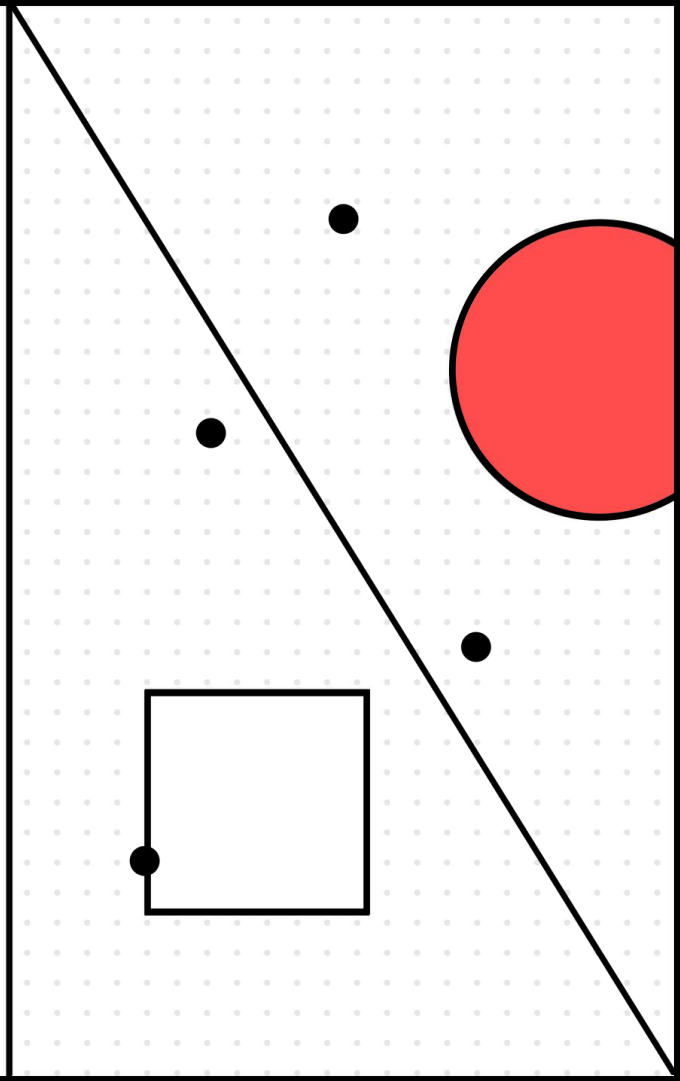
CARD

Cluster-level Adaptation with Reward-guided guided Decoding for Personalized Text Generation

Yutong Song[1], Jiang Wu[2], Weijia Zhang[3], Chengze Shen[4], Shaofan Yuan[4], Weitao Lu[4], Jian Wang[4], Amir Rahmani[1], Nikil Dutt[1], Yu Wang[4]

[1]UC Irvine · [2]Independent Researcher · [3]University of Amsterdam · [4]TikTok

arXiv:2601.06352 · January 2026



Personalization of LLMs Is Critical Yet Challenging

WHY IT MATTERS

Powers dialogue systems, content recommendation, and advertising

Users expect outputs that reflect their own vocabulary, tone, vocabulary, tone, and style

Generic LLM outputs fail to capture individual nuances nuances

CURRENT PARADIGMS & LIMITS

Paradigm	Key Limitation
RAG Retrieval-Augmented Generation	Shallow personalization; sensitive to retrieval quality
PEFT Parameter-Efficient Fine-Tuning	Scales poorly; expensive as user base grows

CORE TENSION

Fine-grained personalization vs. Scalable deployment — existing methods cannot achieve both simultaneously. simultaneously.

Three Fundamental Challenges Drive the Need for CARD

Granularity Trade-off in PEFT

Shared adapters dilute individual nuances, while per-
while per-user adapters incur prohibitive scaling costs
scaling costs and become unstable under sparse user
sparse user histories.

