surrogate serving, SOMA keeps monitoring semantic drift. If recent queries move away from the learned local state for m_{cons} consecutive checks, SOMA rolls back to F and refreshes the summary and the centroid from the turns F serves after the shift. It hands the session back to G only if both gates pass on these turns. The adapter is trained once, so a segment it cannot serve stays on F. Gradual drift also triggers rollback once the margin is exhausted.

4 Theoretical Analysis

This section answers two design questions: when SOMA may accept the session-local surrogate, and when its warm-start overhead is amortized. Both answers assume that later turns stay tied to the warm-start state, though not that they are easier. The results describe when SOMA can operate and claim no advantage over other mining strategies. Appendix E has the proofs and a candidate bound.

4.1 Switching Criterion

At turn t, with shared context $S_t = \mathcal{D}_{t-1} \oplus q_t$, we measure the discrepancy of the adapted surrogate with the bounded score $\mathsf{Gap}_\Theta(S_t) \in [0, 1]$ of Section 2, where Θ is the current adapter. Since

consecutive turns are dependent, we model per-turn statistics as an m-dependent sequence and use the effective sample sizes $W_{\text{eff}} = W/(m+1)$ for the warm-start window and $|B|_{\text{eff}} = |B|/(m+1)$ for an acceptance batch B. Lemma E.1 shows that, with probability at least $1-\delta$, a bounded warm-start statistic lies within $\eta_W(\delta) = \sqrt{\log(2/\delta)/(2W_{\text{eff}})}$ of its mean, so strongly dependent early turns may need a longer warm start before the switch can be assessed.

The fidelity gate evaluates the adapted surrogate on an acceptance batch B through its average gap $\widehat{\mathsf{G}}_B = \frac{1}{|B|}\sum_{S_j \in B} \mathsf{Gap}_\Theta(S_j)$, and the surrogate passes when $\widehat{\mathsf{G}}_B$ is below a threshold ε.

Proposition 4.1 (Acceptance Bound for Switching)**.** *Assume that warm-start and later states share the distribution $\mathcal{Q}_k$ (the locality premise of Section 2), let B be drawn from it and be disjoint from the turns used to fit Θ, and let $\gamma_B(\delta) = \sqrt{\log(1/\delta)/(2|B|_{\text{eff}})}$. With probability at least $1-\delta$, $\mathbb{E}_{S\sim\mathcal{Q}_k}[\mathsf{Gap}_\Theta(S)] \le \widehat{\mathsf{G}}_B + \gamma_B(\delta)$. Hence $\widehat{\mathsf{G}}_B \le \varepsilon - \gamma_B(\delta)$ certifies $\mathbb{E}[\mathsf{Gap}_\Theta] \le \varepsilon$.*

A union bound combines the two results (Corollary E.2). An abrupt topic change typically fails the locality gate and triggers an immediate rollback to F. The guarantee has two practical limits. It needs a held-out acceptance batch, but our implementation reuses the warm-start turns for adaptation and acceptance, so in the reported experiments the gate works as a heuristic check. The bound is also loose. With $\delta = 0.05$ and $|B| = 24$, $\gamma_B \approx 0.25$, which exceeds the thresholds $\varepsilon \in [0.05, 0.12]$ themselves, so we set these thresholds empirically (Appendix E).

4.2 Net Efficiency Gain

Because SOMA pays its warm-start overhead once, its efficiency must be judged per session. Let T be the number of turns, W the switch point, and $c_F(\cdot)$ and $c_G(\cdot)$ the per-turn costs (latency, tokens, or price) of F and of G on its compressed input $\widetilde{S}_t$; the net gain over serving every turn with F is

$$\Delta C(T;W) = \sum_{t=W+1}^{T} \Big[c_F(\mathcal{D}_{t-1} \oplus q_t) - c_G(\widetilde{S}_t)\Big] - C_{\text{probe}}(W) - C_{\text{LoRA}}(W) - C_{\text{rb}}(T;W), \quad (5)$$

where C_{probe} and C_{LoRA} are the mining and adaptation costs and C_{rb} covers drift checks and rollback; SOMA pays off only when the accumulated post-switch savings exceed this one-time overhead. This is consistent with Figure 3(a), where sessions of 1–4 turns do not benefit and locally coherent sessions of 5 or more turns recover the initialization cost.

5 Empirical Studies of the Effectiveness of SOMA

We evaluate SOMA through five questions. **RQ1**: Can a small surrogate preserve the original model's responses? **RQ2**: Does higher response similarity also improve task-grounded quality? **RQ3**: When does SOMA provide net efficiency gains after warm-start and adaptation overhead? **RQ4**: Which components of SOMA are most important? **RQ5**: When is session-local serving reliable, and when should SOMA fall back to the large model?

Datasets. We evaluate SOMA on six multi-turn benchmarks: ShareGPT (ShareGPT, n.d.), ReMeDi (Yan et al., 2022), Craigslist (He et al., 2018), Multi-Char (agentlans, 2024), MATH (Hendrycks et al., 2021), and MT-Bench (Zheng et al., 2023). Dataset details and the context-dependence filtering are in Appendix C.1.

Models and baselines. We test two families: LLaMA, using LLaMA-3.1-70B (Grattafiori et al., 2024) as F and LLaMA-2-7B (Touvron et al., 2023) as G, and Qwen (Yang et al., 2025), using Qwen-3-8B as F and Qwen-3-0.6B as G. The baselines are Original, Surrogate, History-Prefix, History-FT, LLMLingua-2 (Pan et al., 2024), and RouteLLM (Ong et al., 2025) (Appendix C.3).

Metrics. We mainly measure teacher-fidelity similarity, i.e., how closely each method preserves the original model's response, using three LLM judges: GPT-OSS (OpenAI, 2025), DeepSeek-V3 (DeepSeek-AI, 2024), and Gemma-2-27B (Gemma Team, 2024). To complement similarity with task-grounded quality, we report EM accuracy on MATH. Efficiency is measured by token usage, throughput, and end-to-end wall-clock latency, which for SOMA includes soft-prompt mining, LoRA adaptation, and gating. All models run on vLLM 0.10.2 (Kwon et al., 2023) with automatic prefix caching, so every baseline reuses the KV cache across turns.

Table 1: Response similarity to the original model on the LLaMA family. Higher is better. SOMA stays closest to the original model on all six datasets. It raises the average from 75.1 for the unadapted surrogate to 93.1, 0.9 points above the strongest baseline, RouteLLM.

	ShareGPT	ReMeDi	Craigslist	Multi-Char	MATH	MT-Bench	Avg.
Surrogate	79.2 ± 2.18	82.7 ± 1.95	74.3 ± 2.36	70.8 ± 1.73	66.2 ± 2.91	77.5 ± 2.04	75.1 ± 5.98
History-Prefix	86.1 ± 1.67	87.9 ± 1.55	82.4 ± 1.72	84.7 ± 1.69	80.3 ± 3.83	87.2 ± 1.58	84.8 ± 2.94
History-FT	93.4 ± 2.12	91.8 ± 2.09	90.3 ± 1.23	89.1 ± 2.23	87.6 ± 2.26	92.4 ± 1.14	90.8 ± 2.18
LLMLingua-2	84.6 ± 1.72	86.4 ± 1.63	80.8 ± 1.91	82.9 ± 1.58	78.1 ± 2.77	85.3 ± 1.66	83.0 ± 3.11
RouteLLM	95.3 ± 1.44	92.5 ± 1.07	91.0 ± 1.86	91.4 ± 1.23	89.6 ± 1.95	93.2 ± 1.12	92.2 ± 1.98
SOMA	$\mathbf{96.4 \pm 1.91}$	$\mathbf{93.2 \pm 0.98}$	$\mathbf{91.9 \pm 2.49}$	$\mathbf{92.3 \pm 1.05}$	$\mathbf{90.7 \pm 1.12}$	$\mathbf{94.1 \pm 0.91}$	$\mathbf{93.1 \pm 1.99}$

Table 2: Response similarity to the original model on the Qwen family. Higher is better. SOMA again leads on all six datasets. Similarities are lower than for LLaMA, probably because the size gap is wider, and SOMA's average gain over the unadapted surrogate grows to 27.1 points.

