The Shuffle Gap: Diagnosing Personalization Collapse in LLMs

Barak Gahtan
Technion Israel Institute of Technology
barakgahtan@cs.technion.ac.il

Ido Galil
Technion Israel Institute of Technology
idogalil@campus.technion.ac.il

Alex M. Bronstein
Technion Israel Institute of Technology
ISTA Institute of Science and Technology Austria
bron@cs.technion.ac.il

Abstract

Personalized language models are often evaluated by held-out perplexity (PPL) after conditioning on a user's history, but aggregate PPL cannot tell whether the model used the *matched* user's support or merely learned a population-level adapter. We study this ambiguity through the *shuffle gap*: the PPL change when each user's support set is replaced by another user's support. In a hypernetwork case study, WARMSTART maps user histories to LoRA adapters and reduces Blog Authorship PPL from 34.0 to 21.14, yet its random-support shuffle gap at $k{=}5$ is only 0.008 PPL (95% CI $[-0.002, 0.018]$). Under a BM25 topic-controlled shuffle, the gap remains 0.010 PPL; topic-controlled in-context learning has a 3.62 PPL gap under the same audit. No-user global LoRA controls close about four-fifths of the Blog zero-shot$\rightarrow$WARMSTART gain and match WARMSTART on LaMP-7 ($8.02{\pm}0.07$ vs. 8.04 PPL), indicating that most standard-metric improvement is population adaptation rather than matched-support use. We calibrate the audit with noise-floor, discriminability, bottleneck, P2P, WildChat, PersonaFeedback, and synthetic support-signal controls. Together they support a conservative conclusion: in this hypernetwork case study, large PPL gains coexist with aggregate loss-level *personalization collapse*. Margin-based gap-forcing can open the gap, but incurs 8.8% PPL regression and does not improve PersonaFeedback preference sensitivity. The shuffle gap is therefore a necessary support-sensitivity audit for perplexity-based personalization, not a certificate of preference-aligned personalization.

1 Introduction

Personalized language models are usually judged by whether conditioning on a user's history improves held-out quality. For generative models, perplexity (PPL) is the natural loss-level metric: a lower PPL means the adapted model assigns higher probability to the user's held-out text. The difficulty is that a support-conditioned system can reduce aggregate PPL without using the *matched* user's support. It may instead learn a population-level adapter that fits the corpus well on average.

Our focus on PPL is deliberately narrow: it is the loss and headline metric for these systems, not a complete proxy for preference satisfaction, style imitation, or helpfulness. The prerequisite question is whether, under that same loss, the adapted model depends on the matched support; otherwise a useful PPL gain should be described as population adaptation rather than user personalization.

This ambiguity appears sharply in our case study. WARMSTART, a hypernetwork that maps user histories to LoRA adapters, reduces Blog Authorship PPL from 34.0 to 21.14. If each user's support

Preprint.

set is replaced by another user's support, however, PPL changes by only 0.008. The same random shuffle changes in-context learning by 12.33 PPL, or 3.62 PPL after BM25 topic matching. The hypernetwork therefore behaves less like a user-specific adapter and more like an efficient population-style prior.

We call this failure mode *personalization collapse*: under a specified loss and support channel, a model obtains large gains on the aggregate loss while showing near-zero sensitivity to the matched user's support. We diagnose it with the *shuffle gap*. Evaluate each user once with their own support and once with a deranged user-to-support assignment; the shuffle gap is the PPL difference. A large gap means the adapted model depends on which support it receives. A near-zero gap, when paired with a concentrated per-user distribution and evidence that supports are distinguishable, means the model is not materially using the matched support under the evaluated metric.

Our main contribution is an evaluation claim rather than a new personalization method:

- **We formalize matched-vs-mismatched support evaluation.** The shuffle gap tests whether a support-conditioned model's PPL gain depends on the matched user's history, rather than only on corpus-level adaptation.
- **We identify aggregate personalization collapse in competitive hypernetworks.** WARM-START obtains large Blog PPL gains but has a near-zero shuffle gap, and P2P [Tan et al., 2025] shows the same pattern across three seeds.
- **We map the source of collapse with controls and ablations.** Across topic-matched shuffles, population-only adapters, support-discriminability probes, encoder ablations, WildChat, PersonaFeedback, and a synthetic support-signal task, we test whether the null gap reflects real support-insensitivity or a trivial artifact of the data, encoder, or metric.
- **We show collapse can be broken, but not for free.** A dedicated margin term opens the Blog shuffle gap to 7.10 PPL, but matched-support PPL regresses by 8.8% and PersonaFeedback preference sensitivity does not improve, suggesting that current objectives may prefer population-efficient adaptation over using available user signal.

The accompanying theory is intentionally modest: an accounting identity decomposes expected personalization error into population excess risk plus an irreducible consensus gap, clarifying why pooled ERM can prefer population adaptation. Our evidence is most direct for hypernetwork-based perplexity personalization on English stylometric generation. We do not claim that retrieval-based, gradient-based, or preference-aligned personalization collapses in general.

2 Problem Setting and Case Study

Support-conditioned personalization. For each user u, let S_u^k be a support set of k user examples and Q_u be a disjoint query set. A support-conditioned personalization method is an adaptation operator

$$\mathcal{A}(\theta_{\text{base}}, S_u^k) = \theta_u,$$

which returns a user-conditioned model θ_u from a frozen or trainable base model θ_{base}. Standard evaluation reports the mean query loss or PPL of θ_u on Q_u.

The shuffle gap keeps the query users fixed and deranges only the supports. Let π be a derangement over the n_{test} meta-test users, so $\pi(u) \neq u$. For a fixed derangement, the per-user and aggregate gaps are

$$\begin{aligned} \Delta_{\text{shuf}}(u;\pi) &= \text{PPL}\Big(\mathcal{A}(\theta_{\text{base}}, S_{\pi(u)}^k), Q_u\Big) - \text{PPL}\big(\mathcal{A}(\theta_{\text{base}}, S_u^k), Q_u\big)\,, \\ \Delta_{\text{shuf}} &= \frac{1}{n_{\text{test}}} \sum_u \Delta_{\text{shuf}}(u;\pi). \end{aligned} \tag{1}$$

A large positive value means matched support matters for the evaluated loss. A near-zero value is an observational signal: it may reflect ignored support, functionally interchangeable support, or compression of distinguishable support into nearly user-invariant adapters. We report the signed gap plus real and shuffled PPLs; a negative gap simply means matched support has no demonstrated advantage. The experiments therefore pair Equation 1 with discriminability, topic-control, noise-floor, and synthetic-signal checks.

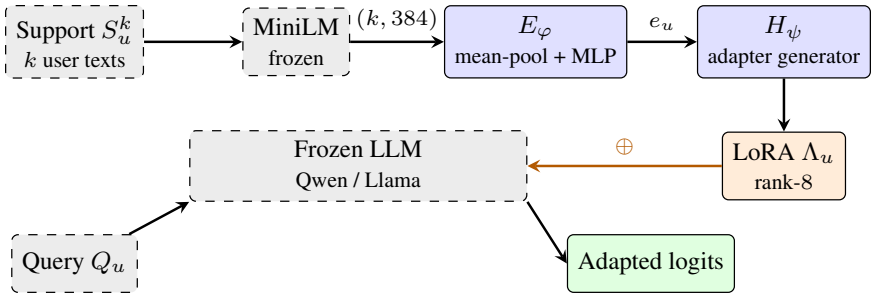


Figure 1: Case-study architecture: WARMSTART maps a support set to a user-conditioned LoRA adapter in one forward pass. The shuffle gap tests whether Λ_u changes in a loss-relevant way when S_u^k is replaced by another user's support.

WARMSTART hypernetwork. WARMSTART is the case-study vehicle, not the paper's primary contribution. It adapts a frozen LLM to a new user in one forward pass: each support item in S_u^k is embedded by frozen `sentence-transformers/all-MiniLM-L6-v2`; the k embeddings are mean-pooled and passed through a two-layer persona encoder E_φ; a hypernetwork H_ψ maps the resulting $e_u \in \mathbb{R}^{256}$ to rank-8 LoRA matrices injected into every $q_{\text{proj}}/v_{\text{proj}}$ layer of a frozen Qwen3-8B or Llama-3.1-8B backbone. Thus

$$\mathcal{A}_{\text{WS}}(\theta_{\text{base}}, S_u^k) = \theta_{\text{base}} \oplus H_\psi(E_\varphi(S_u^k)),$$

where $\oplus$ denotes LoRA composition. Appendix A.5 gives architecture, truncation, posterior, and bounded-loss details.

For training, each meta-batch samples users, disjoint supports S_u^k, and query sets Q_u. The backbone is frozen; gradients update only (φ, ψ). For a query set Q_u, let $\hat{\mathcal{L}}_u^{Q_u}(\theta)$ denote the token-normalized empirical cross-entropy of model θ on that user's query examples. The default objective is mean query cross-entropy plus a worst-tail penalty over users,

$$\mathcal{L}_{\text{meta}}(\varphi, \psi) = \mathbb{E}_u\left[\hat{\mathcal{L}}_u^{Q_u}\big(\mathcal{A}_{\text{WS}}(\theta_{\text{base}}, S_u^k)\big)\right] + \lambda\, \text{CVaR}_{0.5}\Big(\big\{\hat{\mathcal{L}}_u^{Q_u}\big\}_u\Big). \tag{2}$$

All main WARMSTART runs use $\lambda = 1$ unless explicitly marked as ERM ($\lambda = 0$). The $\text{CVaR}_{0.5}$ term is empirical, not a theorem-driven guarantee: it slightly improves tail PPL but does not prevent collapse. We use WARMSTART because it is a competitive hypernetwork personalization system whose standard metrics make the failure mode easy to miss.

3 Experiments

The primary evaluation is Blog Authorship [Schler et al., 2006]: users with at least 60 posts, split 70/10/20 into meta-train/validation/test, with 465 meta-test users. Each test user has a query set Q_u of 5 posts, a support pool of 20 posts, and a disjoint oracle partition used only by the restricted-oracle LoRA (rOR). We report mean per-user PPL and tail metrics. Baselines are zero-shot, ICL, BM25-RAG, support-LoRA trained on the same sampled S_u^k, rOR, P2P [Tan et al., 2025], and no-user global LoRA. Full protocols, seeds, CIs, and LaMP-7/WildChat/PersonaFeedback details are in Appendix A.6–A.12.

The standard PPL reading is simple: WARMSTART reduces mean PPL by nearly 13 points over zero-shot, beats ICL/BM25/support-LoRA by more than 3 PPL at every $k \geq 1$, and is competitive with P2P. Yet the same plots already contain the first warning sign. The support-size curve is saturated at k=1 (Appendix A.13), so adding more matched user history does not measurably improve the hypernetwork. A no-user global LoRA makes the warning sharper. On Blog, a pooled LoRA with no support input closes about four-fifths of the zero-shot→WARMSTART gain; on LaMP-7 it reaches 8.02 ± 0.07 PPL, matching WARMSTART's 8.04 at k=5 (Appendix A.17).

The PPL advantage is not a favorable-seed artifact. Restricting WARMSTART to the same three seeds used by ICL, BM25-RAG, and support-LoRA leaves a 21.00 mean PPL, with paired seed-wise gaps excluding zero against all three baselines (Appendix A.18). The comparison is also query-time compute matched with support-LoRA in the relevant sense: both score queries through the

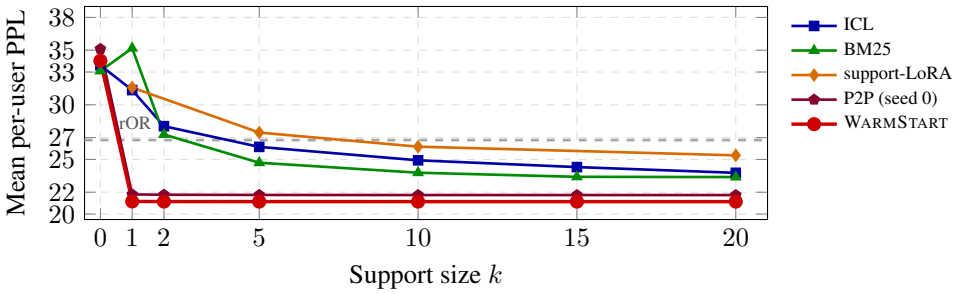


Figure 2: Mean per-user PPL versus support size on Blog. WARMSTART and P2P look excellent by PPL but saturate at k=1; methods that directly read or optimize on support continue improving with larger k. The P2P line is included only to show the shape of the single-seed full-k sweep; Table 1 is the fair k=5 ranking comparison and reports the three-seed P2P aggregate.

Table 1: Per-method perplexity profile at k=5 on Blog meta-test (n=465, Qwen3-8B frozen). Lower is better. WARMSTART/ERM-WARMSTART: n=5 seeds. rOR, ICL/BM25-RAG/support-LoRA, and P2P: n=3 seeds.

Method	Mean	$\text{CVaR}_{0.5}$	$\text{CVaR}_{0.1}$	p_{95}	p_{99}	Worst
Zero-shot (no adapter)	34.03	51.03	101.22	87.66	165.32	639.28
ICL	26.15	37.19	60.90	55.96	87.50	148.78
BM25-RAG	24.70	33.50	52.98	48.80	70.23	129.69
support-LoRA on S_u^k	27.46	40.67	81.03	69.11	158.32	317.78
P2P [Tan et al., 2025]	**20.96**	**29.16**	**46.49**	**41.76**	**67.03**	**116.57**
rOR on O_u (tuned)	26.77	38.91	70.19	62.12	101.96	305.17
ERM-WARMSTART	21.13	29.52	48.24	43.61	70.39	220.07
WARMSTART	21.14	29.50	48.12	43.42	69.91	223.21

same rank-8 LoRA insertion points, but WARMSTART amortizes meta-training while support-LoRA spends per-user gradient steps at deployment. We therefore treat the PPL gain as real amortized adaptation–and then ask whether that adaptation is user-specific.

The remaining Blog gap between global LoRA and WARMSTART is not zero: the mean residual is about 2.65 PPL, and the $\text{CVaR}_{0.1}$ residual is 12.13 PPL (Appendix A.17). We therefore do not claim that the hypernetwork is identical to a global adapter. The narrower, testable claim is that the residual improvement is nearly invariant to *which* user's support is provided. This is the point of the shuffle gap: it distinguishes a support-conditioned parameterization that improves corpus likelihood from one whose improvement is causally tied to the matched support examples.

With that baseline fixed, we next apply the shuffle-gap audit from Equation 1. We report the aggregate gap together with per-user distributions, topic-controlled shuffles, discriminability checks, no-user controls, and synthetic sanity checks, because a small gap alone is observational rather than mechanistic.

Table 2 and Figure 3 overturn the standard-metric reading. In the three-training-seed summary, WARMSTART's k=5 shuffle gap is 0.008 PPL; on the seed-0 checkpoint, four derangement seeds give 0.0099–0.0134 PPL, and deterministic/noise-floor controls are at most 0.0003 PPL (Appendix A.10). The per-user distribution is also concentrated: 91.8% of Blog users are below 0.01 log-CE gap under ERM (Appendix A.27). P2P shows the same pattern despite a different hypernetwork design. By contrast, a separate fixed-sample ICL diagnostic opens a 12.33 PPL gap under random shuffling; after BM25 topic matching, ICL still has a 3.62 PPL gap while WARMSTART remains at 0.010 PPL. The gap is therefore not merely a topic artifact, and Table 1 remains the fair seed-averaged PPL comparison.

The word "near-zero" is practical rather than absolute. The random-derangement gap is above the deterministic evaluation floor, but its scale is roughly 10^{-2} PPL: more than 300× smaller than topic-controlled ICL and about 700× smaller than the gap-forced model. Likewise, the 0.01 log-CE threshold in Figure 4 is a descriptive visual cutoff rather than a decision rule; the conclusion follows

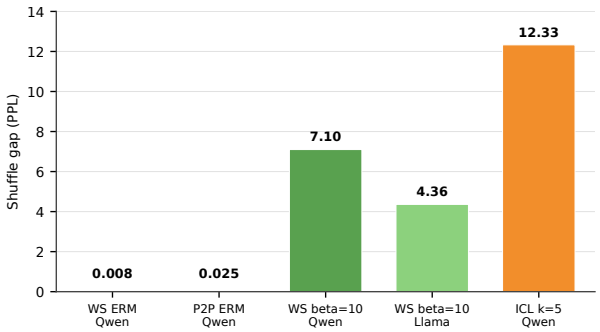


Figure 3: Shuffle gap at k=5. ERM hypernetworks (WARMSTART and P2P) have near-zero support sensitivity despite competitive PPL. ICL and gap-forced hypernetworks have large gaps; the latter pay a real-PPL regression cost.