Scarcity of High-Quality Preference Preference Data

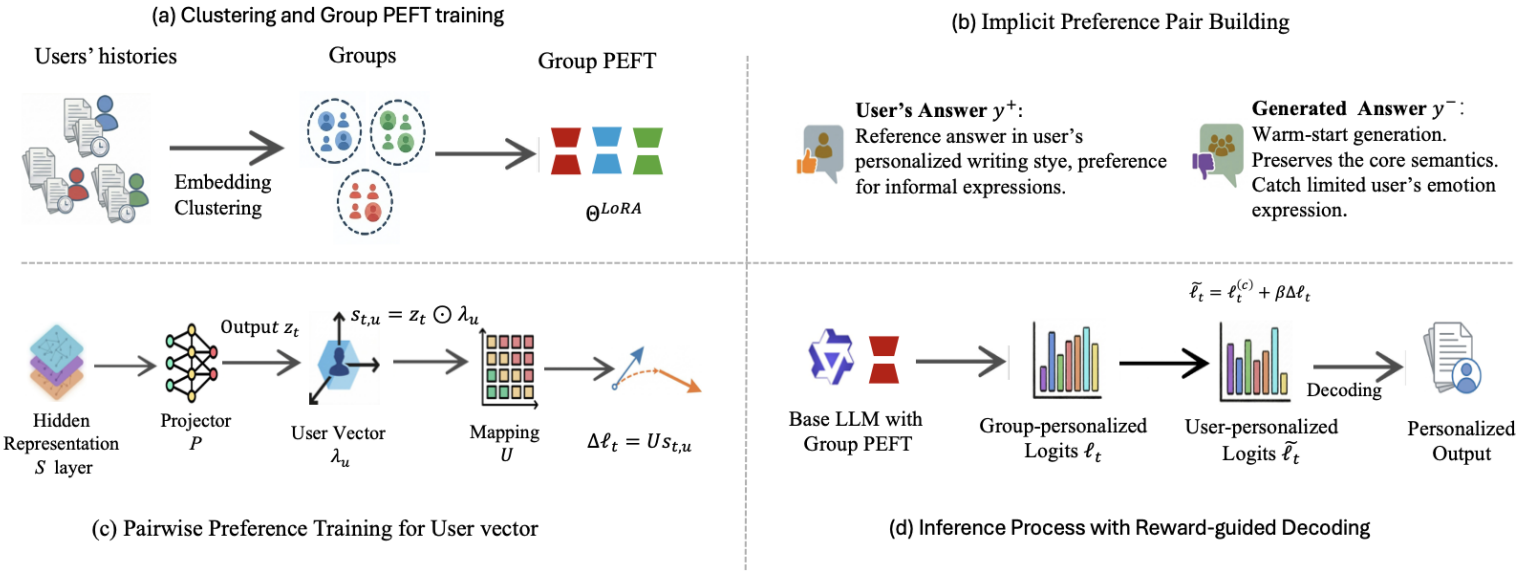
Constructing preference pairs at scale is difficult.
Heuristic constructions often entangle topical content
with stylistic traits, yielding unreliable training signals.

Cold-Start Users

New users with limited interaction history cannot be
effectively personalized by existing PEFT methods, which
require substantial per-user data to converge.

KEY QUESTION

Can we leverage group-level generalization for efficiency while capturing fine-grained individual preferences?


(a) Clustering and Group PEFT training
Users' histories
Groups
Group PEFT
Embedding
Clustering
Θ^{LoRA}
(b) Implicit Preference Pair Building
User's Answer y^+:
Reference answer in user's personalized writing stye, preference for informal expressions.
Generated Answer y^-:
Warm-start generation.
Preserves the core semantics.
Catch limited user's emotion expression.
Hidden Representation S layer
Projector P
Output z_t
$s_{t,u} = z_t \odot \lambda_u$
User Vector λ_u
Mapping U
$\Delta\ell_t = Us_{t,u}$
(c) Pairwise Preference Training for User vector
$\tilde{\ell}_t = \ell_t^{(c)} + \beta\Delta\ell_t$
Base LLM with Group PEFT
Group-personalized Logits ℓ_t
User-personalized Logits $\tilde{\ell}_t$
Decoding
Personalized Output
(d) Inference Process with Reward-guided Decoding

CARD: A Hierarchical Coarse-to-Fine Framework

Core Insight: Personalization signals are hierarchical — broad preferences are shared as group-level priors, while fine-grained nuances manifest as stable individual differences.

STAGE 01

Group-Level Adaptation

Cluster users by stylistic similarity and train one **cluster-specific LoRA adapter** per group. Amortizes adaptation costs and provides robust cold-start priors.