	ShareGPT	ReMeDi	Craigslist	Multi-Char	MATH	MT-Bench	Avg.
Surrogate	50.9 ± 1.21	63.5 ± 2.87	48.2 ± 3.40	44.7 ± 3.54	42.6 ± 1.70	40.4 ± 0.81	48.4 ± 8.31
History-Prefix	70.5 ± 1.15	75.7 ± 2.01	63.0 ± 2.23	65.5 ± 0.91	51.3 ± 2.45	53.6 ± 2.32	63.3 ± 9.47
History-FT	76.4 ± 1.68	79.2 ± 2.61	74.5 ± 3.56	63.2 ± 1.74	65.9 ± 1.70	62.7 ± 0.72	70.3 ± 7.23
LLMLingua-2	68.9 ± 1.32	74.1 ± 2.14	61.2 ± 2.67	63.4 ± 1.18	49.8 ± 2.03	51.7 ± 1.45	61.5 ± 9.49
RouteLLM	79.6 ± 1.04	81.9 ± 1.36	75.1 ± 2.91	72.8 ± 1.62	67.3 ± 1.21	68.0 ± 0.96	74.1 ± 5.95
SOMA	$\mathbf{81.0 \pm 0.93}$	$\mathbf{83.2 \pm 1.21}$	$\mathbf{76.4 \pm 3.85}$	$\mathbf{74.2 \pm 2.53}$	$\mathbf{68.7 \pm 1.28}$	$\mathbf{69.2 \pm 1.08}$	$\mathbf{75.5 \pm 5.97}$

Table 3: MATH exact-match accuracy. SOMA stays below the original model but recovers 77–78% of the gap between the surrogate and the original model and outperforms every other baseline.

Family	Orig.	Surr.	Hist.-P	Hist.-FT	Route	SOMA
LLaMA	48.34 ± 0.32	19.20 ± 0.78	25.03 ± 0.94	31.46 ± 0.87	33.88 ± 0.79	$\mathbf{41.62 \pm 0.66}$
Qwen	36.48 ± 0.41	11.73 ± 0.64	16.21 ± 0.79	22.57 ± 0.83	25.08 ± 0.71	$\mathbf{31.14 \pm 0.74}$

Implementation. We set the warm-start window, acceptance batch, candidate count, and switch threshold from Section 4; after switching, SOMA performs no further LoRA updates and keeps only the drift check. Appendix C.2 lists all hyperparameters.

5.1 Experimental Results

RQ1: Response similarity to the original model. After the warm-start stage, the surrogate should preserve the original model's behavior within the current dialogue state. Table 1 reports response similarity on the LLaMA family. SOMA achieves the highest average similarity across all six datasets. The gap over Surrogate shows the behavior loss caused by directly replacing the large model with a smaller model. The gap over History-FT shows that local adaptation helps, and seems to help more when the warm-start turns are weighted by their anchoring scores. The largest gains appear on MATH and Multi-Char, where later turns depend on accumulated reasoning states, constraints, and roles. Table 2 repeats the evaluation on the Qwen family. Absolute similarities are lower, likely because the capacity gap between Qwen-3-8B and Qwen-3-0.6B is wider and local imitation becomes harder. SOMA is still the best method on every dataset.

RQ2: Task-grounded quality. Response similarity is the primary metric for replacing the original model, but similarity alone does not guarantee task quality. We therefore evaluate MATH using exact-match accuracy. Table 3 shows that SOMA remains below the original large model, as expected, but substantially narrows the gap between the surrogate and the original model. SOMA also outperforms History-Prefix, History-FT, and RouteLLM in both model families. Thus SOMA's gains are not only surface-level matching: local adaptation also preserves more of the original model's reasoning.

RQ3: End-to-end efficiency. SOMA reduces post-switch cost by serving the adapted surrogate with compressed context, but it also introduces one-time costs from soft-prompt probing and LoRA adaptation. Thus, the key question is whether the total session cost decreases after this overhead is included. SOMA reduces average token usage by avoiding repeated full-history serving after switching (Appendix F.1). Figure 3(a) further shows the amortized operating range: SOMA is slightly

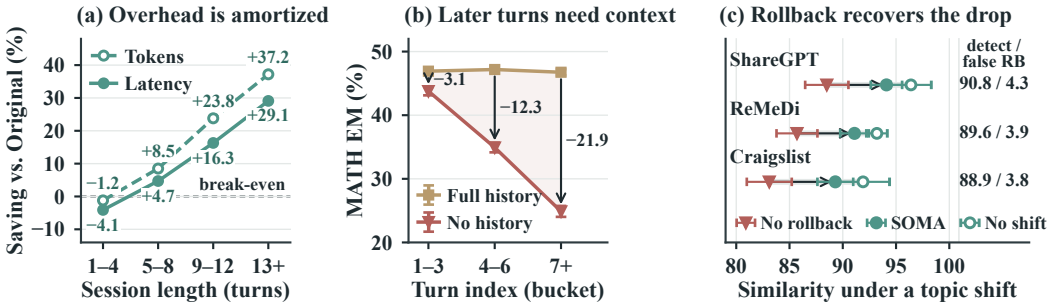
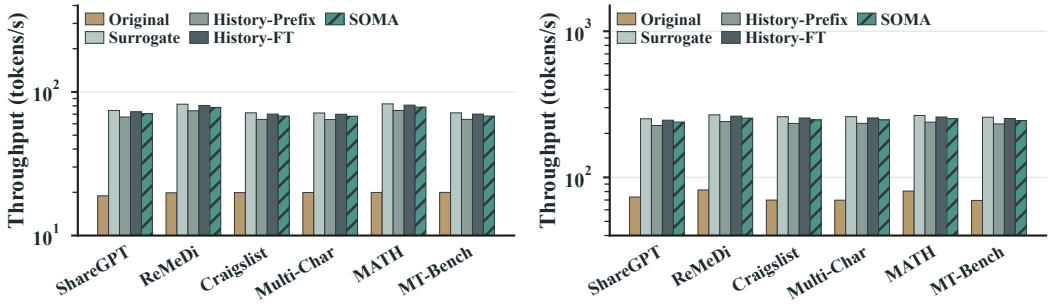


Figure 3: Operating range and reliability. (a) End-to-end saving of SOMA over Original on ShareGPT after all warm-start and adaptation overhead, by session length. (b) MATH exact match with and without the dialogue history: later turns remain context-dependent. (c) Response similarity under an abrupt topic shift without rollback, with SOMA's drift-aware rollback, and without a shift; the right column lists SOMA's shift-detection and false-rollback (RB) rates in %.



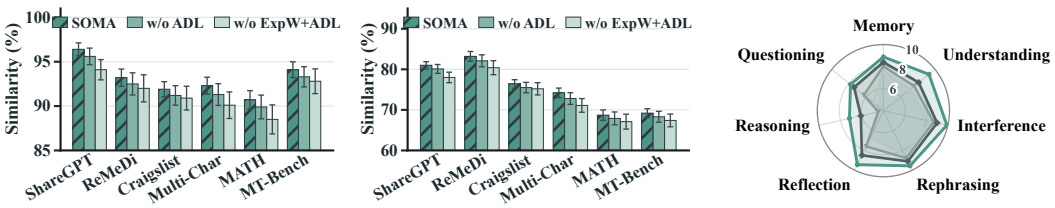
(a) LLaMA family. (b) Qwen family.

Figure 4: Throughput across six datasets. SOMA keeps the serving path close to the surrogate while maintaining higher similarity to the original model.

unfavorable for sessions of 1–4 turns (-4.1% latency) and roughly breaks even at 5–8 turns ($+4.7\%$). From 9 turns on, it saves 16.3% to 29.1% latency and 23.8% to 37.2% tokens. This matches SOMA's intended use case as an amortized serving strategy for conversations where enough later turns remain local to the warm-start state. Figure 4 reports throughput on all six datasets for both families. SOMA runs at 95–97% of the throughput of Surrogate and History-FT, about 5.6% faster than History-Prefix and 3.1–3.9$\times$ faster than Original. Appendix F.1 lists the per-dataset token costs.

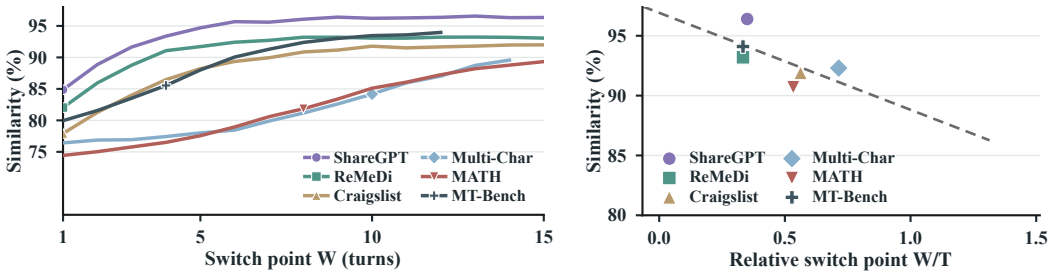
RQ4: Component contributions. We first ablate SOMA on the LLaMA family to identify which components drive the gains. Figure 5a shows that full SOMA performs best across all datasets. Removing the anti-degeneration loss consistently reduces performance, showing that entropy regularization keeps prompt mining stable and prevents collapsed, low-diversity soft prompts. Removing both the expectation-weighted term and the anti-degeneration loss leads to a larger drop, especially on harder datasets. This result suggests that the anchoring scores are most informative when the probe is both stable and semantic. Figure 5b repeats the ablation on the Qwen family and finds the same ordering, so the conclusion does not appear to hinge on the backbone. Figure 5c breaks MT-Bench-101 (Bai et al., 2024) down by ability for the LLaMA family. Relative to the surrogate, SOMA gains most on reasoning, reflection, and questioning, and it also improves understanding, interference handling, memory, and rephrasing to a smaller degree.