Table 2: Collapse and gap-forcing summary at k=5. WARMSTART and P2P are competitive by PPL but nearly support-insensitive. Gap-forcing opens the gap but worsens matched-support PPL. Rows marked as diagnostic probes use fixed raw-shuffle support assignments and are not seed-wise constituents of Table 1's PPL averages.

Condition	Real PPL	Shuf. PPL	Gap	Notes
WARMSTART ERM	21.04	21.04	0.008	3-seed CI $[-0.002, 0.018]$
P2P	20.96	20.99	0.025	3-seed k=5 aggregate
support-LoRA	28.60	29.35	0.75	seed-0 diagnostic probe
ICL	30.07	42.41	12.33	seed-0 diagnostic probe
WARMSTART, β=10	23.00	30.10	7.10	8.8% vs. 5-seed ERM
Llama WARMSTART, β=10	19.81	24.18	4.36	6.5% regression

from the full distribution and the calibrated reference gaps. We also see heterogeneous users rather than universal zero: one discriminability quartile has a positive CI, but the per-user precision@5–gap correlation is essentially zero. We therefore use *aggregate collapse* to mean that matched support is not a material driver of the reported mean-PPL gain, not that every user is perfectly invariant.

The calibration fixes the scale of that statement. Exact repeat gives zero gap, permuting the order of a user's own support changes aggregate PPL by 0.0003, and resampling another support set from the same user changes it by 0.0001 (Appendix A.10). Randomly replacing the support with another user's support is larger than these floors, but stable across derangement seeds (0.0099–0.0134 PPL). The audit can also fire when support contains a causal signal: on a controlled synthetic task where support states the keyword needed by the query, random shuffling increases answer-token loss by 2.27 log CE, while label-matched shuffling gives zero gap (Appendix A.11). Thus the Blog result is best read as calibrated support-insensitivity, not as literal numerical invariance.

A small gap could still mean the supports are interchangeable, not that the model ignores useful support. We therefore measure support discriminability in the same MiniLM space WARMSTART reads. This is a mechanism-faithful check–it asks whether the adaptation pipeline could in principle distinguish the supports–so we pair it with partially independent BM25 topic controls and authorship probes. Blog support items are distinguishable at the item level (precision@5 $\approx$ 0.102, recall@5 $\approx$ 0.027, about 50$\times$ chance), and split-half support centroids identify the same user at 51.6% top-1 and 71.3% top-5 accuracy (chance: 0.2% and 1.1%). Yet per-user discriminability has no monotonic association with per-user shuffle gap (Pearson $r \approx -0.01$, $p = 0.88$). The result is heterogeneous–one quartile has a positive CI–so we describe the finding as aggregate collapse rather than uniform zero gap. Unfreezing the MiniLM encoder opens only a 0.022 PPL gap, and authorship attribution remains near chance. A WildChat-1M stress test, where supports are much more discriminable, still shows no matched-support advantage across four derangement seeds. Finally, a controlled synthetic task confirms the audit can fire: when support states the target keyword, random shuffling raises target-token loss by 2.27 log CE, while label-matched shuffling has zero gap. These checks make support interchangeability an unlikely primary explanation in the evaluated settings.

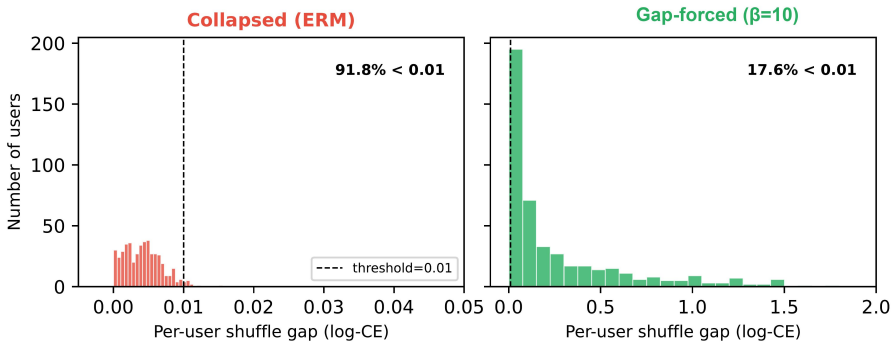


Figure 4: Per-user shuffle-gap distribution (log-CE scale, k=5, n=465). Left: collapsed ERM model – 91.8% of users below 0.01. Right: gap-forced β=10 model – only 17.6% below 0.01, with a heavy right tail.

The same conclusion appears in direct probes of recoverable user identity. In authorship attribution, the adapted model scores one held-out oracle post from the true user against distractor posts. With 20 candidates, rank-1 accuracy stays at chance across all support sizes (3.9–5.6%, chance 5%); with 100 candidates at k=5, rank-1 accuracy is 0.86% with CI $[0.22, 1.72]$%, and MRR equals the random-rank value (0.0519; Appendix A.23). Persona-null counterfactuals give a complementary loss-level probe. For collapsed WARMSTART, replacing every user's support with one fixed support set changes PPL by -0.02, and replacing the persona vector with the population mean changes PPL by 0.00. For the gap-forced model, the same wrong fixed support is catastrophic ($+18.22$ PPL), showing that the intervention really couples the adapter to support identity (Appendix A.24). These probes support the same interpretation as the shuffle gap: the collapsed model's adapter is largely on-manifold and population-like, rather than recoverably user-specific.

The secondary evaluations then keep the scope honest. LaMP-7 is the cleanest population-adaptation control: WARMSTART reaches 8.04 PPL at k=5, but a single no-user LoRA reaches 8.02 ± 0.07 PPL and the LaMP-7 shuffle gap is only 0.009 PPL (Appendix A.15, A.17). WildChat-1M is a heterogeneity stress test rather than a matched-domain replication. Its supports are more discriminable than Blog (precision@5 $\approx$ 0.522 vs. 0.102), yet real-support PPL is 15.49 and four shuffled-support seeds give 15.45–15.46 PPL, with paired log-CE gaps from -0.0026 to -0.0018 (Appendix A.20). We therefore interpret WildChat conservatively: the matched support does not help, but the negative sign is not a mechanism. PersonaFeedback is a preference-format boundary case. Collapsed WARMSTART scores 57.9% accuracy, gap-forced WARMSTART scores 55.7%, and repeated support shuffles change accuracy by less than one point (Appendix A.28). Together, these checks support the paper's scope: the shuffle gap diagnoses loss-level support sensitivity, not preference alignment in general.

Finally, we ask whether collapse is an architectural inevitability or an optimization outcome. We add a shuffled-support hinge term with margin $\delta = 0.5$ and weight β. At $\beta = 10$, the gap opens to 7.10 PPL on Qwen3-8B and 4.36 on Llama-3.1-8B; matched-support PPL regresses by 8.8% relative to the five-seed Qwen ERM mean and by 6.5% on Llama. Figure 5 shows that this is not a one-off point: no tested intervention opens a large gap without a nontrivial real-PPL cost. A preference-based boundary case is also negative: on PersonaFeedback hard tier, collapsed and gap-forced models have overlapping preference accuracy (57.9% vs. 55.7%), and repeated support shuffles leave accuracy nearly unchanged. Opening a PPL shuffle gap is therefore necessary for loss-level support sensitivity, but it is not sufficient for preference-aligned personalization.

Representation diagnostics show what the intervention changes. In collapsed ERM at k=5, the same-user minus different-user cosine gap is only 0.004 in persona space and 1.4×10^{-5} in effective LoRA-update space. Under gap-forcing, the persona-space gap rises to 0.408 and the adapter-space gap rises to 1.9×10^{-3} (Appendix A.25). Thus gap-forcing creates a user-sensitive channel, but most of the adapter direction remains shared. This matches the PPL tradeoff: the intervention proves the architecture can carry user signal, yet it does not by itself find a loss-improving personalized solution.

Two auxiliary probes reinforce this conservative reading. Layer-wise ablations of the gap-forced model localize most of the shuffle-gap contribution to late decoder blocks, especially layers 27

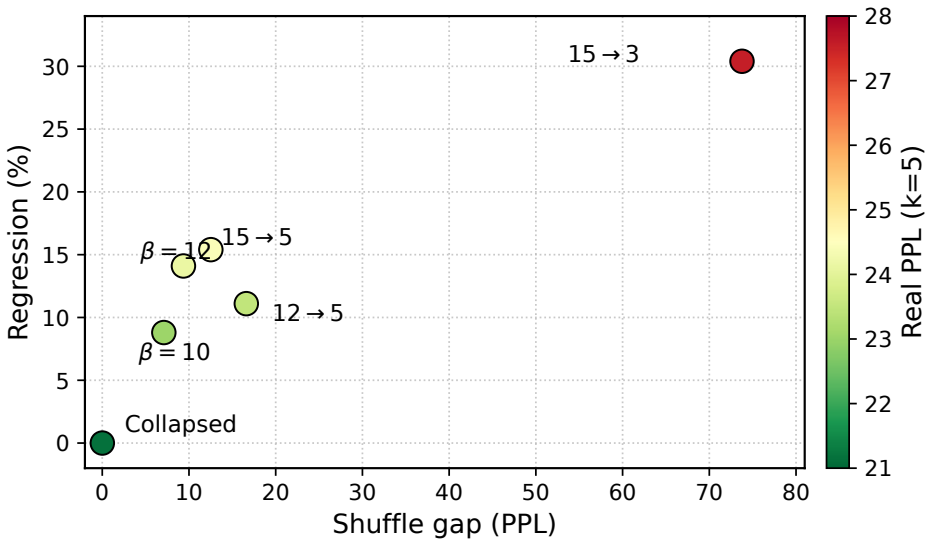


Figure 5: Gap–regression Pareto frontier on Blog (k=5). Regression is computed against the same five-seed collapsed WARMSTART mean PPL from Table 1; color encodes real PPL.

and 32–35; early and middle layers contribute little (Appendix A.26). In short free generations, the collapsed model has almost no matched-vs-shuffled style-proxy difference (MiniLM cosine gap $+0.006$, type-token ratio unchanged), while the gap-forced model shows a larger embedding-space gap ($+0.032$) but mixed surface statistics (lower matched-support type-token ratio and similar Self-BLEU; Appendix A.29). These diagnostics are useful sanity checks, not a claim of solved generation-quality personalization.

Table 3 summarizes the main controls behind the interpretation. The point is not that any single number proves collapse causally. The point is that standard PPL, no-user controls, support shuffles, discriminability checks, synthetic calibration, and preference evaluation all point to the same scoped conclusion: WARMSTART's Blog gain is mostly population adaptation, and a larger PPL shuffle gap is necessary but not sufficient for user-facing personalization.

4 Why Collapse Can Be Loss-Efficient

The empirical story is enough to diagnose collapse; the theory is only accounting intuition for why pooled ERM can prefer it. Let $\mathcal{L}_u(\theta)$ be user u's population loss, $\mathcal{L}(\theta) = \mathbb{E}_u[\mathcal{L}_u(\theta)]$, $\theta^\star = \arg\min_\theta \mathcal{L}(\theta)$, and $\mathcal{L}_u^\star = \min_\theta \mathcal{L}_u(\theta)$. Define personalization error $\Delta_{\mathrm{Pers}}(\theta, u) = \mathcal{L}_u(\theta) - \mathcal{L}_u^\star$, the population excess risk $\Delta_{\mathrm{Gen}}(\theta) = \mathcal{L}(\theta) - \mathcal{L}(\theta^\star)$, and the consensus gap $\Delta_{\mathrm{Cons}} = \mathcal{L}(\theta^\star) - \mathbb{E}_u[\mathcal{L}_u^\star]$.

Proposition 1 (Decomposition identity)**.** *For any $\theta \in \Theta$,*

$$\mathbb{E}_{u\sim\mathcal{P}_U}[\Delta_{\mathrm{Pers}}(\theta, u)] = \Delta_{\mathrm{Gen}}(\theta) + \Delta_{\mathrm{Cons}}.$$

Proof sketch. By linearity of expectation, $\mathbb{E}_u[\mathcal{L}_u(\theta) - \mathcal{L}_u^\star] = \mathcal{L}(\theta) - \mathbb{E}_u[\mathcal{L}_u^\star] = (\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)) + (\mathcal{L}(\theta^\star) - \mathbb{E}_u[\mathcal{L}_u^\star])$. □

The identity is exact for any fixed estimator. For a user-agnostic model trained only on the population mixture, ERM can reduce Δ_{Gen} while leaving Δ_{Cons} untouched. A support-conditioned operator can in principle escape this limitation, but nothing in aggregate PPL forces it to do so: if a population adapter already captures most likelihood gain, a collapsed support channel can remain loss-efficient. Under realizable conditional log loss, Appendix A.3 identifies Δ_{Cons} with the expected conditional divergence $\mathbb{E}_{u,x}[\mathrm{KL}(\mathcal{D}_u(\cdot \mid x) \| \bar{\mathcal{D}}(\cdot \mid x))]$, connecting the term to user heterogeneity in the next-token distribution. A supplementary PAC-Bayes bound for clipped adapted query loss is also in Appendix A.4; it is a meta-generalization statement, not a theory of the CVaR penalty or a sample-complexity-in-k guarantee.

Table 3: Evidence map for the aggregate-collapse interpretation. Full protocols and additional CIs are in the referenced appendix sections.

Main result	Interpretation
PPL improves. Blog PPL drops 34.03→21.14; WARMSTART beats ICL/BM25/support-LoRA at k=5 (Table 1).	The adaptation procedure is loss-effective by standard evaluation.
Matched support barely matters. Random shuffle gap is 0.008 PPL; topic-controlled gap is 0.010 PPL; noise-floor controls are $\leq$ 0.0003 PPL (Appendix A.10).	Matched support has almost no loss-level effect.
A no-user model explains most gain. Global LoRA closes about four-fifths of Blog gain and matches LaMP-7 WARMSTART: 8.02±0.07 vs. 8.04 PPL (Appendix A.17).	Most improvement is population adaptation.
Supports are distinguishable. Blog precision@5 0.102, recall@5 0.027; split-half centroids recover users at 51.6% top-1 and 71.3% top-5 (Appendix A.19).	The null gap is not explained by trivially identical supports.
Identity is not recoverable. 20-way authorship stays at chance (3.9–5.6%); mean-persona and fixed-support nulls change collapsed PPL by 0.00 and −0.02 (Appendix A.23, A.24).	The collapsed adapter behaves like a population adapter rather than a recoverably user-specific one.
The audit can fire. Synthetic random shuffling raises target-token loss by 2.27 log CE; label-matched shuffling gives zero gap (Appendix A.11).	The audit detects causal support signals and stays null for equivalent supports.
Gap does not imply quality. Gap-forcing opens a 7.10 PPL gap but increases matched PPL by 8.8%; PersonaFeedback accuracy remains similar (57.9% vs. 55.7%).	The shuffle gap is necessary for loss-level support sensitivity, not a quality certificate.

5 Related Work

Personalization evaluation. Standard text-generation metrics often miss personalization quality: ROUGE and BERTScore can understate personalized summarization [Vansh et al., 2023], PersonaLens measures variability in assistant personalization [Zhao et al., 2025], and PERG studies robustness in personalized generation [Okite et al., 2025]. Most directly, LaMP-QA validates a benchmark by comparing correct and mismatched user profiles [Salemi and Zamani, 2025]. We turn that sanity check into a model audit: the shuffle gap is paired with per-user distributions, topic controls, no-user LoRAs, support-discriminability checks, and gap-forcing probes.

LLM personalization methods. Existing systems condition on user history through retrieval or ICL [Salemi et al., 2024], train per-user PEFT adapters such as LoRA [Hu et al., 2021, Tan et al., 2024], or amortize adapter generation with hypernetworks. P2P [Tan et al., 2025] is the closest hypernetwork baseline to WARMSTART; our three-seed P2P check shows that the collapse finding is not specific to the WARMSTART persona encoder. Federated and meta-learning methods [Dinh et al., 2020, Deng et al., 2020, Finn et al., 2017, Amit and Meir, 2018] study related adaptation problems, but usually in training-time known-client or abstract task distributions rather than inference-time unseen-user personalization.