STAGE 02

Implicit Preference Learning

Construct preference pairs by contrasting **user-authored text** (preferred) against **cluster-level generations** (dispreferred). Reduces semantic confounding.

STAGE 03

Decoding-Time Personalization

Learn a lightweight **user preference vector** (λ_u). Inject (λ_u). Inject personalization via reward-guided logit logit editing while keeping the backbone frozen. frozen.

KEY DESIGN PRINCIPLE
PRINCIPLE

User-level personalization operates exclusively at decoding time, enabling rapid user switching with minimal per-user storage.

Technical Deep Dive: How CARD Works

CLUSTERING & LORA

01

User Encoding

Encode user history with frozen sentence encoder → embedding e_u.

K-Means Clustering

Partition users into K clusters $\{C_1, C_2, \ldots, C_K\}$ based on stylistic similarity.

Cluster-Specific LoRA

Train one adapter per cluster (rank r=16) via supervised fine-fine-tuning.

PREFERENCE PAIRS

02

Contrastive Construction

For interaction (x, y):

- **Positive (y^+):** User's ground truth.
- **Negative (y^-):** Cluster-LoRA output.

Semantic Alignment

Identical prompts eliminate semantic confounding, isolating stylistic deviation.

REWARD-GUIDED DECODING

03

Personalization Head

Ψ = (P, U) projects hidden states to preference space.

User Vector Modulation

Lightweight vector λ_u modulates representation:

$s_{t,u} = z_t \odot \lambda_u$

Logit Correction

Low-rank adjustment applied to Top-k candidates: candidates:

Ȼt Ě □t□□□ Ą úfĺ fĆEÆ

CARD Achieves State-of-the-Art Performance

Evaluation on LaMP and LongLaMP benchmarks (ROUGE-1 Scores)

TASK	NON-PERS.	RAG	PAD	PPLUG	Î Ï Ï Į	CARD
LaMP4 (News Headline)	0.146	0.166	0.158	0.157	0.152	ĆĒĈĐ
LaMP5 (Scholarly Title)	0.425	0.456	0.442	0.464	0.426	**0.459**
I MÌ Ï Đ Ä RÑÑP Ï MØMŐOØMŒÑÅ	0.497	0.500	0.502	0.511	0.498	**0.521**
LongLaMP1 (Abstract Gen.)	ĆĒĊĈ	0.372	0.355	0.391	0.382	**0.411**
LongLaMP3 (Review Writing)	0.292	ĆĒĊĐČ	0.308	0.396	0.295	**0.405**

CARD ranks 1st in 10 out of 12 evaluation settings. The remaining two settings are near-best (within 0.005 ROUGE-1).

Ablation Study Reveals the Decisive Role of User Vectors

Configuration	[illegible]	LaMP7 R-1
w/o LoRA (User Vector only)	0.207	0.507
w/o User Vector (LoRA only)	0.148	0.498
[illegible]	**0.218**	**0.521**

Impact of Removing User Vector (LaMP4 Drop)

~96% of improvement lost

Group LoRA is Insufficient

Using group-level LoRA alone yields only marginal improvement over non-personalized baselines (+0.002 ROUGE-1). It captures shared patterns but misses individual nuances.

User Vector is the Driver

[illegible]

Semantic & Stylistic Balance

CARD effectively fuses group semantics with individual preferences (e.g., informal wording, emojis), achieving a balance that neither component can reach alone.

Case Study: Balancing Semantics and Style

Task: LaMP-7 (Tweet Paraphrasing)