RQ5: Reliability and fallback. We next test when SOMA's locality assumption holds and how rollback helps when it fails. Figure 3(b) shows that full-history MATH performance stays stable across turn buckets, while no-history performance drops sharply in later turns, from a 3.1-point EM loss on turns 1–3 to a 21.9-point loss on turns 7 and later. This means later turns are not simply easier; they remain strongly dependent on the earlier dialogue context. Figure 3(c) then evaluates abrupt topic shifts. Without rollback, similarity drops substantially. With SOMA's drift-aware rollback, similarity returns to within 2.1–2.6 points of the clean setting (e.g., $88.5 \rightarrow 94.1$ vs. 96.4 on ShareGPT), with 88.9–90.8% of shifts detected and 3.8–4.3% false rollbacks. These results show that SOMA should



(a) LLaMA ablation. (b) Qwen ablation. (c) Ability profile.

Figure 5: Component and capability analysis. Anti-degeneration regularization and expectation-weighted divergence both improve mining on LLaMA (a) and Qwen (b). (c) MT-Bench-101 ability profile of **Surrogate** ▲, **History-Prefix** ▼, **History-FT** ◆, and **SOMA** ●.



(a) Similarity under different switch points. (b) Relative switch point vs. final similarity.

Figure 6: Warm-start and switching behavior. Easier local-dialogue settings switch earlier, while reasoning-heavy or multi-party settings need longer warm starts and achieve lower final similarity.

be viewed as a local serving strategy with automatic fallback. Figure 6 shows the warm-start side. Goal-anchored dialogues (ShareGPT, ReMeDi, Craigslist) reach high similarity after a few turns, while reasoning-heavy or multi-party ones (MATH, Multi-Char) need longer warm starts and level off lower. The best switch point thus seems data-dependent, and the gate picks it per session.

6 Conclusion

We presented SOMA, an efficient multi-turn LLM serving framework that uses early dialogue turns to adapt a small surrogate to the original model's session-local behavior. It probes the surrogate with adversarial soft prompts, distills the robustly anchored warm-start turns with LoRA, and switches once the alignment looks sufficient. Experiments show that SOMA better preserves original-model responses, improves task-grounded quality, and reduces serving cost once warm-start overhead is amortized. These results suggest that session-local surrogate serving is a practical direction for efficient multi-turn LLM deployment when sessions are long and coherent enough to amortize the warm-start cost, as Figure 3(a) indicates. Since SOMA only uses the text outputs of the original model, it should also apply when F is a proprietary model behind an API, as long as the surrogate's embedding layer is available for mining. In practice, SOMA seems most useful for sessions that stay on one task for more than a handful of turns. For short or drifting sessions the gate mostly keeps the original model in charge, which should limit the loss in quality. Two questions remain open, namely how SOMA behaves under gradual topic drift and how it compares with divergence-weighted fine-tuning at the same LoRA budget. We discuss these and other limitations in Appendix I.

AI use statement

We used generative AI tools to polish the writing and to help with parts of the evaluation and plotting code. The research itself is the authors' own work, and the authors checked all AI-assisted text and code and take full responsibility for this paper.

Ethics statement

This work uses public benchmarks under their licenses and involves no human subjects or new personal data. A smaller surrogate may lower response quality when a conversation drifts, especially

in high-stakes domains, so deployments should keep monitoring, conservative thresholds, and privacy-preserving handling of dialogue history (Appendices H and I).

Reproducibility statement

An anonymized implementation of SOMA with scripts for the main experiments is available at `https://anonymous.4open.science/r/SOMA-4175`. Appendix C.1 and the following subsections describe the datasets, filtering, baselines, hyperparameters, judge protocol, and software versions, and Appendix E states the assumptions and proofs of all theoretical results.

References

agentlans. Multi-character dialogue dataset. Hugging Face Dataset, 2024. URL `https://huggingface.co/datasets/agentlans/multi-character-dialogue`.

Anthropic. Introducing Claude. Anthropic Blog, March 2023. URL `https://www.anthropic.com/news/introducing-claude`.

Anthropic. Prompt caching with Claude. `https://www.anthropic.com/news/prompt-caching`, August 2024.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, 2024.

Adarsh Prasad Behera, Jaya Prakash Champati, Roberto Morabito, Sasu Tarkoma, and James Gross. Towards efficient Multi-LLM inference: Characterization and analysis of LLM routing and hierarchical techniques. *arXiv preprint arXiv:2506.06579*, 2025.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 755–773, 2025.

Xueqi Cheng and Yushun Dong. LatentRouter: Can we choose the right multimodal model before seeing its answer? *arXiv preprint arXiv:2605.11301*, 2026.

Xueqi Cheng, Xugui Zhou, Tyler Derr, and Yushun Dong. ReAD: Reinforcement-guided capability distillation for large language models. *arXiv preprint arXiv:2605.11290*, 2026.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *The Twelfth International Conference on Learning Representations*, 2024.

Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. Towards next-generation intelligent assistants leveraging LLM techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5792–5793, 2023.

Bin Gao, Zhuomin He, Puru Sharma, Qingxuan Kang, Djordje Jevdjic, Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo. Cost-Efficient large language model serving for multi-turn conversations with CachedAttention. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pp. 111–126, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569, 2024.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343, 2018.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004.

Jinwoo Jeong and Jeongseob Ahn. Accelerating LLM serving for multi-turn dialogues with efficient resource management. In *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pp. 1–15, 2025.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Haoyang Li, Zhanchao Xu, Yiming Li, Xuejia Chen, Darian Li, Anxin Tian, Qingfa Xiao, Cheng Deng, Jun Wang, Qing Li, et al. LoopServe: An adaptive dual-phase LLM inference acceleration system for multi-turn dialogues. *arXiv preprint arXiv:2507.13681*, 2025a.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. *Advances in Neural Information Processing Systems*, 36:72888–72903, 2023.

Lincan Li, Zheng Chen, and Yushun Dong. LLM as clinical graph structure refiner: Enhancing representation learning in EEG seizure diagnosis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6770–6778, 2026.

Yubo Li, Xiaobin Shen, Yidi Miao, Xinyu Yao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025b.

Fang Liu, Yang Liu, Lin Shi, Zhen Yang, Li Zhang, Xiaoli Lian, Zhongqi Li, and Yuchi Ma. Beyond functional correctness: Exploring hallucinations in LLM-generated code. *IEEE Transactions on Software Engineering*, 52(3):1037–1055, 2026.

Xiang Liu, Hong Chen, Xuming Hu, and Xiaowen Chu. FlowKV: Enhancing multi-turn conversational coherence in LLMs via isolated key-value cache management. *arXiv preprint arXiv:2505.15347*, 2025.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

Eric Melz. Enhancing LLM intelligence with ARM-RAG: Auxiliary rationale memory for retrieval augmented generation. *arXiv preprint arXiv:2311.04177*, 2023.

Dilxat Muhtar, Yelong Shen, Yaming Yang, Xiaodong Liu, Yadong Lu, Jianfeng Liu, Yuefeng Zhan, Hao Sun, Weiwei Deng, Feng Sun, et al. StreamAdapter: Efficient test time adaptation from contextual streams. *arXiv preprint arXiv:2411.09289*, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL `https://arxiv.org/abs/2508.10925`.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024.

Soya Park, Hari Subramonyam, and Chinmay Kulkarni. Thinking assistants: LLM-based conversational assistants that help users think by asking rather than answering. *arXiv preprint arXiv:2312.06024*, 2023.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68658–68685, 2024.

ShareGPT. ShareGPT: Share your wildest ChatGPT conversations with one click. `https://sharegpt.com/`, n.d.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-LoRA: Serving thousands of concurrent LoRA adapters. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 296–311, 2024.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. In *First Conference on Language Modeling*, 2024.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3): 339–373, 2000.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639: 130193, 2025.