Aggregate-learning failure modes. The consensus-gap identity is closest to MiCRo's irreducible-error result for reward-model personalization [Shen et al., 2025]. The empirical failure resembles posterior collapse in variational text models, where a high-capacity decoder ignores a latent channel [Bowman et al., 2016, Lucas et al., 2019]; here the support set is the conditioning channel and the generated adapter is the latent code. Recommender-systems popularity baselines make a parallel point: high average accuracy can come from population-level item popularity rather than individualized preference modeling [Steck, 2011, Cremonesi et al., 2010]. Our no-user global LoRA is the analogous population-only baseline for LLM personalization. Extended comparisons are in Appendix A.30.

6 Discussion and Limitations

What the audit says. The Blog PPL gain is real, but most of it is explainable as population adaptation: no-user LoRA reproduces most of the gain, the hypernetwork saturates at $k = 1$, and matched-vs-

shuffled support barely changes loss. The shuffle gap therefore tests a different claim than standard PPL: not "does the model fit this corpus better?" but "does the adapted model depend on this user's support?" This distinction matters because the two claims lead to different scientific conclusions. A method can be a strong corpus adapter while being a weak user adapter.

What it does not say. A near-zero shuffle gap is observational. It can arise because the model ignores support, because supports are functionally equivalent for the metric, or because the support channel compresses distinguishable examples into nearly user-invariant adapters. Our Blog, BM25, authorship, trainable-encoder, WildChat, and synthetic controls argue against interchangeability as the primary explanation here, but do not eliminate every possible mechanism. Conversely, a large gap is not sufficient evidence of useful personalization: the gap-forced model is support-sensitive yet worse on matched-support PPL and not better on PersonaFeedback. The metric is therefore a diagnostic for loss-relevant support use, not a certificate of human-perceived preference alignment.

What gap-forcing teaches. The gap-forced model is best read as a probe, not as a proposed remedy. It shows that the same architecture can encode user-specific support when the objective explicitly rewards mismatch sensitivity, so collapse is not simply a capacity failure. The cost is that the user-sensitive solution gives up likelihood and does not transfer to preference accuracy. This suggests an optimization tension: under next-token ERM, ignoring available user-specific signal can be loss-efficient. A useful fix must therefore open a gap only in the direction that improves matched-support behavior, not merely punish mismatched support.

How to use the audit. For a perplexity-based personalization benchmark, report matched-support PPL, shuffled-support PPL under a fixed derangement, a measurement floor from repeat, order, and within-user controls, and a support-discriminability check. Each component blocks a different mistaken conclusion: the no-user LoRA separates population adaptation from personalization, discriminability checks rule out trivially identical supports, and downstream metrics prevent treating mismatch sensitivity as a quality certificate. When these reports disagree, the disagreement is informative: it reveals whether the bottleneck is the dataset, the support encoder, the training objective, or the evaluation metric.

Scope. The paper studies hypernetwork-based perplexity personalization on English stylometric generation, with LaMP-7 and WildChat as secondary checks and PersonaFeedback as a boundary case. We do not claim that retrieval-based, gradient-based, or preference-aligned personalization collapses in general, and we do not claim that PPL is the right final objective for every personalization application. The claim is narrower and easier to test: if a paper reports PPL gains from user support, it should also show that the evaluated loss changes when the support belongs to the right user. The open algorithmic problem is not merely to increase the shuffle gap, but to increase it without proportional PPL regression and with corresponding downstream gains.

7 Conclusion

The shuffle gap exposes a failure mode hidden by standard personalization benchmarks: a support-conditioned model can achieve large PPL gains while remaining nearly indifferent to whose support it receives. Our hypernetwork case study exhibits this pattern. WARMSTART looks strong under ordinary PPL, but the audit shows that most of the gain comes from population adaptation rather than loss-relevant user evidence. Reporting PPL without a matched-vs-mismatched support audit can therefore confuse adaptation to a corpus with personalization to a person.

We recommend pairing PPL with the shuffle gap, a measurement floor, a support-discriminability check, and downstream user-facing metrics when claiming preference or style alignment. The next target is not simply to maximize mismatch sensitivity. It is to design objectives and architectures where the right support improves both the evaluated loss and the downstream behavior that users actually experience.

References

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, 2018.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

Hongru Cai, Yongqi Li, Tiezheng Yu, Shengjie Luo, and Yuxuan Li. One adapts to any: Meta reward modeling for personalized LLM alignment. *arXiv preprint arXiv:2601.18731*, 2026.

Daiwei Chen, Yi Chen, Pradeep Ravikumar, and Qi Zhu. PAL: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, 2010.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012.

Guhao Feng, Shengjie Luo, Kai Hua, Ge Zhang, Di He, Wenhao Huang, and Tianle Cai. In-place test-time training. *arXiv preprint arXiv:2604.06169*, 2026.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The PRISM alignment dataset. *arXiv preprint arXiv:2404.16019*, 2024.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models. 2019.

David A. McAllester. PAC-Bayesian model averaging. In *Conference on Learning Theory*, 1999.

Chimaobi Okite, Naihao Deng, Kiran Bodipati, Huaidian Hou, Joyce Y. Chai, and Rada Mihalcea. Benchmarking and improving LLM robustness for personalized generation. In *Findings of EMNLP*, 2025.

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free RLHF. *arXiv preprint arXiv:2405.20304*, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts. In *International Conference on Learning Representations*, 2020.

Alireza Salemi and Hamed Zamani. LaMP-QA: A benchmark for personalized long-form question answering. In *Proceedings of EMNLP*, 2025.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the Association for Computational Linguistics*, 2024.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006.

Jingyan Shen, Bairu Jia, Cong Chen, Tianlu Wang, Huaxiu Yao, and Yixin Chen. MiCRo: Mixture modeling and context-aware routing for personalized preference learning. *arXiv preprint arXiv:2505.24846*, 2025.

Harald Steck. Item popularity and recommendation accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, 2011.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of EMNLP*, 2024.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Instant personalized large language model adaptation via hypernetwork. *arXiv preprint arXiv:2510.16282*, 2025.

Meiling Tao, Chenghao Zhu, Dongyi Ding, Tiannan Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. PersonaFeedback: A large-scale human-annotated benchmark for personalization. *arXiv preprint arXiv:2506.12915*, 2025.

Rahul Vansh, Darsh Rank, Sourish Dasgupta, and Tanmoy Chakraborty. Accuracy is not enough: Evaluating personalization in summarizers. In *Findings of EMNLP*, 2023.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of ACL*, 2025.

Wenting Zhong, Ruixiang Guo, Yifei Kamber, Benjamin Shao, Gyuhak Kim, Linh Phan, Tianhua Xing, and Srijan Bansal. WildChat: 1M ChatGPT interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

A Appendix

Roadmap. This appendix is organized in four blocks. We begin with notation, formal definitions, and full proofs of the main theoretical claims. We then collect method, data, metric, and protocol details needed to reproduce the experiments. Next we group the supplementary empirical results and diagnostics that support the shuffle-gap story. We close with contextual material: extended related work and broader-impact considerations.

Control summary. The collapse claim is supported by defense-in-depth controls, each ruling out a distinct alternative explanation: no-user global LoRAs (Appendix A.17) attributing about four-fifths of Blog WARMSTART's mean-PPL gain to population adaptation and matching WARMSTART on LaMP-7; persona-null counterfactuals (Appendix A.24) showing the collapsed model is invariant to mean/zero/random persona substitutions and to wrong fixed support, while gap-forcing makes wrong fixed support catastrophic ($+18.22$ PPL); authorship attribution at 20 and 100 candidates (Appendix A.23) returning near-chance identity recovery for the collapsed model; a trainable-support-encoder ablation (Appendix A.14) showing that unfreezing MiniLM opens only a tiny gap; adapter and persona geometry (Appendix A.25) showing collapsed and gap-forced representations differ qualitatively in persona space and quantitatively in adapter space; and PersonaFeedback (Appendix A.28) showing that opening a PPL shuffle gap does not transfer to preference accuracy. The shuffle gap is the loss-level synthesis of these controls.

A.1 Notation

Table 4: Summary of notation used throughout the paper.

Symbol	Description
Data & Users	
m	Number of meta-training users
n	Total number of users when a split is not specified
n_{test}	Number of meta-test users (465 on Blog, 1253 on LaMP-7)
k	Support set size (number of adaptation examples)
S_u^k	Support set of k examples for user u
Q_u	Query set for user u
$\mathcal{P}_U$	Distribution over users
O_u	Oracle partition for user u (rOR training data)
Model & Architecture	
$\mathcal{A}$	Adaptation operator $\mathcal{A} : \Theta \times (\mathcal{X} \times \mathcal{Y})^k \to \Theta$
θ_0	Base (frozen) backbone parameters
$\Delta\theta_u$	User-specific LoRA parameters
d_{emb}	Frozen sentence-encoder embedding dimension
d	Persona embedding dimension
r	LoRA rank
E_φ	Persona encoder (parameters φ)
H_ψ	Hypernetwork (parameters ψ)
Training & Loss	
α	CVaR confidence level (fraction of worst users)
λ	CVaR penalty weight
$\ell(u)$	Per-user loss
$\mathcal{L}_u(\theta)$	Per-user population loss
$\mathcal{L}(\theta)$	Population loss $\mathbb{E}_u[\mathcal{L}_u(\theta)]$
β	Margin loss weight
δ	Margin loss threshold
δ_{PB}	PAC-Bayes confidence level (Theorem 2)
Theoretical Quantities	
$\theta^\star$	Population risk minimizer $\arg\min_\theta \mathcal{L}(\theta)$
$\theta_u^\star$	Per-user risk minimizer $\arg\min_\theta \mathcal{L}_u(\theta)$
$\mathcal{L}^\star$	Population-averaged Bayes risk $\mathbb{E}_u[\mathcal{L}_u^\star]$
$\Delta_{\text{Pers}}(\theta, u)$	Personalization error $\mathcal{L}_u(\theta) - \mathcal{L}_u^\star$
$\Delta_{\text{Gen}}(\theta)$	Population excess risk $\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)$
Δ_{Cons}	Consensus gap $\mathcal{L}(\theta^\star) - \mathbb{E}_u[\mathcal{L}_u^\star]$
$\widetilde{\mathcal{L}}_u^{Q_u}(\theta)$	Clipped empirical query loss $\text{clip}(\hat{\mathcal{L}}_u^{Q_u}(\theta), 0, B)$
$\hat{\widetilde{\mathcal{L}}}_{\text{meta}}(Q)$	Empirical clipped meta-loss under meta-posterior Q
C_1	Universal constant in PAC-Bayes bound (Theorem 1)
$C_{\text{PB}}(Q)$	PAC-Bayes complexity term $C_1\sigma\sqrt{(\text{KL}(Q\|P) + \log(2\sqrt{m}/\delta_{\text{PB}}))/m}$
σ	Sub-Gaussian parameter of the clipped task loss (A1)
B	Per-user query-loss clipping bound; gives $\sigma \le B/2$ via Hoeffding
Evaluation	
PPL	Perplexity
rOR	Restricted-oracle per-user LoRA
Δ_{shuf}	Shuffle gap $\overline{\text{PPL}}_{\text{shuf}} - \overline{\text{PPL}}_{\text{real}}$

A.2 Extended preliminaries

During training we access data from a fixed set of meta-training users; at deployment we encounter meta-test users sampled from $\mathcal{P}_U$ for whom we have only a small amount of adaptation data. This distinction is crucial: a deployed LLM must handle unseen users whose preferences and linguistic patterns differ from the training population. Empirical risk minimization targets $\mathcal{L}(\theta)$ because it is the only quantity the learner can estimate from training data, not because it is what users actually care about; this is the gap the decomposition of Section 4 formalizes. Let $\mathcal{L}^\star = \mathbb{E}_u[\mathcal{L}_u^\star]$ denote the population-averaged Bayes risk. None of the oracle quantities $\theta^\star$, $\theta_u^\star$, $\mathcal{L}_u^\star$, $\mathcal{L}^\star$ is computable in

practice–they presuppose access to each user's true distribution–but all are well-defined theoretical targets against which any trained model can be measured.

Definition 1 (Personalization error)**.** $\Delta_{\mathrm{Pers}}(\theta, u) = \mathcal{L}_u(\theta) - \mathcal{L}_u^\star$ is the excess risk experienced by user u under θ relative to that user's oracle-optimal model.

Definition 2 (Consensus gap)**.** $\Delta_{\mathrm{Cons}} = \mathcal{L}(\theta^\star) - \mathbb{E}_u[\mathcal{L}_u^\star]$ is the irreducible price of fitting a single parameter vector to a heterogeneous user population.

Let $S_u^k = \{(x_i^u, y_i^u)\}_{i=1}^k$ be a k-shot support set drawn i.i.d. from $\mathcal{D}_u$ and let Q_u be a disjoint query set. An *adaptation operator* is a map $\mathcal{A} : \Theta \times (\mathcal{X} \times \mathcal{Y})^k \to \Theta$, and the adapted personalization error is $\Delta_{\mathrm{Pers}}(\mathcal{A}, u, S_u^k) = \mathcal{L}_u(\mathcal{A}(\theta_0, S_u^k)) - \mathcal{L}_u^\star$. When k is small, this adapted error governs the sample efficiency of personalization: how many user examples are required before the adapted model meaningfully approaches the user's oracle.

A.3 Full proof of Proposition 1 and consequences

Full proof of Proposition 1. By linearity of expectation,

$$\begin{aligned}\mathbb{E}_u[\Delta_{\mathrm{Pers}}(\theta, u)] &= \mathbb{E}_u[\mathcal{L}_u(\theta) - \mathcal{L}_u^\star] = \mathbb{E}_u[\mathcal{L}_u(\theta)] - \mathbb{E}_u[\mathcal{L}_u^\star] \\ &= \mathcal{L}(\theta) - \mathcal{L}^\star = \big(\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)\big) + \big(\mathcal{L}(\theta^\star) - \mathcal{L}^\star\big) \\ &= \Delta_{\mathrm{Gen}}(\theta) + \Delta_{\mathrm{Cons}}.\end{aligned}$$

□

Proposition 2 (Consensus gap vanishes iff users agree)**.** $\Delta_{\mathrm{Cons}} = 0$ *if and only if there exists a single* $\theta \in \Theta$ *such that* $\mathcal{L}_u(\theta) = \mathcal{L}_u^\star$ *for* $\mathcal{P}_U$*-almost every* u*.*

Proof. ($\Leftarrow$) If such a θ exists, $\mathcal{L}(\theta) = \mathbb{E}_u[\mathcal{L}_u(\theta)] = \mathbb{E}_u[\mathcal{L}_u^\star] = \mathcal{L}^\star$. Since $\mathcal{L}_u(\theta^\star) \geq \mathcal{L}_u^\star$ pointwise (by optimality of $\theta_u^\star$), integrating gives $\mathcal{L}(\theta^\star) \geq \mathcal{L}^\star$. Combined with $\mathcal{L}(\theta^\star) \leq \mathcal{L}(\theta) = \mathcal{L}^\star$, we obtain $\mathcal{L}(\theta^\star) = \mathcal{L}^\star$, so $\Delta_{\mathrm{Cons}} = 0$. ($\Rightarrow$) If $\Delta_{\mathrm{Cons}} = 0$, then $\mathcal{L}(\theta^\star) = \mathcal{L}^\star$ and $\mathcal{L}_u(\theta^\star) \geq \mathcal{L}_u^\star$ pointwise, with the integrated inequality tight. A pointwise non-negative integrand that integrates to zero vanishes almost everywhere, so $\theta^\star$ is the claimed common optimum. □

Corollary 1 (ERM cannot reduce the consensus gap)**.** *Let* $\hat{\theta}$ *be any estimator whose distribution depends only on samples drawn from* $\mathcal{D} = \mathbb{E}_u[\mathcal{D}_u]$*, independently of the downstream user. Then* $\mathbb{E}_u[\Delta_{\mathrm{Pers}}(\hat{\theta}, u)] \geq \Delta_{\mathrm{Cons}}$.