USER HISTORY

ÀÓÞ ŒP ŊŎPÖ Ř ŃŎŅŅÑÃŒŎ OMŐŐŘ ŎŎŊ À

ÀPOÑ Ö ŎQŎÑ R MŒŎÒPÑØMÕÕŘ MÖMŖÒÒŊÆŽÖ ŃŹŘÒŊÀ

ÀŃMŌÃP R MÞ ŅŐØPOÑ R ÑÑÔÑŌÑ AAA À

GROUP LORA ONLY

HMÖ QÑŹŘ ŐÕÑMŒÑ PŎ OMQÑ ŹÑŃÑQÑ Ö Ř ŃŎŅŅÑ PŎÑMŘB

İ OÑ ŅÕÕ R MŒÑŔŃÑÕÑŌP MÕÑ ØÞOÑ Ö ŎQÕŊB

HMÖ ÕÕÔÕÕŊ ŅŐŹŘ MŹÑ PŎ POÑ ÞŐŃŎÖ ÕŊ R ÑÑÔÑŌÑB

FULL CARD

ŊŎPÖ Ř ŃŎŅŅÑ MÕÑ ÕÃ ŒŎ OMŐŐŘ ŎŎŊ

POMP Ö ŎQÕÑ R MŒÖ ŎQÕŊ MÕÑ ÕÒPÑØMÕÕŘ MÖMŖÒÕŊ AAA

ŃMŌÃP RMÒP ŅŐØPOÑ R ÑÑÔÑŌÑ

Insight: Group LoRA preserves core semantics but uses "average" phrasing. **CARD** injects individual preferences (informal wording, emotional cues, emojis) while maintaining semantic stability.

CARD Excels in Low-Resource Settings and Scales Efficiently

LOW-RESOURCE ROBUSTNESS

J ÒPO ŎŌŌŘ D OÒŒPŎØŘ ÒPÑŌŒ

0.216

ĪÎ Į Ġ ĜÆĈ Ï ŃŎØŃ ĂQŒĆĔÇĎ NMŒŃŎŎŃÅ

CARD is highly sample-efficient, peaking at just ~10 history items, while other methods require significantly more data to converge.

ĜĞĞĦ ĦĜÍ F K FÎ Ì ĪĖĪ ĦÎ Í

Metric	RAG	PEFT	FĖĪ G
Training Time / User	Î ĂŠ ŚÅ	O(|Pᵤ|)	Î ĂŠD ŚÅ (Offline Only)
Latency / Query	O(|Pᵤ|)	Î Ă ŎMŇ Ą Ì ŃØJŇÅ	O(k·J) ĂJŇØŘ I ŎRÅ
Storage / User	Î ĂŠ ŚŇ Å	O(r·H·L)	O(D) (Vector Only)

FĖĪ G ØŇØÞØŇŒŎŌŘ MĈČĐÆŇŎ ŐØŅŅØŇŌŃŇ QŇPŎØŐŇØÞŒŇØĘ MŃŌNŎŌŇ Ä F ĐŒPŇØ LoRA are shared.

Ç Þ ÖMŌ MŌŇ I I Ì ÎÞŇŊÑŒ F ŎŌŅÒØÖ Ï ÑØŒŎŌMÕÒŖMPÒŎŌ Ï ÞMÕÒPŘ

Why Beyond ROUGE? Surface-level lexical overlap metrics fail to capture subjective qualities like style, tone, and expressiveness.

I I Ì Ĥ Ç GĠĜĂĠĨ İ ÆĐĿČÅ

CARD consistently receives the **highest or near-highest scores** across all LaMP tasks. The judgments are stable and align with the quantitative ROUGE improvements.

HUMAN EVALUATION

Human evaluators exhibit stronger discrimination. Notably, **CARD sometimes scores higher than reference outputs**. Humans prefer CARD's personalized generations for their stylistic alignment and expressiveness.

GĜĨ HĠÍ FÇĨ HFĜĨ İ Ç Ė İ Ì Ė İ İ GĨ

USER VECTOR DIMENSION

ĤĚ ĈČĐ

Peak performance; larger vectors introduce overfitting.

Ç HGGĜÍ I Ė K GĨ GĜĨ İ Ç

Last 4 Layers

Aggregation outperforms single-layer or 8-layer approaches.

CLUSTER SIZE

Ī ŎNÞŒP Ħ

Stable performance across a wide range of cluster counts. counts.

[illegible]

01 [illegible]

Combines cluster-level LoRA (shared patterns) with decoding-time user vectors (individual deviations) for deviations) for superior quality-efficiency trade-off.

02 IMPLICIT PREFERENCE LEARNING

[illegible]

03 [illegible]

Personalization operates exclusively at decoding time with frozen LLM, enabling rapid user switching and minimal storage.

[illegible] STRONG EMPIRICAL RESULTS

State-of-the-art performance on LaMP and LongLaMP benchmarks, with robust low-resource generalization.

Limitations

- K-means clustering may not capture complex latent user relationships.
- Single preference vector may limit expressiveness for diverse behaviors.

[illegible]

- Modeling temporal preference drift over time.
- [illegible]
- [illegible]

CODE AVAILABLE [illegible]