Yan Wang, Dongyang Ma, and Deng Cai. With greater text comes greater necessity: Inference-time training helps long text generation. In *First Conference on Language Modeling*, 2024.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. ReMeDi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3013–3024, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in LLM-based multi-turn dialogue systems. *ACM Computing Surveys*, 58(6): 1–38, 2025.

Dahai Yu, Lin Jiang, Rongchao Xu, and Guang Wang. HealthMamba: An uncertainty-aware spatiotemporal graph state space model for effective and reliable healthcare facility visit prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7001–7009, 2026.

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-R1: Teaching LLMs multi-round routing and aggregation via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 141233–141265, 2025.

Tianhao Zhang, Jie Yang, Deli Zhao, and Xinliang Ge. Linear local tangent space alignment and application to face recognition. *Neurocomputing*, 70(7-9):1547–1553, 2007.

Yuying Zhao, Yu Wang, Xueqi Cheng, Anne Marie Tumlin, Yunchao Liu, Damin Xia, Meng Jiang, and Tyler Derr. Amplifying your social media presence: Personalized influential content generation with LLMs. *arXiv preprint arXiv:2505.01698*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

APPENDIX

A NOTATIONS

This section summarizes all notations used throughout this paper.

Table 4: Notation summary.

Symbol	Meaning
$\mathcal{D}_k$	Dialogue prefix $\{(q_t, a_t)\}_{t=1}^{k}$
$q_{\le t}$	Full history up to turn t
F, G	Original (large) model; surrogate (small) model
S_t	Dialogue state $\mathcal{D}_{t-1} \oplus q_t$ at turn t
$\mathcal{Q}_k$	Distribution of later dialogue states after $\mathcal{D}_k$ (local region)
$\mathcal{V}$, E	G's vocabulary; embedding matrix $E \in \mathbb{R}^{d\times\lvert\mathcal{V}\rvert}$
Π_t	Next-token distribution of G at turn t; entry $\Pi_t[v]$
$\mathbf{e}_v$	Embedding of token v (column of E)
$\mathbf{P} \in \mathbb{R}^{L\times d}$	Soft prompt (row-wise prompt tokens in embedding space)
$N_k(u)$	k nearest neighbors of token u by cosine in E
$\overline{\mathbf{e}}_{t,i}$	Expected embedding $\overline{\mathbf{e}}_{t,i} = E^\top \Pi_{t,i}$ at position i
$w_{t,i}$	Expectation-weighted factor for semantic divergence loss
$\mathsf{Gap}_G(S)$	Response discrepancy $1 - \cos(\phi(a^G(S)), \phi(a^F(S))) \in [0, 1]$
r_t	Anchoring score of warm-start turn t (residual mining loss)
$\lvert B\rvert_{\text{eff}}$	Effective batch size $\lvert B\rvert/(m+1)$ under m-dependence
$\widehat{\mathsf{G}}_B$	Average gap of the adapted surrogate on a held-out acceptance batch

B RELATED WORK

Efficient LLM for Multi-turn Dialogues. Recent efforts to improve multi-turn LLM serving efficiency fall into two main paradigms. The first involves **single-model approaches** that reduce context length or reuse computation. These include *summarization and context compression* (Wang et al., 2025; Chen et al., 2025; Xiao et al., 2024), *memory augmentation* (Melz, 2023; Gutiérrez et al., 2024), and *caching and attention reuse*, such as CachedAttention (Gao et al., 2024), prompt caching (Anthropic, 2024), resource-aware multi-turn scheduling (Jeong & Ahn, 2025), isolated per-turn KV compression in FlowKV (Liu et al., 2025), and adaptive prefill sparsification with progressive KV compression in LoopServe (Li et al., 2025a). While effective in lowering per-turn cost, these methods still rely on repeated large-model inference, which also leads to high API expenses, latency, and GPU demand. Moreover, they may truncate or underutilize dialogue context, harming performance on complex tasks. The second paradigm adopts **multi-model approach**, using smaller models for simple queries and escalating difficult ones to larger LLMs (Behera et al., 2025; Ding et al., 2024), typically via model routing (Shnitzer et al., 2024; Ong et al., 2025; Cheng & Dong, 2026), multi-round routing and aggregation learned with reinforcement learning (Zhang et al., 2025), and distillation (Hinton et al., 2015; Cheng et al., 2026). However, pre-trained small models often generalize poorly on complex multi-turn dialogues, while switching models adds inefficiency.

Relation to systems-level serving and test-time adaptation. Prefix and KV caching (CachedAttention, prompt caching, FlowKV, LoopServe) reduce the prefill cost of re-processing the history on the *same* large model. They leave the per-token decode cost, memory footprint, and API price of that model unchanged. SOMA works at a different level. After the switch it takes the large model off the serving path, and the surrogate can itself use prefix caching. Our efficiency comparison already serves every method, including Original, through vLLM with automatic prefix caching (Section 5), so the reported token and latency gains come on top of KV-cache reuse. Test-time adaptation methods such as Temp-LoRA (Wang et al., 2024) and StreamAdapter (Muhtar et al., 2024) adapt a model to its *own* generated or in-context text. History-FT and SOMA perform cross-model distillation, adapting the surrogate to the large model's warm-start responses within the session.

Local Manifold Approximation. We use "local manifold approximation" in the informal sense of classical local methods such as LLE (Roweis & Saul, 2000), LTSA (Zhang et al., 2007), and Laplacian eigenmaps (Belkin & Niyogi, 2003). A dialogue prefix induces a local region of dialogue states, and SOMA approximates the large model only inside that region. It computes no intrinsic dimensions, tangent spaces, or geodesics, and it does not match the hidden spaces of the two models. The local view enters the method through two components, the output-space discrepancy of Problem 2.1, which compares responses through a shared sentence encoder, and the locality gate. Its measurable prediction is that warm-start supervision carries over to later turns until the topic shifts, and we test this in Section 5 and Appendix F.2.

C Reproducibility

In this section, we introduce the details of the experiments in this paper for reproducibility. At the same time, we have uploaded all necessary code to our GitHub repository to reproduce the results presented in this paper: `https://anonymous.4open.science/r/SOMA-4175`. All major experiments are encapsulated as shell scripts, which can be conveniently executed. We introduce the details for reproducibility in the subsections below.

C.1 Real-World Datasets

In this section, we briefly present the real-world datasets used in this paper, and all these datasets are commonly used datasets in multi-turn conversation tasks. *ShareGPT* (ShareGPT, n.d.) is a large collection of real conversations between users and ChatGPT that users shared publicly. *ReMeDi* (Yan et al., 2022) is a multi-domain Chinese medical dialogue corpus of doctor–patient conversations. *Craigslist* (He et al., 2018) contains multi-turn buyer–seller negotiation chats from Craigslist, enabling study of bargaining strategies and goal-directed dialogue. *Multi-Char* (agentlans, 2024) provides multi-character conversational scenarios with role specifications to evaluate coordination and role consistency in multi-party dialogue. *MATH* (Hendrycks et al., 2021) is a benchmark of competition-style math problems with step-by-step solutions designed to assess mathematical reasoning in language models. *MT-Bench* (Zheng et al., 2023) is a multi-turn benchmark to assess response quality across diverse tasks. In this study, we filter out the dialogues in these datasets whose later turns do not depend on the earlier context.

C.2 Implementation of SOMA

We implement SOMA based on PyTorch with HuggingFace Transformers, serve inference via vLLM with FlashAttention (Shah et al., 2024), and run on Nvidia A6000 GPUs. Soft–prompt tuning optimizes only the prompt tensor $\mathbf{P} \in \mathbb{R}^{L \times d}$ on the surrogate G using AdamW, cosine decay, gradient clipping, and a single forward pass per warm-start turn and step. The objective combines unlikelihood on a semantic neighborhood, an expectation–weighted penalty using the truncated top–m expectation, and a light anti-degeneration entropy term. The warm-start context–response pairs, weighted by their anchoring scores r_t, are then used to adapt G with LoRA on attention and MLP projections (rank r), keeping the base weights frozen; early stopping is triggered when the training loss stops improving, since all warm-start turns are used for adaptation. At inference, a cosine closeness gate with a lightweight sentence encoder decides a one–time switch to the adapted surrogate. We set all hyperparameters below, including W, $|B|$, M, and (k, m), empirically within the listed ranges. The bounds in Section 4 indicate how the required margins scale, but at these sample sizes they are too loose to set the values directly. Prompt length $L \in \{4, 8, 16, 32, 64\}$; learning rate $\eta \in [1 \times 10^{-4},\ 5 \times 10^{-3}]$; AdamW weight decay $\lambda_{\text{wd}} \in [0,\ 10^{-2}]$; gradient clip $\in\ [0.5,\ 1.0]$; neighborhood size $k \in \{20, 50, 100\}$; expectation truncation $m \in \{50, 100, 200\}$; temperature $\tau \in [0.4,\ 1.2]$; expectation weight $\lambda \in [0.5,\ 2.0]$; anti–degeneration weight $\beta \in [0.02,\ 0.15]$; LoRA rank $r \in \{4, 8, 16, 32\}$ with scale $\alpha_{\text{lora}} \in \{4, 8, 16, 32\}$; LoRA LR $\eta_{\text{lora}} \in [5 \times 10^{-5},\ 2 \times 10^{-3}]$; warm–start window $W \in [3,\ 12]$ turns; acceptance batch $|B| \in [6,\ 24]$ contexts; parallel candidates $M \in \{3, 4, 5\}$; switch threshold $\varepsilon \in [0.05,\ 0.12]$ with $m_{\text{cons}} \in \{2, 3\}$.