Proof. $\Delta_{\mathrm{Gen}}(\hat{\theta}) \geq 0$ since $\theta^\star$ minimizes $\mathcal{L}$. Applying Proposition 1, $\mathbb{E}_u[\Delta_{\mathrm{Pers}}(\hat{\theta}, u)] = \Delta_{\mathrm{Gen}}(\hat{\theta}) + \Delta_{\mathrm{Cons}} \geq \Delta_{\mathrm{Cons}}$. □

Proposition 2 formalizes the intuition that a zero consensus gap is the hallmark of a homogeneous user population. In any realistic deployment–users with different preferences, communication registers, or domain vocabularies–$\Delta_{\mathrm{Cons}} > 0$, and no amount of additional population data or algorithmic refinement of $\theta^\star$ can reduce it. Corollary 1 applies to estimators whose deployed parameters do not depend on the downstream user's support. It therefore rules out closing the consensus gap with pooled, user-agnostic training alone. User-conditioned procedures such as support-dependent retrieval, in-context conditioning, or support-generated adapters fall outside the scope of the corollary and must be analyzed separately–which is precisely the role of Theorem 1. The consensus gap connects to a growing literature on irreducible error in multi-party learning: Shen et al. [2025] prove an analogous result for reward-model personalization under mixture preferences, and Proposition 1 is the generative-loss analogue lifting the phenomenon from bounded preference pairs to unbounded next-token prediction.

Proposition 3 (Conditional log-loss identity under realizability)**.** *Under conditional log-loss* $\ell(\theta, x, y) = -\log p_\theta(y \mid x)$*, suppose the evaluated query-input distribution is fixed as* $\mathcal{P}_X$ *(or that evaluation conditions on the observed prompts), and write* $\mathcal{D}_u(\cdot \mid x)$ *for user* u*'s conditional next-token distribution. Let* $\bar{\mathcal{D}}(\cdot \mid x) = \mathbb{E}_u[\mathcal{D}_u(\cdot \mid x)]$*. If* Θ *is expressive enough to realize both* $\bar{\mathcal{D}}(\cdot \mid x)$ *and* $\mathcal{D}_u(\cdot \mid x)$ *for* $\mathcal{P}_U$*-almost every* u*, then*

$$\Delta_{\mathrm{Cons}} = \mathbb{E}_u \mathbb{E}_{x \sim \mathcal{P}_X}\left[\mathrm{KL}\big(\mathcal{D}_u(\cdot \mid x) \big\| \bar{\mathcal{D}}(\cdot \mid x)\big)\right].$$

Proof. Under the realizability assumptions, $\mathcal{L}(\theta^\star) = \mathbb{E}_{x\sim\mathcal{P}_X}[H(\bar{\mathcal{D}}(\cdot \mid x))]$ and $\mathcal{L}_u^\star = \mathbb{E}_{x\sim\mathcal{P}_X}[H(\mathcal{D}_u(\cdot \mid x))]$ for $\mathcal{P}_U$-almost every u. Therefore, applying the mixture-entropy identity pointwise in x,

$$\Delta_{\mathrm{Cons}} = \mathbb{E}_x[H(\bar{\mathcal{D}}(\cdot \mid x))] - \mathbb{E}_{u,x}[H(\mathcal{D}_u(\cdot \mid x))] = \mathbb{E}_{u,x}\left[\mathrm{KL}\big(\mathcal{D}_u(\cdot \mid x)\big\|\bar{\mathcal{D}}(\cdot \mid x)\big)\right].$$

□

If user prompt marginals also differ, this proposition should be read as the conditional component of the consensus gap under the evaluated query prompts. A joint KL over (x, y) would add input-marginal divergence, which ordinary conditional LM scoring does not directly measure.

A.4 Full proof of Theorem 1 and extended assumption commentary

Assumptions used by Theorem 1. *(A1, sub-Gaussian clipped task loss.)* For every hypothesis $h = (\theta_0, \mathcal{A})$ in the support of the analyzed meta-posterior, the centered clipped task loss $\widetilde{Z}_h(u, S_u^k, Q_u) - \mathbb{E}[\widetilde{Z}_h(u, S_u^k, Q_u)]$ is σ^2-sub-Gaussian over the joint draw of $u \sim \mathcal{P}_U$, $S_u^k \sim \mathcal{D}_u^k$, and a disjoint query set Q_u. A sufficient condition is the bounded-loss instantiation described in Appendix A.5. *(A3, finite KL.)* The analyzed meta-posterior Q and a data-independent meta-prior P satisfy $\mathrm{KL}(Q\|P) < \infty$. We record *(A2, support-Lipschitz operator)* separately in Appendix A.5; A2 is not used in Theorem 1 itself, but underlies Proposition 4.

Theorem 1 (PAC-Bayes bound for clipped adapted query loss)**.** *Under A1 and A3, fix $k \in \mathbb{N}$. Let $u_1, \ldots, u_m$ be i.i.d. meta-training users from $\mathcal{P}_U$, each with support set $S_{u_i}^k$ and disjoint query set Q_{u_i}. Then with probability $\geq 1 - \delta_{\mathrm{PB}}$, for every meta-posterior Q with $\mathrm{KL}(Q\|P) < \infty$,*

$$\mathbb{E}_{(\theta_0,\mathcal{A})\sim Q}\,\mathbb{E}_u\,\mathbb{E}_{S_u^k,Q_u}\left[\widetilde{\mathcal{L}}_u^{Q_u}\big(\mathcal{A}(\theta_0, S_u^k)\big)\right] \leq \hat{\widetilde{\mathcal{L}}}_{\mathrm{meta}}(Q) + C_1\sigma\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(2\sqrt{m}/\delta_{\mathrm{PB}})}{m}}, \tag{3}$$

where $\widetilde{\mathcal{L}}$ is the clipped empirical query loss and $\hat{\widetilde{\mathcal{L}}}_{\mathrm{meta}}(Q)$ is its meta-training average under Q.

Full proof of Theorem 1. For a hypothesis $h = (\theta_0, \mathcal{A})$, define the bounded task loss

$$\widetilde{Z}_h(u, S_u^k, Q_u) = \widetilde{\mathcal{L}}_u^{Q_u}\big(\mathcal{A}(\theta_0, S_u^k)\big) = \mathrm{clip}\Big(\hat{\mathcal{L}}_u^{Q_u}\big(\mathcal{A}(\theta_0, S_u^k)\big), 0, B\Big).$$

Under A1, the centered random variable $\widetilde{Z}_h - \mathbb{E}[\widetilde{Z}_h]$ is σ^2-sub-Gaussian over the joint draw of a task $T = (u, S_u^k, Q_u)$, where $u \sim \mathcal{P}_U$, $S_u^k \sim \mathcal{D}_u^k$, and Q_u is a disjoint query set. Let $T_1, \ldots, T_m$ be the i.i.d. meta-training tasks. Applying the McAllester-style PAC-Bayes inequality [McAllester, 1999] in the meta-learning form of Amit and Meir [2018] to the loss family $\{\widetilde{Z}_h\}_h$ yields that with probability at least $1 - \delta_{\mathrm{PB}}$ over $T_1, \ldots, T_m$, simultaneously for every Q with $\mathrm{KL}(Q\|P) < \infty$,

$$\mathbb{E}_{h\sim Q}\,\mathbb{E}_T[\widetilde{Z}_h(T)] \leq \mathbb{E}_{h\sim Q}\,\frac{1}{m}\sum_{i=1}^m \widetilde{Z}_h(T_i) + C_{\mathrm{PB}}(Q),$$

where $C_{\mathrm{PB}}(Q) = C_1\,\sigma\sqrt{(\mathrm{KL}(Q\|P) + \log(2\sqrt{m}/\delta_{\mathrm{PB}}))/m}$. Expanding the definitions of $\widetilde{Z}_h(T)$ and the empirical average gives (3). □

What the theorem does and does not say. Theorem 1 controls transfer of the clipped empirical query loss actually optimized in the bounded-loss instantiation of WARMSTART. It does not by itself imply a high-probability bound on the unclipped population loss $\mathcal{L}_u$ or on personalization error Δ_{Pers}. Upgrading the statement would require additional concentration over finite query sets together with control of the bias introduced by clipping.

Proposition 4 (Mean-pooled support concentration)**.** *Assume that each support example is mapped to an embedding $z_{u,j} \in \mathbb{R}^{d_{\mathrm{emb}}}$ with $\|z_{u,j}\|_2 \leq C_{\mathrm{emb}}$ almost surely, that the support summary is the mean $\bar{z}_u^k = \frac{1}{k}\sum_{j=1}^k z_{u,j}$, and that there is an idealized mean-conditioned parameter map $\widetilde{\mathcal{A}}(\theta_0, \bar{z})$ that is $L_{\mathcal{A}}$-Lipschitz in $\bar{z}$. Let $\mu_u = \mathbb{E}[z_{u,1} \mid u]$. Then*

$$\mathbb{E}_{u,S_u^k}\left[\left\|\bar{z}_u^k - \mu_u\right\|_2^2\right] \leq \frac{C_{\mathrm{emb}}^2}{k} \qquad \textit{and} \qquad \mathbb{E}_{u,S_u^k}\left[\left\|\widetilde{\mathcal{A}}(\theta_0, \bar{z}_u^k) - \widetilde{\mathcal{A}}(\theta_0, \mu_u)\right\|_\Theta\right] \leq \frac{L_{\mathcal{A}}C_{\mathrm{emb}}}{\sqrt{k}}.$$

Proof. Because $\|z_{u,j}\|_2 \le C_{\mathrm{emb}}$, $\mathbb{E}[\|z_{u,j} - \mu_u\|_2^2 \mid u] = \mathbb{E}[\|z_{u,j}\|_2^2 \mid u] - \|\mu_u\|_2^2 \le C_{\mathrm{emb}}^2$. Independence inside the support set gives

$$\mathbb{E}\Big[\big\|\bar{z}_u^k - \mu_u\big\|_2^2 \Big| u\Big] = \frac{1}{k^2}\sum_{j=1}^{k} \mathbb{E}\Big[\big\|z_{u,j} - \mu_u\big\|_2^2 \Big| u\Big] \le \frac{C_{\mathrm{emb}}^2}{k}.$$

Jensen's inequality yields $\mathbb{E}\|\bar{z}_u^k - \mu_u\|_2 \le C_{\mathrm{emb}}/\sqrt{k}$, and the Lipschitz assumption then gives the stated bound on the adapted parameter perturbation. □

Remarks. *Meta-generalization, not sample-complexity-in-k.* See the discussion following Theorem 1 in the main text. Proposition 4 records an $O(k^{-1/2})$ concentration rate in embedding and parameter space under the stronger mean-pooled bounded-embedding model used by WARMSTART; turning that into a loss-level statement would require additional regularity of $\mathcal{L}_u(\theta)$. *The variance-aware* CVaR *meta-objective is empirical, not part of the theorem.* The CVaR penalty introduced in Section 2 is an empirically motivated training choice: its formal role is to change the empirical clipped meta-loss by shifting Q, and it does not appear directly in $C_{\mathrm{PB}}(Q)$ except through the dependence of $\mathrm{KL}(Q\|P)$ on the chosen posterior family. Crucially, there is no λ-indexed posterior Q_λ in the theorem statement: Q is universally quantified. Stating a formal result that the bound tightens as a function of λ would require defining a λ-indexed family $\{Q_\lambda\}_{\lambda \ge 0}$ together with a uniform argument across that family, which is outside the scope of the present work. Whether the empirical effect of CVaR training translates into a deployment-time improvement is an empirical question evaluated in Section 3. *Lipschitz property recorded but not used in Theorem 1.* A2 plays no role in the PAC-Bayes proof; we record it because it yields the separate parameter-space concentration statement of Proposition 4.
Limitations of A1 and A3. A1 is strong for unbounded log-likelihood losses: it does not in general follow from gradient clipping on the operator parameters or from bounded sentence-encoder outputs, and may be violated by heavy-tailed user distributions in which a small fraction of users produce extremely long or atypical sequences. The operative analytic content we rely on in experiments is the bounded-loss sufficient condition (Section 2); we state this explicitly so that what the bound actually rests on is transparent to the reader. An extension to empirical-Bernstein PAC-Bayes bounds could weaken A1 to a finite-variance condition at the cost of more intricate constants. A3 is the standard PAC-Bayes regularity hypothesis and is enforced not by the optimizer alone but by an explicit choice of posterior family with controlled mean and covariance; the WARMSTART instantiation makes these choices explicit (Appendix A.5).

Reproducibility details. The next five subsections collect the implementation, data, metric, baseline, and protocol details needed to reproduce the main experiments and interpret the reported shuffle-gap numbers.

A.5 Method details and assumption compatibility

Support encoder implementation. All main WARMSTART and P2P runs use the frozen `sentence-transformers/all-MiniLM-L6-v2` checkpoint. Each support item is encoded independently with the checkpoint tokenizer and default truncation window (256 wordpiece tokens for this checkpoint), and embeddings are returned with `normalize_embeddings=False`. This yields one 384-dimensional vector per support item; the persona encoder then mean-pools over the k support vectors and applies a two-layer MLP (hidden dimension 512, output dimension 256). The query text is not provided to this mean-pooled persona encoder. For long Blog posts, early truncation by the frozen sentence encoder is therefore a plausible bottleneck and part of the collapse mechanism being audited.

Bounded per-user query loss as the operative sufficient condition for A1. We clip the per-user query loss appearing in (2) to $[0, B]$ before averaging and before feeding it into $\mathrm{CVaR}_{0.5}$; all main runs use $B = 10$ in token-averaged cross-entropy units and $\lambda = 1$ unless the row is explicitly marked ERM. Reported evaluation PPLs are computed from the unclipped query cross-entropy. By Hoeffding's lemma, the clipped training loss is σ-sub-Gaussian jointly in (u, S_u^k, Q_u) with $\sigma \le B/2$, which is a fully sufficient condition for A1. Without this explicit clip, gradient clipping on (φ, ψ) to $\|\nabla\| \le 1$ and a bounded sentence-encoder output $\|e_u\|_2 \le 1$ make sub-Gaussianity plausible but do not by themselves prove a sub-Gaussian moment generating function over the joint randomness of

users and per-user sequences; we therefore treat the loss clip as the operative sufficient condition, not a cosmetic engineering detail.

Parameter-map Lipschitz property of $\mathcal{A}_{\mathrm{WS}}$ (A2). The persona encoder E_φ and the hypernetwork H_ψ are Lipschitz MLPs; we apply optional spectral normalization to both during meta-training. Their composition yields $\|\mathcal{A}_{\mathrm{WS}}(\theta_{\mathrm{base}}, S) - \mathcal{A}_{\mathrm{WS}}(\theta_{\mathrm{base}}, S')\|_\Theta \le L_{\mathcal{A}} \cdot d_{\mathrm{sup}}(S, S')$ with $L_{\mathcal{A}}$ controlled by the product of layer operator norms. This is a Lipschitz condition on the adapted parameter, not on the per-user query loss: translating it into a bound on the loss would additionally require that $\mathcal{L}_u(\theta)$ is regular as a function of θ, a hypothesis we do not impose. Because A2 is not used in the proof of Theorem 1, A2 is therefore recorded only for completeness.

Posterior family for finite KL (A3). A sufficient concrete instantiation is to choose Q as a diagonal Gaussian over (φ, ψ) with finite mean $(\bar{\varphi}, \bar{\psi})$ and strictly positive finite covariance $\mathrm{diag}(\boldsymbol{\tau}^2)$, paired with a data-independent isotropic Gaussian prior $P = \mathcal{N}(0, \sigma_P^2 I)$. For finite-dimensional Gaussians, $\mathrm{KL}(Q\|P)$ is then finite. In practice we encourage this regime by combining ℓ_2 weight decay on $(\bar{\varphi}, \bar{\psi})$ with a per-coordinate variance floor $\boldsymbol{\tau}^2 \ge \tau_{\min}^2 > 0$ enforced after each optimizer step. SGD on (φ, ψ) alone does not itself prove A3; A3 comes from the explicit posterior family one chooses to analyze.

Positioning. WARMSTART adapts in a single forward pass (one pass through the persona encoder and hypernetwork to generate the adapter, then one pass through the adapted backbone to produce logits); unlike per-user LoRA fine-tuning [Hu et al., 2021] it avoids $O(\text{epochs})$ gradient steps per user at deployment. Unlike in-context learning it learns a task-adaptive feature space and produces explicit parametric adapters. Unlike In-Place Test-Time Training [Feng et al., 2026] it maintains a stable persona embedding that survives across sessions. Unlike Profile-to-PEFT [Tan et al., 2025], whose architectural template it closely resembles, WARMSTART is trained with the variance-aware meta-objective (2) rather than vanilla ERM and is covered by the PAC-Bayes meta-generalization bound of Theorem 1.