You are an expert conversation analyst.

Task – Classification:
* **context-dependent**: The response of every turn relies on information from earlier turns (the conversation is a single chain of related reasoning steps).
* **context-independent**: The response of every turn is independent of earlier turns (the conversation is a collection of independent questions and answers).

Conversation (verbatim):
{dialogue}

Question: Is the conversation context-dependent or context-independent?
Please answer with exactly one word:
Either **context-dependent** or **context-independent** — nothing else.

Figure 7: Instructions for the LLM to filter context-dependent dialogue.

C.3 Implementation of Baselines

Original: We query the large model (family-appropriate: LLaMA-3.1-70B or Qwen-3-8B) with the full dialogue history at every turn to get the responses.

Surrogate: We query the small model (LLaMA-2-7B or Qwen-3-0.6B) with the full dialogue history at every turn to get the responses.

History-Prefix: The surrogate receives the entire conversation produced by the Original up to $t-1$ and generates a_t; no parameter updates are performed. The switching is the same as SOMA.

History-FT: We fine-tune the surrogate on (S_t, a_t^F) pairs where S_t is the Original's full context up to t and a_t^F is the next reply, using LoRA on attention projections with early stopping; inference then runs the fine-tuned surrogate without the Original. The switching is the same as SOMA.

LLMLingua-2 (Pan et al., 2024): We compress the Original's history at each turn with LLMLingua-2 and feed the compressed summary to the surrogate; we follow the authors' recommended settings.

RouteLLM (Ong et al., 2025): We adopt the released router to choose between the Original and the surrogate per turn (complex vs. simple queries), following the original settings from the paper.

C.4 Data Filtering

To ensure that SOMA is evaluated on context-dependent dialogues, we prompt multiple strong LLMs, including GPT-OSS, DeepSeek-V3, and Gemma-2-27B, with the classification prompt shown in Figure 7, and retain only conversations that all three models recognize as context-dependent. This avoids model-specific biases and yields a high-quality subset of dialogues where later turns meaningfully depend on earlier turns, matching the use cases in this paper.

C.5 Implementation of LLM Judge

The prompt for the LLM judge is shown in Figure 8. It adapts the single-answer grading protocol of MT-Bench (Zheng et al., 2023) and MT-Bench-101 (Bai et al., 2024) to reference-based similarity

You are an impartial evaluator.
You are evaluating the AI assistant's capabilities in solving a problem posed by 'Human'.
Assess the assistant's performance based on the following criteria:

1. Accuracy: Verify the correctness of the AI assistant's answer against the ground truth (reference solution).
2. Reasoning: Assess the completeness, clarity, and logical soundness of the step-by-step reasoning process.
3. Context Integration: Consider whether the assistant integrates relevant prior dialogue or context that influences the solution.
4. Communication: Appraise how clearly and effectively the assistant communicates its reasoning to aid understanding.

Score Guidelines (0.0 to 1.0):
- 1.0: Completely correct answer and meticulously clear, step-by-step reasoning that is logically sound and instructional.
- 0.7–0.9: Correct answer with a well-articulated reasoning process that includes the necessary steps and promotes understanding.
- 0.4–0.6: Partially correct answer with some minor reasoning flaws or omissions.
- 0.0–0.3: Incorrect answer and/or poor reasoning that lacks clarity or logic.

Only respond with a single score between 0.0 and 1.0. Do not include any explanation.

Reference Solution:
{Original Response}
Candidate Answer:
{Surrogate Response}

Score:

Figure 8: Instructions for the LLM judge to evaluate the response similarity.

scoring. The judge receives the original model's response as the reference and scores the surrogate's response on a 0–1 scale along accuracy, reasoning, context integration, and communication. The reference and the 0–1 scale are our modifications, since the MT-Bench protocols grade single answers on a 1–10 scale without a reference. Three judges (GPT-OSS, DeepSeek-V3, and Gemma-2-27B) score each response, and we report the mean.

C.6 Packages Required for Implementation

We perform all experiments on a server equipped with Nvidia A6000 GPUs. Below we list the key packages and their versions used in our implementation:

- **Python** == 3.10
- **pytorch** == 2.8.0 + CUDA 12.8
- **torchvision** == 0.19.0
- **torchaudio** == 2.8.0
- **numpy** == 1.26.x
- **pandas** == 2.2.x
- **scipy** == 1.12.x
- **cmake** == 3.28+
- **ninja** == 1.11+
- **ipython** == 8.x
- **psutil** == 5.9+
- **vllm** == 0.10.2
- **transformers** == 4.56.1

- **accelerate** == 1.9.0
- **bitsandbytes** == 0.46.1
- **sentencepiece** == 0.2.0
- **tiktoken** == 0.11.0
- **einops** == 0.8.1
- **datasets** == 4.0.0
- **huggingface-hub** == 0.34.2
- **safetensors** == 0.5.3
- **ray** == 2.49.1
- **scikit-learn** == 1.7.1
- **fastapi** == 0.116.2
- **uvicorn** == 0.35.0

D MATHEMATICAL CONCEPTS

Here we provide a more comprehensive view of the relevant concepts that can be helpful to understand the idea of stratification as discussed in the main body of the paper.

Definition D.1 (Preimage)**.** Let $f : X \mapsto Y$ be a function from a set X (*domain*) to a set Y (*codomain*). For any subset $N \subseteq Y$, the *preimage* of N under f, denoted $f^{-1}(N)$, is defined as:

$$f^{-1}(N) = \{x \in X \mid f(x) \in N\}.$$

In other words, $f^{-1}(N)$ consists of all elements in the domain X that are mapped into the subset N of the codomain Y, regardless of whether every element of N has a preimage.

Definition D.2 (Metric Space)**.** A set X, whose elements are called points, is said to be a *metric space* if for any two points $p, q \in X$, there is an associated real number $d(p, q)$, called the *distance* from p to q, such that the following three conditions hold:

1. $d(p, q) \geq 0$, and $d(p, q) = 0 \iff p = q$;
2. $d(p, q) = d(q, p)$ (symmetry);
3. $d(p, q) \leq d(p, r) + d(r, q)$ for any $r \in X$ (triangle inequality).

Any function satisfying these properties is called a *distance function*, or a *metric*.

Definition D.3 (Neighborhood)**.** Let X be a metric space. A set $N_r(p) \subset X$ is called a *neighborhood* of a point $p \in X$ if it consists of all points $q \in X$ such that $d(p, q) < r$ for some radius $r > 0$. The number r is called the *radius* of the neighborhood.

Definition D.4 (Continuous)**.** Let X and Y be two topological spaces. A function $f : X \mapsto Y$ is *continuous* if for each point $x \in X$ and each neighborhood N of $f(x)$ in Y, the set $f^{-1}(N)$ is a neighborhood of the point $x \in X$ in the domain.

Definition D.5 (Topological Equivalence or Homeomorphism)**.** A function $h : X \mapsto Y$ is called a *homeomorphism* if it is one-to-one, continuous, and has a continuous inverse function. When such a function exists, X and Y are called *homeomorphic* (or *topologically equivalent*) spaces.

Definition D.6 (Open Set)**.** A subset $U \subseteq X$ is called an *open set* if for every point $p \in U$, there exists a neighborhood $N_r(p) \subseteq U$. That is, each point in U has some "wiggle room" around it that still lies entirely within the set U itself.

Definition D.7 (Countable Base (Second Countability))**.** Let X be a topological space. A collection $\mathcal{B}$ of open subsets of X is called a *base* (or *basis*) for the topology on X if for every open set $U \subseteq X$ and every point $x \in U$, there exists a set $B \in \mathcal{B}$ such that

$$x \in B \subseteq U.$$

If there exists a base $\mathcal{B}$ that is *countable*, then the space X is said to be *second countable* or to have a *countable base*; Euclidean spaces, with balls of rational radii around rational points, are a standard example of a second countable space.