A.6 Data and secondary-benchmark protocols

The Blog Authorship Corpus [Schler et al., 2006] collects posts from 19,320 bloggers; we filter to users with at least sixty posts and partition $70/10/20$ into meta-training/meta-validation/meta-test splits. For each test user u we form three pairwise-disjoint slices: a query set Q_u of 5 posts, a support pool of 20 posts, and an oracle partition O_u of at least 35 posts. The three slices are constructed once per user before any sampling of S_u^k so no leakage is possible. The filtered Blog meta-test split has n=465 users (filter: ≥ 60 posts), tokenization uses the Qwen3 tokenizer with no per-post normalization beyond model defaults, and all filtering/splitting choices needed for reimplementation are specified here.

LaMP-7 (secondary). From the *LaMP* benchmark [Salemi et al., 2024] we use the LaMP-7 tweet paraphrasing task with the published train/dev/test files mapped to meta-train/meta-validation/meta-test respectively. The reported evaluation set is the filtered meta-test split: users with at least ten profile tweets (n=1253). For each user the query is the official (input, target) pair and the support set is the first k entries of `profile_texts` for $k \in \{1, 2, 5, 10\}$, byte-identical across WARMSTART, ICL, BM25-RAG, and support-LoRA so that the WS-vs-support-LoRA comparison is a literal architectural A/B on identical data. The target paraphrase is materialized only as the query row, while profile tweets are materialized only as support rows; our loader therefore does not copy the target into the support pool. Per-user cross-entropy masks the LaMP-7 instruction prompt so that only the target paraphrase contributes to the loss (labels = input ids on the target span, -100 on the prompt span), using the same row builder for WARMSTART, zero-shot, ICL, BM25-RAG, and support-LoRA query evaluation. During support-LoRA training the profile rows are trained as plain continuations with no prompt mask, making support-LoRA a support-only adaptation comparator rather than an equally meta-trained task adapter. rOR is not run on LaMP-7, since the three-way per-user partition required by H1 is not meaningful for tweet-length profiles. Support-LoRA uses 5 inner steps, learning rate 5×10^{-4}, batch size 8, LoRA rank 8 on $\{q_{\mathrm{proj}}, v_{\mathrm{proj}}\}$, and a per-item training-token cap of 64 (a 256-token control on k=5 reproduces the headline number to within 10^{-4}, 1252/1253 users bit-for-bit identical; the cap is not active on tweet-length profile items). This 64-token cap is specific to LaMP-7's tweet-profile support-LoRA training; Blog support-LoRA follows the Blog support/query windows used by the other Blog baselines. WARMSTART is run for five seeds

and ERM-WARMSTART for two seeds, each with MAX_STEPS $= 400$, BATCH_USERS $= 4$, learning rate 2×10^{-4}, and the same backbone, persona encoder, and hypernetwork as on Blog.

Secondary benchmarks beyond LaMP-7. The *WildChat-4.8M* corpus [Zhong et al., 2024], grouped by hashed IP address, supplies a realistic distribution of user interaction lengths and conversation contexts. We use n=200 English users with 15 turns per user for the discriminability analysis and a larger n=1000 English-user subset for the PPL shuffle-gap probe, using paired bootstrap CIs over users. The *PersonaFeedback* hard tier [Tao et al., 2025] provides pairwise preference judgments (x, y_+, y_-) in which y_+ is the user-preferred and y_- the user-rejected response. We map PersonaFeedback to our framework via two pairwise metrics computed from the adapted model's per-token log-probabilities: the *pairwise log-odds* $\Delta_\theta(x, y_+, y_-) = \log p_\theta(y_+ \mid x) - \log p_\theta(y_- \mid x)$, which provides a continuous adaptation signal sensitive to the relative ordering of the two candidates, and the *preference accuracy* $\mathrm{Acc}_\theta(x, y_+, y_-) = \mathbf{1}\{\Delta_\theta(x, y_+, y_-) > 0\}$, which provides a calibrated headline number on $[0, 1]$ that we report as the primary PersonaFeedback metric. Both metrics use both the chosen and the rejected response. We do not report the negative log-probability of the preferred response alone, since that quantity conflates personalization with raw fluency on y_+ and is not invariant to the overall sequence-likelihood scale of the adapted model. *PRISM* [Kirk et al., 2024] enables demographic-stratified evaluation and provides the basis for our group-level analysis. WildChat shuffle-gap results are reported in Section 3, and PersonaFeedback preference results are reported in Appendix A.28; PRISM and other LaMP subtasks are left for future work.

A.7 Metrics, CIs, and the sample-to-convergence threshold

For each test user u and support size k we report $\mathrm{PPL}_\theta(Q_u \mid S_u^k) = \exp(-|Q_u|^{-1} \sum_{(x,y)\in Q_u} \log p_\theta(y \mid x))$. The per-user PPL reduction is $\rho_u(k) = \mathrm{PPL}_{\theta_{\mathrm{base}}}(Q_u) - \mathrm{PPL}_{\text{adapted}}(Q_u \mid S_u^k)$; positive values indicate improvement, while negative values indicate that adaptation hurt this user. We define the empirical sample-to-convergence threshold as $k^\star(\epsilon) = \min\{k : \mathbb{E}_u[\mathrm{PPL}_{\text{adapted}}(Q_u \mid S_u^k)] - \mathbb{E}_u[\mathrm{PPL}_u^{\mathrm{rOR}}] \le \epsilon\}$, where the per-user reference $\mathrm{PPL}_u^{\mathrm{rOR}}$ is the perplexity on Q_u of the per-user restricted oracle defined in Section 3. The rOR is a model-class-restricted (rank-r LoRA over a frozen backbone), finite-data (O_u) empirical benchmark rather than an estimator of any population-level quantity from Proposition 1.

We report two kinds of confidence intervals because they answer different questions. A *within-seed paired bootstrap* over the 465 meta-test users (with 2000 resamples unless otherwise stated) controls user-sampling noise at a fixed seed and is the appropriate CI when we ask "is the method delta real at this seed?" A *between-seed Student-t* interval over the random seeds (five for WARMSTART and ERM-WARMSTART, three for the rOR reference and the seed-matched ICL/BM25-RAG/support-LoRA/P2P k=5 summaries where reported) controls seed-level noise from data partition, hypernet initialization, and meta-training stochasticity, and is the appropriate CI when we ask "is the method delta robust to re-running the experiment?" The two CIs can disagree: the paired bootstrap is typically tighter because it pools user variance at a fixed seed, while the Student-t interval is more conservative but directly addresses reproducibility. For the WARMSTART-vs-rOR comparisons we report the Student-t CI because it is the more demanding statistic. The raw-shuffle ICL and support-LoRA probes in Table 2 are separate fixed-sample diagnostics, not the seed-wise components of Table 1's aggregate PPL rows.

A.8 Baseline implementation details

BM25-RAG. We use Okapi BM25 with k_1=1.5, b=0.75 over lowercase-alnum tokenization. For each query post, the support pool is ranked independently and the top-k posts are prepended to the prompt. *Hyperparameter robustness.* At k=5 on Blog (seed 0, n=465), mean PPL with the default $(k_1, b) = (1.5, 0.75)$ is 25.97. Two off-default settings span the standard range used in the information retrieval (IR) literature: $(1.2, 0.6)$ gives 26.03 PPL and $(2.0, 0.9)$ gives 25.90 PPL, a spread of only 0.13 PPL. The BM25 baseline is therefore not tuned into the comparison reported in Table 1.

Support-LoRA. A fresh per-user rank-r LoRA is trained from scratch on the exact sampled S_u^k that WARMSTART sees at each (u, k), using a byte-identical RNG sequence. The number of inner steps and learning rate are tuned on a 20-user meta-validation bench at k=5 and frozen at (num_steps=10, lr=5×10^{-4}) before running on the full meta-test split. On Blog, support-LoRA

Table 5: Shuffle-gap calibration on Blog (Qwen3-8B, seed 0, k=5). Log-CE CIs are paired bootstraps over users (2000 resamples). Aggregate PPL delta is $\exp(\bar{\ell}_{\text{condition}}) - \exp(\bar{\ell}_{\text{real}})$.

Condition	Mean log-CE gap	95% CI	Aggregate PPL delta
Exact repeat	0	$[0, 0]$	0
Within-user support-order permutation	1.4×10^{-5}	$[-1.6, 1.9]\times10^{-4}$	0.0003
Within-user support resample	4.7×10^{-6}	$[-2.9, 3.1]\times10^{-4}$	0.0001
Random derangement seed 17	6.1×10^{-4}	$[1.2, 11.6]\times10^{-4}$	0.0134
Random derangement seed 42	5.1×10^{-4}	$[-0.7, 10.9]\times10^{-4}$	0.0112
Random derangement seed 73	5.1×10^{-4}	$[-0.3, 10.7]\times10^{-4}$	0.0111
Random derangement seed 99	4.5×10^{-4}	$[-1.2, 10.3]\times10^{-4}$	0.0099

uses the same sampled Blog support posts and model-token windows as the other Blog baselines; no LaMP-style 64-token support cap is applied.

rOR (restricted-oracle LoRA). The rOR is fitted on each test user's oracle partition O_u (at least 35 posts, disjoint from the support pool and query set) with the same LoRA insertion points and rank as WARMSTART. The optimization budget (number of gradient steps, learning rate schedule) is selected on meta-validation users before being frozen and applied to meta-test. This is a finite-data, from-scratch per-user LoRA reference, not a true upper bound: unlike WARMSTART, it does not inherit a learned population prior or meta-initialization, which is one plausible reason it can underperform the amortized hypernetwork despite seeing more same-user text.

A.9 Shuffle-gap evaluation protocol

The shuffle-gap diagnostic (Section 3) proceeds as follows. Given a trained WARMSTART checkpoint and the n=465 meta-test users, for each user u we (1) sample a support set S_u^k from the support pool using the same RNG as standard evaluation, (2) compute the real per-user PPL $\text{PPL}_{\text{real}}(u, k)$ by adapting with S_u^k, (3) draw a permutation π over the meta-test users such that $\pi(u) \neq u$ for all u (a derangement), (4) compute the shuffled per-user PPL $\text{PPL}_{\text{shuf}}(u, k)$ by adapting u's query with $S_{\pi(u)}^k$, and (5) report the shuffle gap $\Delta_{\text{shuf}}(k) = \frac{1}{n}\sum_u [\text{PPL}_{\text{shuf}}(u, k) - \text{PPL}_{\text{real}}(u, k)]$. The derangement is drawn with a fixed seed (default 17) and is identical across all methods and checkpoints so that differences in shuffle gap reflect only the model, not the permutation. For each k we report both the aggregate gap and per-user paired differences to enable bootstrap CIs. To verify robustness to the permutation choice, we repeat the collapsed WARMSTART evaluation with four independent derangement seeds (17/42/73/99); the resulting gaps (0.008/0.011/0.011/0.010 PPL at k=5) are indistinguishable, confirming the near-zero finding is not an artifact of a particular permutation.

Margin-loss training. The shuffled-support margin penalty is computed inside each meta-training step. Given a meta-batch of B users, we draw a random derangement π over the batch (uniform random, not within-cluster) and compute the shuffled-support loss $\mathcal{L}_u(\theta;\, S_{\pi(u)}^k)$ for each user u. The hinge loss $\max(0,\, \delta - [\mathcal{L}_u(\theta;\, S_{\pi(u)}^k) - \mathcal{L}_u(\theta;\, S_u^k)])$, weighted by β ("shuffle_margin_lambda" in configs), is added to the meta-objective. We set δ=0.5 throughout ("shuffle_margin" in configs). A short tuning sweep varies $\beta \in \{2, 3, \ldots, 15\}$ at 200 steps; the full-budget runs use β=10 at 500 steps. Schedule variants anneal β linearly from a high initial value to a low final value over training.

A.10 Shuffle-gap calibration and measurement floor

To calibrate the scale of the Blog shuffle gap, we run four controls on the seed-0 WARMSTART checkpoint at k=5 (n=465). The first two estimate deterministic evaluation noise: exact repeat of the same support assignment and a permutation of support order within each user's support set. The third draws a fresh k-shot support set from the same user's support pool; this is not measurement noise, but it estimates the variability induced by same-user support sampling. The final rows are the usual random user-to-user derangements.

The exact-repeat control is identically zero, and the order-permutation and same-user-resampling controls have aggregate PPL deltas below 0.001. The random derangement gap is therefore above the deterministic evaluation floor, but its practical magnitude remains tiny: roughly 0.01 PPL, more

Table 6: Controlled synthetic support-signal check. Random shuffling replaces a user's support with a different user's support; label-matched shuffling replaces it with another support from the same keyword class. Loss and PPL are computed only on answer tokens.

Condition	Mean loss	Target-token PPL	Gap vs. real
Zero-shot	1.491	4.44	–
Real support	0.035	1.04	–
Random shuffled support	2.309	10.06	$+2.27$ log CE, CI $[0.44, 4.67]$
Label-matched shuffled support	0.035	1.04	0.00 log CE, CI $[-0.0008, 0.0008]$

than two orders of magnitude below the topic-controlled ICL reference (3.62 PPL) and far below the gap-forced intervention. This calibration supports the conservative reading used in the main text: the collapsed hypernetwork is not literally invariant to support at infinite precision, but it is nearly insensitive to *whose* support it receives under the evaluated loss.

A.11 Synthetic support-signal sanity check

The shuffle gap is observational on real corpora, so we add a controlled synthetic check to verify that the diagnostic can distinguish genuine causal support use from functional support interchangeability. We create 12 synthetic keyword classes and 20 users per class. Each support item states the user's keyword in plain text; each query asks for the keyword, and only the answer tokens are scored. The Qwen3-8B WARMSTART architecture is unchanged (rank-8 LoRA on all $q_{\text{proj}}/v_{\text{proj}}$ layers, d=256 persona embedding); we train for 300 meta-steps with k=5 support items and evaluate on 48 held-out users.

This experiment is not evidence about Blog itself; it is a calibration of the diagnostic and architecture. When the support channel contains a simple query-relevant signal, the same hypernetwork produces a large random-shuffle gap. When the swapped support is functionally equivalent (same keyword class), the gap vanishes. This is the behavior the audit should have: large gaps for causal support mismatch and null gaps for truly interchangeable support.

A.12 Reproducibility details

All reported experiments were run on NVIDIA H100 GPUs, with Qwen3-8B and Llama-3.1-8B-Instruct as frozen backbones and rank-r LoRA adapters. The main WARMSTART, ERM-WARMSTART, P2P, and gap-forcing runs use one H100 per run; main training runs use 500 steps unless otherwise stated, while support-LoRA and rOR consume per-user inner-loop optimization at evaluation time. Concrete configurations for every reported number are documented in this appendix: LoRA insertion points and rank (r=8 on $\{q_{\text{proj}}, v_{\text{proj}}\}$), persona dimension ($d$=256), main-training budget (500 steps, batch size 4, learning rate 2×10^{-4}), seed grid (5 seeds for WARMSTART, 3 for Table 1's ICL/BM25-RAG/support-LoRA and P2P k=5 summaries, 1 for the raw/topic-controlled shuffle probes and the full P2P support-size curve, 5 for the LaMP-7 global/no-user LoRA), margin-loss intervention (δ=0.5, β=10), BM25 settings (k_1=1.5, b=0.75 with the $(1.2, 0.6)/(2.0, 0.9)$ robustness sweep in Appendix A.8), shuffle-gap derangement seeds (17/42/73/99), and meta-test split sizes (n=465 Blog, n=1253 LaMP-7, n=1000 WildChat PPL probe). We do not currently plan to release code or trained checkpoints; the parameters above are provided so the reported numbers can be independently audited and reimplemented from the public datasets.

Supplementary empirical results and diagnostics. The remaining empirical subsections collect ablations, cross-benchmark tables, and auxiliary diagnostics that support the main shuffle-gap claim without interrupting the main-paper narrative.

A.13 Ablations

We ablate four design axes of WARMSTART to isolate each component's independent contribution to personalization performance. Results are reported in the ablation tables accompanying this section; full hyperparameter details for each sweep match the main experimental configuration described in Appendix A.12.

Table 7: Ablation study on Blog meta-test (n=465, Qwen3-8B, k=20, seed 0). Each group varies one hyperparameter from the default (α=0.5, r=8, d=256); all others are held fixed. Bold marks the default row.

Axis	Setting	Mean PPL	$\mathrm{CVaR}_{0.1}$ PPL
Default (α=0.5, r=8, d=256)		**21.82**	**49.54**
CVaR α	0.10	21.86	49.59
	0.25	21.85	49.64
	0.75	21.80	49.51
LoRA rank r	4	21.79	49.65
	16	21.93	49.72
	32	22.01	50.54
Persona dim d	64	21.86	49.91
	128	21.83	49.86
	512	21.79	49.61
+ Spectral normalization		22.80	53.28

CVaR level α. We ablate $\alpha \in \{0.1, 0.25, 0.5, 0.75\}$ to isolate the contribution of the tail-reweighting severity to personalization quality and worst-user robustness. The CVaR level controls what fraction of worst-performing users the penalty targets during meta-training: smaller α focuses training pressure on a narrower, harder tail of the user distribution, whereas larger α approaches plain-mean ERM as $\alpha \to 1$. This sweep tests whether variance-aware training yields improved worst-user robustness across a finer grid than the binary ERM-vs-WARMSTART comparison in the main body, and reveals whether gains saturate or reverse as the tail becomes more aggressively targeted.

LoRA rank r. We ablate $r \in \{4, 8, 16, 32\}$ to isolate the contribution of adapter capacity to the personalization–generalization trade-off. The rank r determines the number of free parameters the hypernetwork must generate per attention layer, directly setting the expressiveness of the per-user adapted model. Lower ranks reduce the risk of overfitting a user's support set but limit how far the adapter can move the backbone toward a user's idiosyncratic distribution; higher ranks increase expressiveness at the cost of a larger hypernetwork output head and potentially noisier adaptation when support sets are small. The sweep identifies the operating point at which increasing rank no longer yields PPL gains on the meta-test query set, informing the choice of r for the main experiments.

Persona embedding dimension d. We ablate $d \in \{64, 128, 256, 512\}$ to isolate the contribution of the persona encoder's bottleneck width to representation quality. The embedding dimension d controls the information capacity of the fixed-size persona vector $e_u = E_\varphi(S_u^k)$ that summarizes a user's support set before being passed to the hypernetwork H_ψ. Too small a d forces the encoder to discard user-specific signal; too large a d may cause the hypernetwork to overfit to idiosyncratic surface features of the support set rather than stable user style. This ablation reveals the sensitivity of WARMSTART to the bottleneck width and whether the shuffle gap–a diagnostic of genuine personalization versus support-set memorization–varies systematically with d.

Meta-training set size m. We train WARMSTART from scratch at $m \in \{100, 300, 600, 1000, 1628\}$ (where m=1628 is the full meta-training split) with all other hyperparameters held at their default values. This sweep is motivated by the $1/\sqrt{m}$ dependence in Theorem 1 and tests how many meta-training users are needed before the hypernetwork produces useful per-user adapters.

Table 7 reports results for all ablation configurations on Blog (n=465, seed 0). The default configuration (α=0.5, r=8, d=256) sits near the Pareto optimum across all three hyperparameter axes.

Key observations. (i) *CVaR α*: all four levels produce near-identical mean PPL (range 21.80–21.86). The default α=0.5 offers marginally better $\mathrm{CVaR}_{0.1}$ than the more aggressive α=0.10 or α=0.25; this is consistent with the collapsed-shuffle-gap finding (Section 3) that the persona encoder has not yet learned user-specific signal that the CVaR penalty could differentially exploit. (ii) *LoRA rank*: ranks 4 and 8 perform similarly; ranks 16 and 32 progressively degrade, with r=32 incurring +1.0 PPL on $\mathrm{CVaR}_{0.1}$, suggesting that the hypernetwork's parameter-generation burden outpaces the adaptation capacity gained. (iii) *Persona dim*: the range 128–512 is indistinguishable on mean PPL (≤ 0.04 spread); only d=64 shows a mild $\mathrm{CVaR}_{0.1}$ penalty (+0.37). (iv) *Spectral normalization* is the

Table 8: Per-seed WARMSTART mean PPL on Blog (n=465) and LaMP-7 (n=1253, k=5). Low across-seed variance confirms the meta-training procedure is stable.

	Blog			LaMP-7 (k=5)	
Seed	k=0	k=1	k=20	Mean PPL	$\mathrm{CVaR}_{0.1}$
0	35.12	21.84	21.82	8.04	35.77
1	32.92	20.71	20.70	7.99	35.04
2	32.82	20.59	20.58	8.01	35.90
3	34.79	21.34	21.32	8.11	34.87
4	34.52	21.29	21.28	8.06	35.50
Mean	34.03	21.15	21.14	8.04	35.42
Std	1.08	0.51	0.51	0.05	0.45

Table 9: Meta-training sample complexity (Blog, Qwen3-8B, seed 0). Mean and $\mathrm{CVaR}_{0.1}$ per-user PPL at k=1 and k=5 versus number of meta-training users m. Zero-shot (no adapter): 35.12 mean PPL.

	k=1		k=5	
m	Mean	$\mathrm{CVaR}_{0.1}$	Mean	$\mathrm{CVaR}_{0.1}$
100	912.70	4065.56	909.34	3992.64
300	39.28	114.83	39.13	114.10
600	26.39	62.79	26.38	63.35
1000	22.01	50.05	22.00	50.07
1628	21.84	49.57	21.83	49.56

only variant with a large effect: $+0.98$ mean PPL and $+3.74$ $\mathrm{CVaR}_{0.1}$, indicating that constraining the operator norms of the hypernetwork layers materially limits adapter expressiveness without a compensating stability benefit in this regime. Overall, all non-spectral-norm variants fall within a 0.22-PPL band on mean and 1.1-PPL band on $\mathrm{CVaR}_{0.1}$, confirming that WARMSTART is not brittle with respect to these design choices.

Single-seed caveat. All ablation configurations are evaluated at seed 0; the within-ablation spread (0.22 PPL on mean) is smaller than the between-seed standard deviation (0.51 PPL, Table 8), so the reported insensitivity should be interpreted as a single-seed observation pending multi-seed confirmation.

Seed stability. Table 8 reports per-seed results across five independent meta-training runs on both Blog and LaMP-7, confirming that the meta-training procedure is reproducible: across-seed standard deviation is ≤ 0.51 PPL on Blog and ≤ 0.05 PPL on LaMP-7 at $k \geq 1$.

Meta-training sample complexity. Table 9 reports WARMSTART mean PPL at k=1 as a function of the number of meta-training users m, as a qualitative check on the $1/\sqrt{m}$ dependence in Theorem 1. A sharp phase transition occurs between m=300 and m=600 (Figure 6): at m=100 the hypernetwork produces catastrophically bad LoRA weights (mean PPL 912.7, far above zero-shot 35.1); at m=300 the adapted model remains worse than zero-shot (39.3 vs. 35.1); by m=600 the hypernetwork becomes useful (26.4, on par with the restricted-oracle LoRA); and m=1000 nearly matches the full-data result (22.0 vs. 21.8). A control with $10\times$ lower learning rate (10^{-5}) at m=100 eliminates the catastrophic failure (mean PPL 35.2 at k=1, essentially zero-shot), confirming that the 912 PPL is an optimization instability rather than a clean statistical overfitting effect. The phase transition between m=300 and m=600 remains robust: at m=300 the adapter is worse than zero-shot regardless of learning rate, while by m=600 it becomes useful.

Compute-matched comparison. A natural concern is whether WARMSTART's PPL gains simply reflect more compute than the baselines. At inference time, WARMSTART and support-LoRA are already compute-matched: both apply an identical rank-r LoRA at the same insertion points to score the query on the same k support items; the only asymmetry is *amortized meta-training*. WARMSTART pays a one-time meta-training budget (500 steps over 1,628 users, Qwen3-8B backbone) to learn the persona encoder and hypernetwork, which is then reused across all test users at negligible marginal cost. Support-LoRA spends zero training compute ahead of time but pays a per-user per-query cost

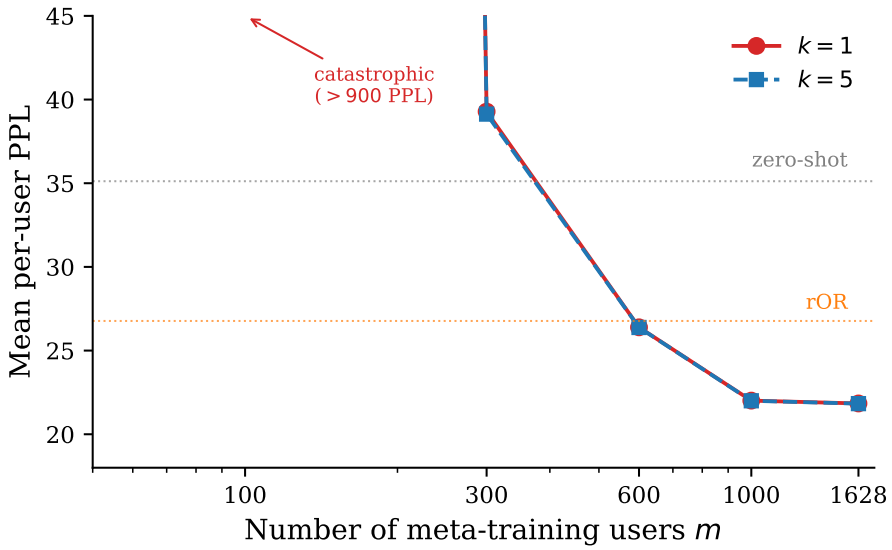


Figure 6: Mean per-user PPL versus number of meta-training users m (Blog, Qwen3-8B, seed 0). A phase transition between m=300 and m=600 separates a regime where the hypernetwork is actively harmful from one where it is useful. The overall trend is broadly consistent with the $1/\sqrt{m}$ dependence of Theorem 1, though the m=100 failure (PPL $>$ 900, off-scale) is primarily an optimization instability (Section A.13).

Table 10: Trainable support-encoder ablation on Blog (Qwen3-8B, seed 0, n=465). MiniLM is unfrozen during training. Bootstrap CIs use 2000 resamples over users.

k	Real PPL	Shuffled PPL	Gap (PPL)	Gap log-CE (95% CI)
1	21.811	21.840	0.029	0.00134 [0.00023, 0.00249]
5	21.792	21.814	0.022	0.00100 [0.00030, 0.00172]
20	21.786	21.805	0.019	0.00089 [0.00030, 0.00152]

(10 AdamW steps on S_u^k); this cost accumulates linearly with the number of users served. By the time $n_{\text{deploy}}\approx$1,628 users have been served, support-LoRA has consumed roughly the same cumulative compute as WARMSTART's meta-training, yet WARMSTART still outperforms it by 4.2 PPL at k=20 (Table 1). ICL uses zero extra training compute by definition and converges to 23.77 PPL at k=20, still 2.6 PPL above WARMSTART's k=1 asymptote. rOR consumes a large per-user fine-tuning budget (fitted on each user's O_u of $\geq$ 35 posts) yet reaches only 26.77 mean PPL–5.6 PPL above WARMSTART. These comparisons confirm that the PPL advantage does not arise from a compute advantage at query time; it arises from amortizing meta-training across users, which is the intended operating regime of any deployed personalization system.

A.14 Trainable support-encoder ablation

The main WARMSTART pipeline uses a frozen MiniLM support encoder. To test whether collapse is merely a consequence of freezing this encoder, we train a variant in which the MiniLM support encoder is updated jointly with the persona MLP and hypernetwork for 500 steps, then evaluate the same real-vs-shuffled protocol on Blog. Table 10 shows that the model remains essentially collapsed: at k=5 the shuffled-support penalty is only 0.022 PPL, with a log-CE gap of 0.0010. The confidence intervals exclude zero because the user count is large, but the effect size is two orders of magnitude below the ICL reference and far below the gap-forced intervention. Thus the frozen support encoder is not the sole explanation for the near-zero shuffle gap.

Table 11: Per-k mean PPL on LaMP-7 (Qwen3-8B frozen, n=1253). WARMSTART: five seeds. ERM-WARMSTART: two seeds. ICL/BM25-RAG/support-LoRA: three seeds.

Method	k=1	k=2	k=5	k=10
Zero-shot	30.27	30.27	30.27	30.27
ICL	31.04	26.28	23.34	21.59
BM25-RAG	29.32	25.07	22.75	21.60
support-LoRA	25.85	–	22.52	22.57
P2P	–	–	8.00	–
ERM-WARMSTART	–	–	8.02	–
WARMSTART	8.05	8.05	8.04	8.04

Table 12: Meta-generalization gap: mean per-user PPL of the same WARMSTART checkpoint evaluated on meta-training vs. meta-test users (Blog, Qwen3-8B, seed 0). The observed gap is approximately constant across k.

k	Meta-train (m=1628)	Meta-test (n=465)	Gap
0	35.11	35.12	0.01
1	19.57	21.84	2.27
5	19.54	21.83	2.29
20	19.54	21.82	2.28

A.15 Extended LaMP-7 results

Table 11 reports per-k PPL for all methods on LaMP-7 (n=1253, Qwen3-8B frozen). ICL k=1 (31.04) is *worse* than zero-shot (30.27), a second independent instance of the "one noisy support example acts as a misaligned topical prior" non-monotonicity seen for BM25-RAG k=1 on Blog. The CVaR penalty is direction-consistent at two ERM-WARMSTART seeds: ERM-WARMSTART is marginally better on mean ($\Delta_{\text{mean}} \approx +0.02$–$0.03$ PPL in WARMSTART's disfavor) while WARMSTART is strictly better on $\text{CVaR}_{0.5}$, $\text{CVaR}_{0.1}$, p_{99}, and worst-user at both seeds ($\Delta\text{CVaR}_{0.1} \approx -0.7$–$-0.8$ PPL at k=5); we report this as a two-seed direction replication, not a significance claim. The support-LoRA token cap and its control are described in Appendix A.6. Because the hypernetwork methods are both much lower PPL than non-hypernetwork baselines and essentially flat in k, the LaMP-7 result indicates strong task-level amortization. The explicit learned global/no-user adapter in Appendix A.17 strengthens that interpretation: it reaches 8.02 ± 0.07 PPL over five seeds, matching the WARMSTART band without receiving support. The split, target-mask, and support-query materialization checks above are included to rule out the most direct leakage and masking explanations for the unusually large gap.

A.16 Meta-generalization gap

As a qualitative complement to Theorem 1, we evaluate the seed-0 WARMSTART checkpoint on meta-training users (m=1628) and meta-test users (n=465) at identical k values.

At k=0 (no adaptation, base model only) the gap vanishes (0.01 PPL), confirming that the backbone generalizes equally well to both populations. Once the hypernetwork is activated ($k \geq 1$), the gap opens to ≈ 2.3 PPL and remains flat across k. This is a descriptive sanity check rather than a direct identification of the abstract quantities in Theorem 1.

A.17 Global (no-user) LoRA control

To separate the contribution of *population* adaptation from *per-user* adaptation, we train a single global rank-8 LoRA on the same q/v insertion points and the same backbone as WARMSTART, using 500 steps of pooled meta-training data with no user conditioning whatsoever (lr 5×10^{-4}, 307,702 training documents, batch size 4, $\leq$ 512 training tokens per document). At evaluation, the same global LoRA is applied unchanged to every meta-test user. Table 13 reports mean per-user PPL and $\text{CVaR}_{0.1}$ alongside zero-shot and WARMSTART references.

Table 13: Global (no-user) LoRA control on Blog meta-test (n=465, Qwen3-8B, seed 0). Same architecture and insertion points as WARMSTART, trained on pooled documents with no user conditioning.

Method	Mean PPL	$\mathrm{CVaR}_{0.1}$ PPL
Zero-shot	35.12	108.29
Global (no-user) LoRA	23.79	60.25
WARMSTART (k=5, 5-seed)	21.14	48.12

Table 14: Seed-matched comparison at k=5 on Blog meta-test (n=465, Qwen3-8B frozen, seeds $\{0, 1, 2\}$ for all four methods). Per-method 95% confidence intervals are between-seed Student-t on the seed-wise mean PPL (n=3 seeds, df=n−1=2, two-sided $t^*_{0.025,2}$=4.303). Paired comparisons report the within-seed difference $d_s = \overline{\mathrm{PPL}}_{\mathrm{WS},s} - \overline{\mathrm{PPL}}_{\mathrm{baseline},s}$ for $s \in \{0, 1, 2\}$ with paired Student-t CI $\bar{d} \pm t^*_{0.025,2}\, s_d/\sqrt{3}$ (df=2, two-sided t^*=4.303); negative means WARMSTART is better. All three paired CIs exclude zero, so the ranking is robust under matched-seed evaluation. Seed-wise paired differences (verifiable from the per-method PPL columns): WS−ICL $= (-5.506, -4.942, -5.004)$; WS−BM25-RAG $= (-4.256, -3.565, -3.280)$; WS−support-LoRA $= (-6.879, -6.173, -6.336)$.