Definition D.8 (Hausdorff Space)**.** A topological space with the property that two distinct points can always be surrounded by disjoint open sets is called a *Hausdorff space*.

Essentially, Hausdorff spaces are the spaces where any two points being "far off" is defined.

Definition D.9 (Manifold)**.** A *manifold* of dimension n is a second-countable Hausdorff topological space in which each point has a neighborhood homeomorphic to Euclidean space $\mathbb{R}^n$.

Definition D.10 (Smooth Manifold)**.** A *smooth manifold* is a manifold $\mathcal{M}$ equipped with a collection of coordinate charts (i.e., homeomorphisms $\varphi : U \to \mathbb{R}^n$ for open sets $U \subseteq \mathcal{M}$) such that all *transition maps* between overlapping charts,

$$\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j),$$

are infinitely differentiable (i.e., C^∞). This structure is known as a *smooth atlas*, and it allows calculus to be performed on the manifold.

E Theoretical Results and Proofs

This appendix states and proves the results used in Section 4. Each statement lists its assumptions, and Section 4.1 discusses where our implementation departs from them.

Dependence and effective sample size. Consecutive dialogue turns are not independent. We assume that the per-turn statistics $Z_1, \dots, Z_N$ form an m-dependent sequence, that is, $\{Z_s\}_{s \le t}$ and $\{Z_s\}_{s > t+m}$ are independent for every t, and we define the effective sample size $N_{\text{eff}} = N/(m+1)$, with $m = 0$ for independent samples. For such sequences Hoeffding's inequality holds with N replaced by N_{eff}. If each Z_s takes values in an interval of length R, then for every $u > 0$

$$\Pr\Big(\frac{1}{N}\sum_{s=1}^{N}\big(Z_s - \mathbb{E}Z_s\big) \ge u\Big) \le \exp\Big(-\frac{2N_{\text{eff}}\,u^2}{R^2}\Big),$$

and the same bound holds for the lower tail. This follows from Janson's inequality for sums of partly dependent variables (Janson, 2004), applied to the dependency graph of the sequence, whose fractional chromatic number is at most $m+1$.

Locality premise. The results below compare statistics measured on warm-start turns with their expectations over the later turns of the session. We therefore assume, as in Section 2, that warm-start states and later states share the marginal distribution $\mathcal{Q}_k$. The locality gate monitors this premise, and Figures 3(c) and 6 test it. Without it, the bounds describe only the warm-start distribution.

Switching Criterion: Warm-Start Concentration and Acceptance

Lemma E.1 (Warm-Start Concentration)**.** *Let $h(S) \in [0,1]$ be a local statistic that is fixed before the window is observed, such as $\mathsf{Gap}_\Theta(S)$ for a fixed adapter Θ or the distance of a query to a fixed centroid. Let $h(S_1), \dots, h(S_W)$ be m-dependent with common mean $h^\star = \mathbb{E}_{S\sim\mathcal{Q}_k}[h(S)]$, and write $\hat{h}_W = \frac{1}{W}\sum_{t=1}^{W} h(S_t)$. Then, with probability at least $1-\delta$,*

$$\big|\hat{h}_W - h^\star\big| \le \eta_W(\delta) := \sqrt{\frac{\log(2/\delta)}{2W_{\text{eff}}}}.$$

Proof of Lemma E.1. Apply the m-dependent Hoeffding inequality with $R = 1$ to each tail at level $\delta/2$. Setting $\exp(-2W_{\text{eff}}u^2) = \delta/2$ gives $u = \eta_W(\delta)$, and a union bound over the two tails gives the claim. □

Statement: Proposition 4.1. Let the adapter Θ be fixed given the adaptation data, and let the acceptance batch $B = \{S_1, \dots, S_{|B|}\}$ be disjoint from that data and drawn from $\mathcal{Q}_k$ as an m-dependent sequence. Let $\widehat{\mathsf{G}}_B = \frac{1}{|B|}\sum_j \mathsf{Gap}_\Theta(S_j)$ and $\mathsf{G}^\star = \mathbb{E}_{S\sim\mathcal{Q}_k}[\mathsf{Gap}_\Theta(S)]$. Then, with probability at least $1-\delta$,

$$\mathsf{G}^\star \le \widehat{\mathsf{G}}_B + \gamma_B(\delta), \qquad \gamma_B(\delta) = \sqrt{\frac{\log(1/\delta)}{2|B|_{\text{eff}}}}.$$

In particular, if $\widehat{\mathsf{G}}_B \le \varepsilon - \gamma_B(\delta)$, then $\mathsf{G}^\star \le \varepsilon$ with probability at least $1-\delta$.

Proof of Proposition 4.1. Condition on the adaptation data. Because B is disjoint from it, Θ is fixed, and $\mathsf{Gap}_\Theta(S_1), \ldots, \mathsf{Gap}_\Theta(S_{|B|})$ are m-dependent with mean $\mathsf{G}^\star$ and range $R = 1$. The lower-tail inequality gives $\Pr(\widehat{\mathsf{G}}_B - \mathsf{G}^\star \le -u) \le \exp(-2|B|_{\text{eff}} u^2)$, and setting the right-hand side to δ gives $u = \gamma_B(\delta)$. On the complementary event, $\mathsf{G}^\star < \widehat{\mathsf{G}}_B + \gamma_B(\delta) \le \varepsilon$ if $\widehat{\mathsf{G}}_B \le \varepsilon - \gamma_B(\delta)$. □

When the acceptance batch is reused. If B overlaps the adaptation data, Θ depends on the S_j in B. The gaps then no longer have mean $\mathsf{G}^\star$ given Θ, $\widehat{\mathsf{G}}_B$ is biased downward by the fit, and Proposition 4.1 does not apply. Our implementation reuses the warm-start turns in this way for the initial switch (Appendix G). Holding out part of the warm-start window would restore the bound, at the cost of fewer adaptation examples. The bound is also loose at the sample sizes used here. For $\delta = 0.05$, $m = 0$, and $|B| = 24$, $\gamma_B = 0.25$, which already exceeds the thresholds $\varepsilon \in [0.05, 0.12]$, so the bound mainly shows how the required margin scales with $|B|_{\text{eff}}$.

Corollary E.2 (Switching Rule)**.** *Let the locality statistic h satisfy the conditions of Lemma E.1, and let B satisfy those of Proposition 4.1. If $\widehat{\mathsf{G}}_B \le \varepsilon - \gamma_B(\delta)$ and $\hat{h}_W \le \tau - \eta_W(\delta)$, then, with probability at least $1 - 2\delta$, both $\mathsf{G}^\star \le \varepsilon$ and $h^\star \le \tau$.*

Proof of Corollary E.2. By Lemma E.1, $h^\star \le \hat{h}_W + \eta_W(\delta) \le \tau$ with probability at least $1 - \delta$. By Proposition 4.1, $\mathsf{G}^\star \le \varepsilon$ with probability at least $1 - \delta$. A union bound shows that both hold with probability at least $1 - 2\delta$. □

Coverage and Suboptimality to Determine the Number of Soft-Prompt Candidates

Assume that the mining objective is locally quadratic with curvature $\mathbf{H}_T \succeq 0$ whose top eigenvector $\boldsymbol{v}_1$ lies in an r_{act}-dimensional active subspace. SOMA draws its M candidate prompts from an isotropic Gaussian, so their projections onto this subspace point in independent, uniformly random directions. The results below bound the probability that one of them starts within angle θ of $\boldsymbol{v}_1$, and the fraction of the top curvature that such a direction retains. These results cover only the initial directions of the candidates, and we set $M \in \{3, 4, 5\}$ empirically (Appendix C.2).

Proposition E.3 (Coverage of the Best Local Direction)**.** *Let $r_{\text{act}} \ge 2$, let $\boldsymbol{v}_1$ be a unit vector in an r_{act}-dimensional subspace, and draw $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ i.i.d. from $\mathrm{Unif}(\mathbb{S}^{r_{\text{act}}-1})$ in that subspace. For $\theta \in (0, \pi/2]$, let*

$$p_{r_{\text{act}}}(\theta) = \Pr\left(\angle(\boldsymbol{u}_1, \boldsymbol{v}_1) \le \theta\right) = \tfrac{1}{2}\, I_{\sin^2\theta}\Big(\tfrac{r_{\text{act}}-1}{2}, \tfrac{1}{2}\Big),$$

where $I_x(\cdot,\cdot)$ is the regularized incomplete beta function. Then

$$\Pr\Big(\min_{m \le M} \angle(\boldsymbol{u}_m, \boldsymbol{v}_1) \le \theta\Big) = 1 - \big(1 - p_{r_{\text{act}}}(\theta)\big)^M, \qquad p_{r_{\text{act}}}(\theta) \ge \frac{(\sin\theta)^{r_{\text{act}}-1}}{(r_{\text{act}}-1)\,\mathrm{B}\big(\frac{r_{\text{act}}-1}{2}, \frac{1}{2}\big)}.$$

For example, $p_{r_{\text{act}}}(\pi/2) = 1/2$ for every r_{act}. Isotropic Gaussian initializations satisfy the sampling assumption, because the normalized projection of $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ onto a fixed subspace is uniformly distributed on the unit sphere of that subspace.