Method	Mean PPL (3-seed) ± 95% CI	Paired Δ (WARMSTART−method)
WARMSTART (seeds 0–2)	21.00 ± 1.55	n/a
ICL	26.15 ± 2.31	-5.15 ± 0.77 CI $[-5.92, -4.38]$
BM25-RAG	24.70 ± 2.79	-3.70 ± 1.25 CI $[-4.95, -2.45]$
support-LoRA	27.46 ± 2.44	-6.46 ± 0.92 CI $[-7.38, -5.54]$

The global LoRA closes most of the zero-shot→WARMSTART mean-PPL gap without ever being told who the user is. Under the seed-0 comparison, it closes about 85% of the gap (35.12→23.79 vs. 35.12→21.83); relative to the five-seed WARMSTART mean in Table 1, it closes about 81%. We therefore describe the control as explaining about four-fifths of the Blog gain. We treat it as a *global adapter control* rather than evidence of personalization: most of WARMSTART's headline PPL improvement reflects a population-level distributional shift toward Blog-Authorship-style text that any sufficiently expressive single LoRA could learn from the same pooled meta-training data. The user-conditioned residual is about 2–3 PPL on the mean, depending on the seed aggregation, and 12.13 PPL on $\mathrm{CVaR}_{0.1}$ in the five-seed comparison; this is the part of the gain that the persona encoder and hypernetwork must claim. The shuffle-gap diagnostic of Section 3 shows that this residual is itself essentially independent of *whose* support the model receives.

On LaMP-7, the same no-user control is even more direct. We train a rank-8 global LoRA on pooled LaMP-7 profile/query material with the same backbone and insertion points, no support conditioning, 400 steps, batch size 4, and lr 5×10^{-4}. Across five seeds it reaches mean PPL 8.02 ± 0.07 (SD) and $\mathrm{CVaR}_{0.1}$ PPL 36.63 ± 0.46, while the zero-shot reference is 30.27 and WARMSTART reports 8.04 at k=5 in Table 11. Thus the LaMP-7 hypernetwork gain is reproducible by a single no-user adapter; together with the 0.009 PPL shuffle gap, this makes LaMP-7 a particularly clean population-adaptation control.

A.18 Seed-matched robustness comparison

Table 1 reports WARMSTART with five seeds and ICL/BM25-RAG/support-LoRA with three seeds, raising the concern that a portion of WARMSTART's lead might reflect favorable seed coverage rather than a genuine effect. To check this, we re-aggregate WARMSTART using only seeds $\{0, 1, 2\}$ so all four methods are matched at n=3 seeds, and report (a) per-method mean PPL with between-seed Student-t confidence intervals, and (b) paired Student-t confidence intervals on the within-seed difference $\overline{\mathrm{PPL}}_{\mathrm{WS},s} - \overline{\mathrm{PPL}}_{\mathrm{baseline},s}$ at the same seed s (both methods use the same 465 meta-test users at the same seed, so within-seed differences are paired across methods).

The 3-seed-matched WARMSTART mean of 21.00 PPL is essentially indistinguishable from the 5-seed value of 21.14 PPL reported in Table 1 (paired $|\Delta| < 0.15$ PPL), confirming that the additional seeds 3 and 4 do not materially shift the headline number. Per-method between-seed Student-t intervals are

Table 15: Mean shuffle gap of the collapsed WARMSTART (Blog, Qwen3-8B, seed 0, k=5) binned by per-user precision@5 quartile. The gap is not uniformly zero across bins–Q3's CI excludes zero–but does not show a monotonic increase with discriminability (Pearson $r \approx -0.01$, p=0.88, n=465). 95% bootstrap CIs (2000 resamples).

Quartile	Range	n	Mean gap (PPL)	95% CI
Q1 (low)	precision@5 < 0.020	82	0.009	$[-0.017, 0.036]$
Q2	precision@5 $\in [0.020, 0.050)$	130	-0.006	$[-0.028, 0.017]$
Q3	precision@5 $\in [0.050, 0.120)$	131	0.060	$[\ 0.011, 0.128]$
Q4 (high)	precision@5 ≥ 0.120	122	0.014	$[-0.015, 0.044]$

wide because n=3 inflates the t-critical value (4.303), but the within-seed paired comparison–which controls for shared user-level noise across methods at a fixed seed–is much tighter and excludes zero for all three baselines: WARMSTART is statistically distinguishable from each baseline at 95% confidence under matched-seed paired evaluation. We therefore retain the headline 5-seed numbers in Table 1 and treat the seed-matched table here as the robustness check against the asymmetric-coverage concern.

A.19 Support discriminability and binned shuffle gap

A near-zero shuffle gap admits two readings: the model could be ignoring support information (collapse), or the available supports could be functionally interchangeable. We separate these by measuring *support discriminability* on the Blog meta-test split using the same MiniLM-L6-v2 encoder that WARMSTART reads support through. The candidate pool contains exactly the $465 \times 20 = 9{,}300$ Blog meta-test support items: no query items and no oracle items. Each support item is one Blog post, embedded independently with the MiniLM checkpoint tokenizer and its default truncation; cosine similarity is computed on the resulting 384-dimensional embeddings. For an anchor support item, we retrieve its 5 nearest neighbors from this pool with the anchor itself excluded. Since each anchor has 19 same-user positives, if h of the top-5 neighbors belong to the same user, then item-level $\text{precision@5} = h/5$ and item-level $\text{recall@5} = h/19$. Random retrieval gives expected precision@5 $19/(9{,}300-1) \approx 2.0 \times 10^{-3}$ and expected recall@5 $5/(9{,}300-1) \approx 5.4 \times 10^{-4}$.

Aggregate (n=465 users): item-level precision@5 has mean 0.102 and median 0.000; item-level recall@5 has mean 0.027 and median 0.000. Averaging precision over each user's 20 support items gives mean user precision@5 0.102 and median 0.050, about 50× the random baseline. A split-half nearest-centroid test gives a more direct set-level check: we split each user's 20 support items into two halves, mean-pool each half, and ask whether one half's centroid retrieves the other half's user among all 465 candidates. Across five split seeds, top-1 accuracy is $51.6 \pm 1.3\%$ and top-5 accuracy is $71.3 \pm 0.8\%$ (chance: 0.2% and 1.1%). Thus Blog supports are not interchangeable in the representation available to the adaptation pipeline. These observations are facts about the encoder, not the WarmStart head.

Binning users by mean precision@5 quartile and reporting the mean shuffle gap of the collapsed WARMSTART (seed 0, k=5) per bin gives Table 15. The pattern is *not uniformly zero*: bin Q3 has 95% bootstrap CI $[0.011, 0.128]$, which excludes zero. The other three bins have CIs that include zero. We therefore report the result as "no monotonic association between discriminability and shuffle gap," rather than as "flat across quartiles." Across the per-user sample (n=465 Blog meta-test users), the Pearson correlation between per-user precision@5 and per-user shuffle gap is $r = -0.007$ (PPL scale; $t(463) = -0.15$, p=0.88, Fisher 95% CI $[-0.098, 0.084]$). Scaling precision@5 to recall@5 leaves the correlation unchanged. The data do not support a monotonic discriminability→gap relationship, but they also do not certify uniform zero gap across the discriminability dimension.

In the Blog evaluated setting, the absence of a monotonic discriminability→gap relationship argues against support interchangeability as the *primary* explanation for the small shuffle gap: even users whose supports are most distinguishable from other users' (top precision@5 quartile) receive a mean shuffle gap of 0.014 PPL, comparable to Q1 and Q2. We do not claim binned estimates are uniformly zero (Q3 is positive). We claim the data do not support an interpretation in which the small shuffle gap arises because supports are functionally interchangeable; if that were the dominant mechanism, the gap should rise monotonically with discriminability, which it does not.

Table 16: Support discriminability across two corpora using the same MiniLM encoder and the same metrics. WildChat is materially more heterogeneous than Blog on every measure. The WildChat PPL shuffle-gap probe is reported separately in Table 17.

Metric	Blog (n=465)	WildChat (n=200)
Within-user mean cosine	0.218	0.282
Between-user centroid mean cosine	0.489	0.159
Discriminability ratio (within/between)	**0.53** (median 0.40)	**2.65** (median 1.62)
Item-level precision@5 (mean)	0.102 ($\sim 50\times$ chance)	**0.522** ($\sim 112\times$ chance)
Item-level precision@5 (median)	0.050	**0.520**
WARMSTART shuffle-gap probe	0.008 PPL	see Table 17

Table 17: WildChat-1M PPL shuffle-gap probe (n=1000 English users grouped by hashed IP, Qwen3-8B, k=5). The paired CIs are over users on the additive log-CE scale.

Condition	Mean log CE	Mean PPL	Paired log-CE gap
Zero-shot	2.8588	17.44	–
Real support	2.7399	15.49	–
Shuffled, seed 17	2.7373	15.45	-0.0026 $[-0.0044, -0.0008]$
Shuffled, seed 42	2.7373	15.45	-0.0026 $[-0.0044, -0.0008]$
Shuffled, seed 73	2.7374	15.45	-0.0025 $[-0.0043, -0.0006]$
Shuffled, seed 99	2.7382	15.46	-0.0018 $[-0.0036, 0.0000]$

Mechanism faithfulness vs. independent semantic measurement. Our discriminability metric uses the same MiniLM-L6-v2 encoder that WARMSTART reads support through. This is intentional: we ask whether supports are distinguishable in the representation space available to the adaptation pipeline, which is the relevant quantity for testing whether WARMSTART could in principle distinguish them. It is not, however, an independent semantic measurement of user distinctiveness, and we do not present it as one. Two partially-independent corroborations are reported elsewhere in the paper: (i) the BM25-matched topic-controlled derangement (Section 3), which uses a lexical (non-MiniLM) similarity score and shows the topic-controlled WARMSTART gap remains 0.010 PPL while ICL retains a 3.62 PPL gap; and (ii) the authorship-attribution probe (Appendix A.23) at 20 and 100 candidates, which scores candidate posts under the adapted model rather than under MiniLM. A non-MiniLM lexical or classifier-based discriminability metric could be added as further robustness; we leave this as a refinement that does not change the present conclusions.

A.20 Heterogeneous-corpus check: WildChat-1M discriminability and shuffle gap

To stress-test the discriminability conclusion on a genuinely more heterogeneous corpus, we re-run the same per-user discriminability diagnostic on WildChat-1M [Zhong et al., 2024] (n=200 English users grouped by hashed IP, k=15 user-turns per user as the discriminability pool, 3,000 pooled items total). Table 16 compares the two corpora.

The contrast is sharp. On WildChat each user's turns are over $5\times$ more self-similar relative to other users (discriminability ratio 2.65 vs. 0.53), and item-level precision@5 rises from $\sim 10\%$ on Blog to $\sim 52\%$ on WildChat. By any standard measure WildChat is the more heterogeneous, more discriminable corpus.

Despite the higher discriminability, the WARMSTART shuffle-gap probe on WildChat shows no matched-support advantage across four derangement seeds: shuffled support is slightly better or statistically indistinguishable from real support. We interpret the negative sign conservatively as another collapse signal rather than as evidence that mismatched users help; the key point is that the matched user's support does not improve the adapted model. The interchangeable-supports interpretation of a small shuffle gap is therefore not supported as the primary explanation in either evaluated regime: where supports are weakly discriminable (Blog), the gap is small; where supports are strongly discriminable (WildChat), the gap is also small. The diagnostic measures model-level support-insensitivity, conditional on supports being non-trivially distinguishable–a condition both corpora satisfy.

Table 18: Shuffle gap versus support size k for the gap-forced β=10 model (Blog, Qwen3-8B, seed 0, n=465). Bootstrap 95% CIs (2000 resamples) on the mean gap in log-CE units.

k	Gap (PPL)	Gap (log-CE)	95% CI (log-CE)
1	7.56	0.255	$[0.165, 0.360]$
2	7.98	0.285	$[0.204, 0.372]$
4	6.09	0.227	$[0.161, 0.298]$
8	5.62	0.214	$[0.156, 0.277]$
16	5.02	0.194	$[0.137, 0.257]$

Table 19: Gap–regression tradeoff across tested variants (Blog, Qwen3-8B, k=5). Regression is computed against the same five-seed collapsed WARMSTART mean from Table 1 (21.14 PPL). Rows marked seed 0 are exploratory single-seed variants; β=10 reports the 5-seed mean.

Variant	Real PPL	Gap (PPL)	Regression
Collapsed WARMSTART (5-seed)	21.14	0.008	0%
β=10 (5-seed)	23.00	7.10	8.8%
Sched 12→5 (seed 0)	23.48	16.61	11.1%
β=12 (seed 0)	24.12	9.36	14.1%
Sched 15→5 (seed 0)	24.40	12.52	15.4%
Sched 15→3 (seed 0)	27.56	73.77	30.4%
InfoNCE (seed 0)	≈46	27.36	≈118%

A.21 Shuffle gap versus support size k (gap-forced model)

Table 18 reports the shuffle gap of the gap-forced (β=10) model at varying support sizes. The gap is significantly nonzero at all k (all bootstrap CIs exclude zero) and slightly decreases from k=1 (7.56 PPL) to k=16 (5.02 PPL). The decrease is consistent with the law of large numbers: at larger k, both matched and shuffled support means converge toward the population embedding centroid, narrowing the distinguishable signal available to the hypernetwork. The monotonic decrease also suggests that the gap-forced model's user sensitivity is strongest at low k, where the support set is most idiosyncratic.

A.22 Gap–regression Pareto frontier

Figure 5 visualizes the same tradeoff in the main text. The point is not a strict ordering over all variants; rather, no tested intervention opens a large shuffle gap while preserving the collapsed model's real-support PPL, and the InfoNCE point lies well off the efficient frontier.

A.23 Authorship attribution

As an additional diagnostic of whether the adapted model encodes user-specific signal, we evaluate authorship attribution: for each meta-test user u, the model adapted with u's support scores one oracle probe post from u against 19 distractor posts from other users, and we report rank-1 accuracy and mean reciprocal rank (MRR). Results are in Table 20.

Performance is indistinguishable from random chance at every k, providing independent confirmation of the personalization-collapse finding in Section 3: the adapted model's per-user query loss depends on the *number* of support items but not on *whose* support items they are.

100-candidate stress test. As a stronger control we re-run the audit with 100 distractor candidates at k=5, lowering chance to 1.00%. Rank-1 accuracy is 0.86% (95% bootstrap CI $[0.22\%, 1.72\%]$, n=465) and MRR is 0.0519 (CI $[0.0424, 0.0631]$); the random-rank MRR for 100 candidates is also ≈0.0519. Both metrics are indistinguishable from chance, strengthening the claim that the collapsed adapter does not preserve recoverable authorship identity under this scoring protocol.

Table 20: Authorship attribution (Blog, Qwen3-8B, seed 0, 20 candidates per query). Random chance: accuracy $= 5\%$, MRR ≈ 0.18.

k	Rank-1 Accuracy	MRR
0	4.1%	0.172
1	4.3%	0.173
2	5.6%	0.188
5	4.1%	0.169
10	5.6%	0.183
15	3.9%	0.168
20	4.5%	0.173

Table 21: Persona-null counterfactuals at k=5 (Blog meta-test, n=465, Qwen3-8B, seed 0). Mean per-user PPL under five conditions; "Δ vs real" is the PPL increase relative to using the user's actual support. The "gap-forced" row uses the β=10 checkpoint.

Condition	Real	Fixed support	Mean persona	Zero persona	Random persona
Collapsed (ERM)	21.83	21.80	21.83	24.78	24.82
Δ vs real	–	-0.02	0.00	$+2.95$	$+2.99$
Gap-forced (β=10)	23.63	41.85	24.55	25.56	25.65
Δ vs real	–	$+18.22$	$+0.92$	$+1.93$	$+2.02$

A.24 Persona-null counterfactuals

To complement the user-vs-user shuffle gap, we test the adapted model's response to four *null* replacements of the support information: *(i) fixed support*: the same support set is used for every user (a single random user's S^k); *(ii) mean persona*: the persona embedding e_u is replaced by the mean over meta-train users $\bar{e} = \mathbb{E}_u[E_\varphi(S_u^k)]$; *(iii) zero persona*: e_u is replaced by the zero vector; *(iv) random persona*: e_u is replaced by an isotropic Gaussian sample matched to the empirical $\bar{e}$ moments. A model that is genuinely user-conditioned should incur a meaningful PPL increase under (i)–(iv); a collapsed model should be approximately invariant.