Proof of Proposition E.3. **Step 1: Exact cap probability.** Write $r = r_{\text{act}}$ and $T = \langle \boldsymbol{u}_1, \boldsymbol{v}_1 \rangle$. For $\boldsymbol{u}_1$ uniform on $\mathbb{S}^{r-1}$, T is symmetric about 0 and $T^2 \sim \mathrm{Beta}(\frac{1}{2}, \frac{r-1}{2})$, so $1 - T^2 \sim \mathrm{Beta}(\frac{r-1}{2}, \frac{1}{2})$. For $\theta \le \pi/2$ we have $\cos\theta \ge 0$, hence

$$\Pr(T \ge \cos\theta) = \tfrac{1}{2}\Pr(T^2 \ge \cos^2\theta) = \tfrac{1}{2}\Pr(1 - T^2 \le \sin^2\theta) = \tfrac{1}{2}\, I_{\sin^2\theta}\big(\tfrac{r-1}{2}, \tfrac{1}{2}\big).$$

Step 2: Lower bound. Let $a = (r-1)/2$ and $x = \sin^2\theta$. Since $(1-s)^{-1/2} \ge 1$ on $[0, 1)$,

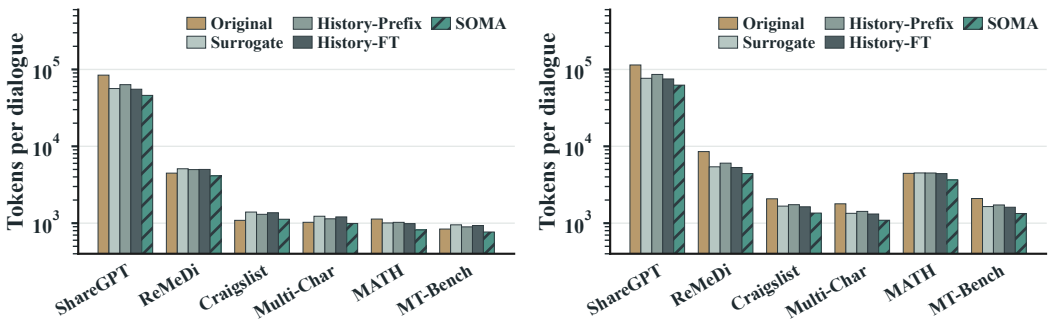
$$I_x\big(a, \tfrac{1}{2}\big) = \frac{1}{\mathrm{B}(a, \frac{1}{2})} \int_0^x s^{a-1}(1-s)^{-1/2}\,\mathrm{d}s \;\ge\; \frac{1}{\mathrm{B}(a, \frac{1}{2})} \int_0^x s^{a-1}\,\mathrm{d}s = \frac{x^a}{a\,\mathrm{B}(a, \frac{1}{2})},$$

so $p_r(\theta) \ge x^a / \big(2a\,\mathrm{B}(a, \frac{1}{2})\big) = (\sin\theta)^{r-1} / \big((r-1)\,\mathrm{B}(\frac{r-1}{2}, \frac{1}{2})\big)$.

Step 3: Independence across M samples. The M draws are independent, so none of them falls in the cap with probability $(1 - p_r(\theta))^M$, which gives the coverage probability. □

Lemma E.4 (Directional Suboptimality)**.** *Let $\mathbf{H}_T \succeq 0$ have largest eigenvalue λ_1 and top eigenvector $\boldsymbol{v}_1$. For any unit $\hat{\boldsymbol{u}}$ with $\angle(\hat{\boldsymbol{u}}, \boldsymbol{v}_1) \le \theta \le \pi/2$,*

$$\hat{\boldsymbol{u}}^\top \mathbf{H}_T \hat{\boldsymbol{u}} \;\ge\; \lambda_1 \cos^2\theta, \qquad \lambda_1 - \hat{\boldsymbol{u}}^\top \mathbf{H}_T \hat{\boldsymbol{u}} \;\le\; \lambda_1 \sin^2\theta.$$



(a) LLaMA family (b) Qwen family

Figure 9: Average tokens per dialogue across six datasets. SOMA reduces token usage by avoiding repeated full-history serving after switching.

Proof of Lemma E.4. Let $\varphi = \angle(\hat{\boldsymbol{u}}, \boldsymbol{v}_1) \le \theta$ and decompose $\hat{\boldsymbol{u}} = \cos\varphi\, \boldsymbol{v}_1 + \sin\varphi\, \boldsymbol{w}$ with $\|\boldsymbol{w}\|_2 = 1$ and $\boldsymbol{w} \perp \boldsymbol{v}_1$. Because $\mathbf{H}_T \boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$, the cross term vanishes, and

$$\hat{\boldsymbol{u}}^\top \mathbf{H}_T \hat{\boldsymbol{u}} = \lambda_1 \cos^2\varphi + \sin^2\varphi \sum_{j \ge 2} \lambda_j \langle \boldsymbol{v}_j, \boldsymbol{w} \rangle^2 \;\ge\; \lambda_1 \cos^2\varphi \;\ge\; \lambda_1 \cos^2\theta,$$

since $\lambda_j \ge 0$ and $\cos^2$ is decreasing on $[0, \pi/2]$. Subtracting from λ_1 gives the residual bound. □

Corollary E.5 (Candidate Budget)**.** *Under the conditions of Proposition E.3, the coverage probability within angle θ is at least $1-\delta$ whenever $M \ge \log(1/\delta) \big/ \log\big(1/(1-p_{r_{\mathrm{act}}}(\theta))\big)$. Because $\log(1/(1-p)) \ge p$, the simpler condition $M \ge (r_{\mathrm{act}} - 1)\, \mathrm{B}\big(\frac{r_{\mathrm{act}}-1}{2}, \frac{1}{2}\big) \log(1/\delta)/(\sin\theta)^{r_{\mathrm{act}}-1}$ also suffices. Together with Lemma E.4, this trades candidate count against the quality of the best direction.*

Proof of Corollary E.5. Coverage fails with probability $(1-p)^M$, where $p = p_{r_{\mathrm{act}}}(\theta)$, and $(1-p)^M \le \delta$ holds if and only if $M \log(1/(1-p)) \ge \log(1/\delta)$. The second condition follows from $\log(1/(1-p)) \ge p$ and the lower bound on p in Proposition E.3. □

F Additional Experimental Results

This appendix provides supporting results that complement the main experiments: dataset-level token costs, the effect of warm-start length on switching, and the empirical variance concentration behind the switching bound in Section 4.1.

F.1 Additional Efficiency Results

The main text reports the amortized break-even analysis and throughput. Here we provide the dataset-level token results for both model families. Figure 9 shows that SOMA uses the fewest tokens per dialogue on 11 of the 12 dataset–family pairs. The exception is Craigslist with LLaMA, where Original uses 3% fewer. The reason is that after switching, SOMA serves later turns with the adapted surrogate and a compressed context rather than repeatedly sending the full growing history. This avoids the large token cost of Original and History-Prefix while retaining stronger response similarity than directly using the unadapted surrogate.

F.2 Warm-Start Length and Switching Behavior

The main text (Figure 6) reports how final similarity depends on the switch point. The sweep uses warm-start windows $W \in [1, 15]$ and measures the final response similarity after switching. Once the curve flattens, adding more warm-start turns brings little quality gain but delays the efficiency benefit, so short warm starts suffice for general or goal-anchored chats, while compositional reasoning and multi-party dialogues need more evidence before switching.