The collapsed model is essentially invariant to (i) replacing u's support with a single shared support set ($\Delta = -0.02$ PPL) and to (ii) replacing e_u with the mean persona ($\Delta = 0.00$ PPL); the only nulls that hurt are (iii) zeroing the persona and (iv) replacing it with random noise off the persona manifold ($\Delta \approx +2.97$ PPL). The almost-zero Δ on fixed-support and mean-persona is the strongest single indicator that the collapsed adapter ignores user-specific signal: the adapter performs equivalently with *any* on-manifold persona embedding, including the population mean. The gap-forced model behaves qualitatively differently. (i) Wrong fixed support is now catastrophic ($+18.22$ PPL), confirming that the gap-forcing intervention couples the adapter to the matched user's support identity rather than to a global style. (ii) Even the mean persona costs a non-trivial $+0.92$ PPL, and (iii)/(iv) zeroing or randomizing the persona costs $+1.9$–2.0 PPL. The asymmetry between fixed support ($+18.22$) and the embedding-space nulls ($+0.9$–$+2.0$) suggests that the gap-forced model is sensitive to an on-manifold support-derived representation; arbitrary global, zero, or random persona substitutions do not reproduce the same behavior. This is consistent with the layer-wise attribution result (Appendix A.26) that places the user-sensitive signal in late decoder blocks.

A.25 Persona and adapter similarity

The persona-null ablation (Appendix A.24) shows that the gap-forced model is sensitive to which support it sees, but that ablation is conducted at the loss level. Here we open the adapter and ask how user-distinguishable the intermediate representations are. For each meta-test user u we compute the persona embedding $e_u = E_\varphi(S_u^k)$ and the effective rank-8 LoRA update $\Delta W_u = B_u A_u$ produced by the hypernetwork (concatenated across all 36 layers), then for two independent draws of S_u^k for the same user we compute the same-user cosine, and for pairs of distinct users in a meta-batch we compute the diff-user cosine. Reporting both mean cosines and their difference isolates how much of the geometry depends on *whose* support the operator received, in contrast to the constant background of structure shared by all users.

Table 22: Persona embedding and effective LoRA-update ($\Delta W = BA$) cosine similarities (Blog meta-test, Qwen3-8B, seed 0). "Same" is two draws of S_u^k for the same user; "diff" is across-user pairs from the same meta-batch. The same−diff column is the only support-dependent signal.

	Persona embedding			Effective LoRA update ΔW		
	Same	Diff	Same−Diff	Same	Diff	Same−Diff
Collapsed (ERM WARMSTART*)*						
k=1	0.9715	0.9669	0.0046	0.99993	0.99991	1.4×10^{-5}
k=5	0.9941	0.9900	0.0041	0.99999	0.99997	1.4×10^{-5}
k=10	0.9970	0.9928	0.0042	0.99999	0.99998	1.5×10^{-5}
Gap-forced (β=10)						
k=1	0.1908	0.0117	0.1791	0.9962	0.9933	2.9×10^{-3}
k=5	0.4752	0.0673	0.4079	0.9993	0.9974	1.9×10^{-3}
k=10	0.6355	0.0861	0.5493	0.9997	0.9977	1.9×10^{-3}

The two checkpoints differ qualitatively in persona space and only quantitatively in adapter space. Under collapsed ERM, the persona embedding is a near-rigid direction: same- and diff-user cosines are both ≈ 0.97–0.99 with a same−diff gap of 0.004 across all k, indicating per-user variation that is essentially noise around a single population mean. Under gap-forcing, the persona embedding becomes strongly user-specific: same-user cosine rises from 0.19 at k=1 to 0.64 at k=10 while diff-user cosine stays near zero, opening a same−diff gap of 0.18 at k=1 and 0.55 at k=10. The persona encoder is visibly and increasingly user-conditioned with growing k. The effective rank-8 LoRA update $\Delta W = BA$ tells a more restrained story: same- and diff-user cosines remain ≥ 0.99 for both checkpoints; gap-forcing increases the same−diff gap from 1.4×10^{-5} to 1.9×10^{-3}, a two-order-of-magnitude rise, but the absolute gap is still small relative to the shared ≈ 0.997 background. Most of the trained adapter direction is therefore shared globally, and gap-forcing carves out only a small user-specific component within an otherwise common subspace. This nuance matters when read alongside the persona-null ablation (Appendix A.24), where substituting wrong fixed support raises gap-forced PPL by $+18.22$ at k=5: the behavioral sensitivity is large even though the cosine geometry of the adapter remains highly collinear, suggesting that gap-forcing operates within a subspace where small angular separations correspond to substantial loss differences (consistent with the late-layer concentration in Appendix A.26).

A.26 Layer-wise attribution of the shuffle gap

To localize which transformer layers carry the user-sensitive signal in the gap-forced (β=10, seed 0) model, we zero out the LoRA adapter at one layer at a time and re-measure the shuffle gap at k=5 across all 465 meta-test users. Figure 7 reports the residual gap when each layer group is individually ablated, relative to the all-layers baseline gap of 5.76 PPL.

The user-sensitive signal is concentrated in the final $\sim$10 layers of the 36-layer backbone, with layers 27 and 32–35 individually responsible for the largest gap reductions. Early-layer LoRA (0–10) contributes negligibly; layer 0 slightly *increases* the gap when zeroed, consistent with a mild population-level adaptation that slightly competes with user-specific conditioning in later layers. The concentration of persona signal in late decoder blocks is consistent with user-sensitive conditioning being implemented as a higher-level modulation rather than a surface-level lexical shift.

A.27 Per-user shuffle-gap distribution

The aggregate shuffle gap (Section 3) summarizes collapse as a single scalar; Figure 4 shows the full per-user distribution of $\Delta_{\text{shuf}}(u) = \text{PPL}_{\text{shuf}}(u) - \text{PPL}_{\text{real}}(u)$ at k=5 for all 465 meta-test users. Under ERM (left panel), the distribution is sharply concentrated near zero: 91.8% of users fall below the 0.01 log-CE threshold, confirming that collapse is not driven by a few outliers but is a population-wide phenomenon. Under the gap-forced β=10 model (right panel), the distribution shifts rightward and develops a heavy right tail: only 17.6% of users remain below 0.01, and the median gap rises from $\sim10^{-3}$ to ~0.15 log CE. The tail users with the largest gaps tend to have more distinctive writing styles, consistent with the intuition that users far from the population prior benefit most from genuine personalization.

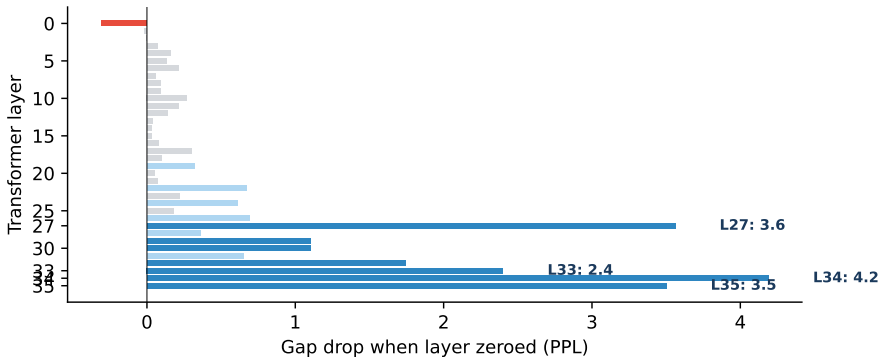


Figure 7: Layer-wise shuffle-gap attribution (gap-forced β=10 model, Qwen3-8B, seed 0, k=5). Each bar shows the gap reduction when that layer's LoRA is zeroed (baseline gap 5.76 PPL). Layers 27 and 32–35 carry the majority of the user-sensitive signal; early and middle layers contribute negligibly.

Table 23: PersonaFeedback hard tier (101 users, 1010 triplets, Qwen3-8B). Preference accuracy and mean log-odds with bootstrap 95% CIs (2000 resamples).

Condition	Pref. Accuracy	Mean Log-Odds
Collapsed	57.9% [54.9, 61.0]	0.106 [0.078, 0.134]
Collapsed shuffled, seed 17	57.8% [54.8, 60.9]	0.106 [0.078, 0.135]
Collapsed shuffled, seed 42	58.2% [55.1, 61.2]	0.107 [0.079, 0.135]
Gap-forced (β=10)	55.7% [52.6, 58.8]	0.096 [0.069, 0.124]
Gap-forced shuffled, seed 42	56.5% [53.4, 59.6]	0.096 [0.069, 0.124]

A.28 PersonaFeedback preference evaluation

As a preference-based complement to the PPL-based shuffle gap, we evaluate WARMSTART on the PersonaFeedback hard tier [Tao et al., 2025]: 1010 pairwise preference triplets (x, y_+, y_-) from 101 users, where y_+ is the user-preferred response and y_- is the rejected response. Each user's persona profile (a textual description of demographics, personality, and preferences) serves as the support input to WARMSTART. Table 23 reports preference accuracy $\text{Acc} = \mathbf{1}\{\log p_\theta(y_+ \mid x) > \log p_\theta(y_- \mid x)\}$ and mean pairwise log-odds.

Both models exceed chance (50%), indicating that the persona text provides some preference-relevant signal. However, the collapsed and gap-forced models achieve similar accuracy (overlapping CIs), and repeated preference-based shuffle gaps are near zero: collapsed seed-17 gap $+0.1\%$, collapsed seed-42 gap -0.3%, and gap-forced seed-42 gap -0.8%. This null result is consistent with two factors: (i) PersonaFeedback uses textual persona descriptions rather than the post-history supports that WARMSTART was trained on, creating a distribution shift; (ii) the Blog-trained gap-forcing run may not transfer to preference-based conditioning without domain-specific fine-tuning. The result underscores that breaking PPL-based collapse (Section 3) does not automatically yield preference-aligned personalization, and that preference-based benchmarks provide a complementary and more demanding evaluation axis.

A.29 Generation quality (diagnostic)

As a complementary diagnostic to the PPL-based shuffle gap, we sample short continuations (≤ 50 tokens, k=5, derangement seed 17) from the collapsed ERM and gap-forced (β=10) checkpoints on 300 meta-test users, and compare three reference-free style proxies for the matched-support and shuffled-support generations: MiniLM cosine similarity to a held-out user reference, type-token ratio (TTR), and Self-BLEU-2. Results are in Table 24.

The two checkpoints behave differently on the embedding-space proxy. The collapsed ERM model has only a tiny positive MiniLM cosine gap ($+0.006$), consistent with little detectable preference for

Table 24: Generation-quality probe (n=300 meta-test users, k=5, derangement seed 17, max 50 generated tokens). Reference-free style proxies. "Gap" is matched−shuffled.

Metric	Real	Shuffled	Gap
Collapsed (ERM WARMSTART*, seed 0)*			
MiniLM cosine	0.315	0.309	+0.006
TTR	0.815	0.815	−0.000
Self-BLEU-2		0.053	
Gap-forced (β=10, seed 0)			
MiniLM cosine	0.315	0.283	+0.032
TTR	0.809	0.827	−0.018
Self-BLEU-2		0.057	

matched-user style. The gap-forced model shows a larger positive MiniLM cosine gap of $+0.032$: matched-support generations are closer in MiniLM space to the held-out reference than shuffled-support generations, indicating that the gap-forcing intervention couples the adapter to the support's semantic neighborhood. TTR, by contrast, is unchanged under ERM and moves in the opposite direction under gap-forcing (matched generations have lower TTR than shuffled), while Self-BLEU-2 remains very low in both settings. We interpret the table as evidence that the gap-forcing intervention transfers weakly to the embedding-space style proxy but not to surface lexical statistics; we do not interpret it as a settled claim about generation-quality personalization, since we have no ground-truth user-style metric.

Contextual material. We end with background comparisons and broader-impact considerations that are useful for reviewers but are not part of the core empirical argument.

A.30 Extended related work

Federated learning with training-time known clients. Methods such as pFedMe [Dinh et al., 2020] and APFL [Deng et al., 2020] formalize personalization as a bi-level optimization problem in which each client solves a local objective that includes a regularizer pulling their parameters toward a global consensus model. These approaches operate in a training-time known-client regime: the population of clients is fixed, per-client data is available throughout the federated optimization loop, and personalization is tuned to clients that are known to the optimizer at training time. Our setting is different: WARMSTART performs personalization at deployment time for users that are *unseen* during meta-training, using only a small support set S_u^k provided at inference. The two regimes are complementary rather than directly comparable, and the consensus gap Δ_{Cons} that our framework names as a population-level invariant is orthogonal to the federated objective.

Meta-learning and PAC-Bayes. Gradient-based meta-learning, exemplified by model-agnostic meta-learning (MAML) [Finn et al., 2017], optimizes a shared initialization so that a few gradient steps on a new task lead to rapid convergence. Baxter's foundational framework [Baxter, 2000] established the bi-level structure in which a meta-objective aggregates task-specific losses, paving the way for modern PAC-Bayesian analyses. PAC-Bayes theory provides generalization guarantees for learning a distribution over hypotheses rather than a single hypothesis; Amit and Meir [2018] extends PAC-Bayes to the meta-learning regime, bounding the task-average loss of a meta-posterior in terms of the KL divergence from a meta-prior, building on Pentina and Lampert [2014]'s PAC-Bayesian lifelong-learning bound and McAllester [1999]'s foundational PAC-Bayes inequality. Theorem 1 specializes this template to the adapted-personalization case in which the bounded query loss is evaluated after applying $\mathcal{A}$ to a support-set-dependent perturbation of the base model. The theorem establishes the standard $1/\sqrt{m}$ dependence on the number of meta-training users; deriving an explicit support-size dependence would require additional structural assumptions beyond A1 and A3 and is left to future work.

LLM personalization methods. The dominant algorithmic approach is per-user fine-tuning via parameter-efficient methods, most commonly LoRA [Hu et al., 2021]. Non-parametric alternatives such as retrieval-augmented generation and in-context learning condition the backbone on user history without updating weights; the LaMP benchmark [Salemi et al., 2024] has become a standard evaluation protocol for the retrieval-based approach. Hypernetwork-based adapters generate user-

specific weights in a single forward pass. Profile-to-PEFT [Tan et al., 2025] is the closest architectural template to WARMSTART; we treat P2P as a hypernetwork baseline run with its authors' released pipeline. In-Place Test-Time Training [Feng et al., 2026] is persona-agnostic and performs no learning across users, making it a complementary baseline rather than a competitor.

Preference learning and reward models. Personalized Soups [Jang et al., 2023] interpolates pre-trained and individually fine-tuned checkpoints. Meta Reward Modeling [Cai et al., 2026] applies MAML to reward models together with a Robust Personalization Objective (RPO) that reweights hard users during meta-training; MRM's RPO is presented as an empirical heuristic, and our variance-aware CVaR penalty plays the analogous role in a next-token-prediction setting under A1. Complementary approaches such as Group Robust Preference Optimization [Ramesh et al., 2024] and PAL [Chen et al., 2024] interpolate or reweight among parameter sets trained on different objectives but do not formalize the consensus gap as an irreducible population-level invariant.

DRO and fairness. Group DRO [Sagawa et al., 2020] targets worst-group training loss over predefined groups, whereas our CVaR_α penalty operates over the meta-objective for per-user test-time adaptation; individual fairness [Dwork et al., 2012] imposes a global Lipschitz constraint that is in tension with aggressive personalization, and the consensus gap provides a principled measure of when that tension is irreducible.

A.31 Broader impact and privacy

Deployment of WARMSTART-style adaptation should be paired with user-visible controls over stored data, explicit opt-out mechanisms, and awareness that the persona embedding e_u may encode protected attributes or enable membership inference; a formal privacy analysis is left to future work. Concretely, our linear-probe analysis shows that the persona embedding e_u reaches 69.0% gender-prediction accuracy at $k{=}20$ and near-zero R^2 on age. While these values are below what a purpose-built classifier would achieve, they demonstrate that the embedding retains demographic signal that could enable membership inference or group-level profiling. Any deployment of persona-embedding-based personalization should include a formal privacy audit – for instance, differential privacy guarantees on the embedding or post-hoc scrubbing of protected-attribute signal – before serving users.