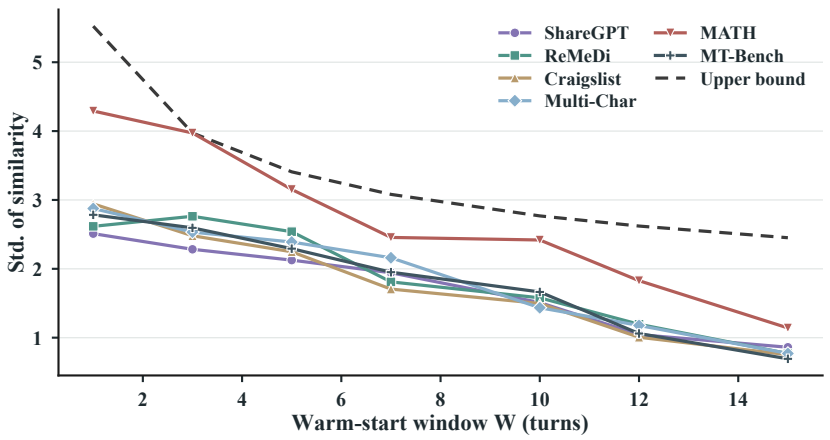


Figure 10: Variance of response similarity versus warm-start window W. Larger warm-start windows reduce variance and support more stable switching decisions.

F.3 Empirical Variance Concentration

Finally, we examine how the variance of surrogate–teacher similarity changes with the warm-start window W. For each dataset, we compute the empirical standard deviation across dialogues at each truncation point. Figure 10 shows that variance decreases as W grows: early turns are highly variable, while larger warm-start windows enter a more stable regime. This trend is consistent with the concentration behavior used in Lemma E.1 and the switching bound in Proposition 4.1. In other words, early turns provide the largest information gain, while later warm-start turns mainly reduce uncertainty until the switching decision becomes stable.

G Design Rationale and Serving Considerations

This section collects design questions raised in review that the main text only touches on.

Choice of soft prompts and of an adversarial objective. Soft prompts are a convenient probe for three practical reasons. They are differentiable, so a handful of gradient steps on the surrogate G can search a continuous neighborhood of the warm-start context without enumerating discrete paraphrases. They live in G's own embedding space, which lets the semantic neighborhood $\mathcal{N}_k(\cdot)$ and the expected-embedding weight share the same token embeddings and reuse the logits of a single forward pass. The precomputed neighborhoods also keep the extra per-step cost proportional to $k + m$, independent of $|\mathcal{V}|$. Other perturbations, such as input noise, paraphrasing, or adversarial token substitution, could also explore beyond the finite warm-start turns. They need an external paraphraser or a discrete search, however, and they do not directly show which *directions* in G's input space move G away from F. We push G away from F because the score is meant to measure how robust the existing agreement is. A constructive, divergence-minimizing prompt would raise the agreement on every turn, and its signal would be hard to separate from the capacity of the prompt itself. An adversarial prompt removes the agreement that is easy to remove, and r_t records what is left. The prompts are then discarded. Only the anchoring scores r_t enter the LoRA objective in Eq. 4, where they serve as per-turn training weights.

Relation to divergence-weighted History-FT and Random-FT. The LoRA objective in Eq. 4 weights every warm-start turn by its anchoring score $\omega_t \propto r_t$. SOMA can therefore be read as a weighted History-FT whose weights come from prompt probing, whereas a divergence-weighted History-FT would use the raw output divergence between G and F on each warm-start turn. The two weightings measure different things and push in opposite directions. Raw output divergence emphasizes turns that the *unperturbed* surrogate currently gets wrong. The anchoring score emphasizes turns whose teacher response the surrogate still predicts *after* an adversarial perturbation, that is, responses that seem to follow from the dialogue state and may therefore transfer to later turns. The ablation in Figure 5a isolates the two terms of the mining objective. It includes no baseline that

weights History-FT by raw output divergence, and none that samples training turns uniformly at random with the same LoRA budget. We regard both as key controls and leave them, together with a prompt-length sweep, to future work.

Robustness, divergence, and the perturbation space. We do not prove that robustness under prompt perturbations predicts adaptation utility better than direct output divergence. Nor do we prove that the continuous perturbation space of $\mathbf{P}$ covers the space of natural-language future turns. We treat both as modeling assumptions and test them indirectly. The topic-shift stress test (Figure 3(c)) probes future turns that leave the mined region, and the warm-start sweep (Appendix F.2) shows how much of the local behavior the window captures as it grows. A direct comparison of the two weightings would probably be the most informative next step for validating the mining objective.

Number of large-model calls and target consistency. The original model F is queried once per warm-start turn to serve the user. We store these W responses and reuse them as the mining targets a_t^F, as the LoRA targets, and as the references for the acceptance gate. Since F is not re-queried while $\mathbf{P}$ is optimized, the targets stay consistent with the dialogue state, and the soft prompt perturbs only the surrogate side of the objective. Because the acceptance gate reuses these turns, the acceptance bound of Section 4.1 does not certify the gate as implemented. Holding out part of the warm-start window would restore the bound, at the cost of fewer adaptation examples. Prompt mining and LoRA adaptation run entirely on the surrogate and take a small number of forward–backward passes over the W warm-start turns (M candidate prompts times the mining iterations, followed by the LoRA steps). Their wall-clock cost is included in the end-to-end latency of Figure 3(a).

Drift detection. The drift statistic is the cosine distance between recent queries and the warm-start centroid, monitored over a sliding window, and rollback triggers after m_{cons} consecutive violations. Since the statistic is continuous, gradual drift raises it step by step and can trigger rollback before any abrupt shift occurs. Corollary E.2 would bound the probability of accepting a surrogate whose true gap exceeds the margin if the acceptance batch were held out and the centroid fixed in advance. In our implementation both come from the warm-start turns, so the initial switch is a heuristic check (Section 4.1). Re-switching after a rollback is different. Its acceptance batch and refreshed centroid come from the turns F serves after the shift, which the adapter never saw, so the held-out condition of Proposition 4.1 holds for that decision. The gate also cannot guarantee that the centroid distance tracks every kind of quality loss, so Figure 3(c) also reports false-rollback and detection rates.

Serving many session-local adapters. A LoRA adapter of rank r on the attention and MLP projections of LLaMA-2-7B has about $2.5r$ million parameters, roughly 20M parameters or 40 MB in bfloat16 for $r = 8$, well under 1% of the surrogate. Multi-adapter serving systems batch requests from many adapters over a single copy of the base weights and page adapters between host and GPU memory (Sheng et al., 2024). GPU residency is then bounded by the number of concurrently *switched* sessions, which is usually far below the total number of sessions, and an adapter is discarded as soon as its session ends. The mining and adaptation steps of a session do not depend on other sessions and can run as short background jobs on the surrogate replica. We did not benchmark multi-node adapter scheduling, and the reported latencies are for the single-session path of Section 3.

H Broader Impact

SOMA aims to reduce the cost and latency of multi-turn LLM serving by replacing repeated large-model inference with session-local surrogate serving. This can make interactive LLM systems more efficient, accessible, and environmentally sustainable. At the same time, using a smaller surrogate may introduce quality degradation if the dialogue leaves the learned local state, especially in high-stakes settings such as healthcare (Li et al., 2026; Yu et al., 2026), legal, or financial assistance. SOMA mitigates this risk through semantic acceptance and drift-aware rollback, but deployment should still include monitoring, conservative thresholds, and human oversight where errors may cause harm. Since SOMA uses dialogue history for local adaptation, practical deployments should also apply data minimization, retention limits, and privacy-preserving processing of the dialogue history.

I LIMITATIONS AND FUTURE WORK

SOMA has several limitations. First, the surrogate's capacity limits how closely it can match the original model. When the small–large model gap is very large, some behaviors cannot be recovered by session-local adaptation alone. Second, soft-prompt mining requires access to the surrogate tokenizer and embedding space, which makes SOMA harder to deploy in strict black-box API settings. Third, SOMA is designed for locally coherent dialogues. Abrupt topic changes, noisy warm-start windows, or strong paraphrase shifts can weaken the learned local directions, although the drift gate is designed to reduce persistent degradation. Finally, SOMA introduces a one-time warm-start cost, so it is most suitable for medium-to-long sessions where post-switch savings can amortize the initial adaptation overhead. Two further limitations concern evaluation. The same continuous centroid statistic detects gradual drift and abrupt shifts, yet our stress test covers only abrupt shifts (Figure 3(c)), and quality may degrade before the rollback margin is exhausted. We also have not compared SOMA with a History-FT variant weighted by raw output divergence, or with uniformly sampled local turns under the same LoRA budget (Appendix G). The added value of prompt mining over such simpler weightings therefore rests on the component ablation alone.

Future work can improve SOMA along several directions. Better drift detectors could make rollback more reliable under topic shifts. Approximate mining methods could reduce the need for internal surrogate access. Multi-region adaptation could support conversations with multiple topics, and privacy-preserving mining could make the method safer for sensitive dialogue settings. Another promising direction is to amortize soft-prompt search across related sessions and extend SOMA to multimodal serving, where early turns often carry images or documents.