LSRP: A Leader–Subordinate Retrieval Framework for Privacy-Preserving Cloud–Device Collaboration

Yingyi Zhang
Dalian University of Technology
City University of Hong Kong
Dalian, China
yingyizhang@mail.dlut.edu.cn

Xianneng Li*
Dalian University of Technology
Dalian, China
xianneng@dlut.edu.cn

Abstract

Cloud-device collaboration leverages on-cloud Large Language Models (LLMs) for handling public user queries and on-device Small Language Models (SLMs) for processing private user data, collectively forming a powerful and privacy-preserving solution. However, existing approaches often fail to fully leverage the scalable problem-solving capabilities of on-cloud LLMs while underutilizing the advantage of on-device SLMs in accessing and processing personalized data. In this paper, we propose a **L**eader–**S**ubordinate **R**etrieval framework for **P**rivacy-preserving cloud–device collaboration (LSRP), a novel solution that bridges these gaps by 1) enhancing on-cloud LLM guidance to on-device SLM through a dynamic selection of task-specific leader strategies and 2) integrating the data advantages of on-device SLMs through small model feedback for aligning the on-cloud LLM. Experiments on two datasets demonstrate that LSRP outperforms state-of-the-art baselines, improving question-answer relevance and personalization, while preserving user privacy through efficient on-device retrieval.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Security and privacy** → **Privacy protections**; • **Human-centered computing** → **Personal digital assistants**.

Keywords

Cloud-Device Framework; Large Language Model; Privacy-Preserving;

ACM Reference Format:
Yingyi Zhang and Xianneng Li. 2025. LSRP: A Leader–Subordinate Retrieval Framework for Privacy-Preserving Cloud–Device Collaboration. In *Proceedings of Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding authors.

KDD '25, Toronto, ON, Canada

ACM ISBN 978-1-4503-XXXX-X/2025/08
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advancements in Large Language Models (LLMs) have enabled enterprises to collect extensive user data through interactive Q&A sessions to enhance personalization [1, 7]. However, growing global privacy regulations—such as the EU's GDPR [4]—impose strict restrictions on data collection, preventing private data from leaving user devices. This presents a major challenge for AI service providers, especially given that on-device deployment of LLMs remains infeasible due to computational limitations.

To address this, leading tech companies (e.g., Apple, Google, Huawei) have developed lightweight on-device Small Language Models (SLMs) [3, 5, 6] to process sensitive data locally. While privacy-preserving, these SLMs have limited capacity for complex reasoning, resulting in degraded user experience. To bridge this gap, cloud-device collaborative frameworks have emerged, where the cloud LLM generates an initial sketch response for refinement by the on-device SLM [8].

However, existing sketch-based frameworks reamians two key issues: **(1) Limited utilization of on-cloud LLMs:** Responses generated from fixed prompts lack task specificity and fail to meet personalized needs. **(2) Inadequate integration of on-device data:** Cloud-side generation is often blind to user-specific context, limiting personalization and overall effectiveness. This raises a central question: *How can cloud-device collaboration be designed to maximize on-device performance while preserving privacy and aligning with user needs?*

To this end, we propose **LSRP**—a **L**eader–**S**ubordinate **R**etrieval framework for **P**rivacy-preserving collaboration. Our method is inspired by path-goal leadership theory [2], treating the cloud LLM as a "leader" that generates structured guidance, and the on-device SLM as a "subordinate" that refines outputs using private data. Our framework consists of two key modules: **(1) U-U-RAG:** To better personalize guidance, we design a *leadership-style-based retrieval mechanism*. **(2) SMFB-DPO:** To improve alignment between leader strategies and user needs, we introduce a *small model feedback loop*.

2 Method Overview

We address the problem of input-output-level privacy-preserving collaboration between on-cloud LLMs and on-device SLMs. Let $\mathcal{T}$ be the user task, $\mathcal{P}$ the private data retained locally, θ the parameters of the cloud LLM, and ϕ the parameters of the device-side SLM. The cloud LLM processes $\mathcal{T}$ and generates intermediate guidance $O_{\text{LLM}} = \mathcal{F}_{\text{LLM}}(\mathcal{T}, \theta)$, which is then refined by the on-device SLM together with $\mathcal{P}$ to produce the final output: $O_{\text{SLM}} = \mathcal{F}_{\text{SLM}}(O_{\text{LLM}}, \mathcal{P}, \phi)$. The objective is two-fold: (1) preserve privacy by ensuring $\mathcal{P}$ is never shared externally, and (2) improve

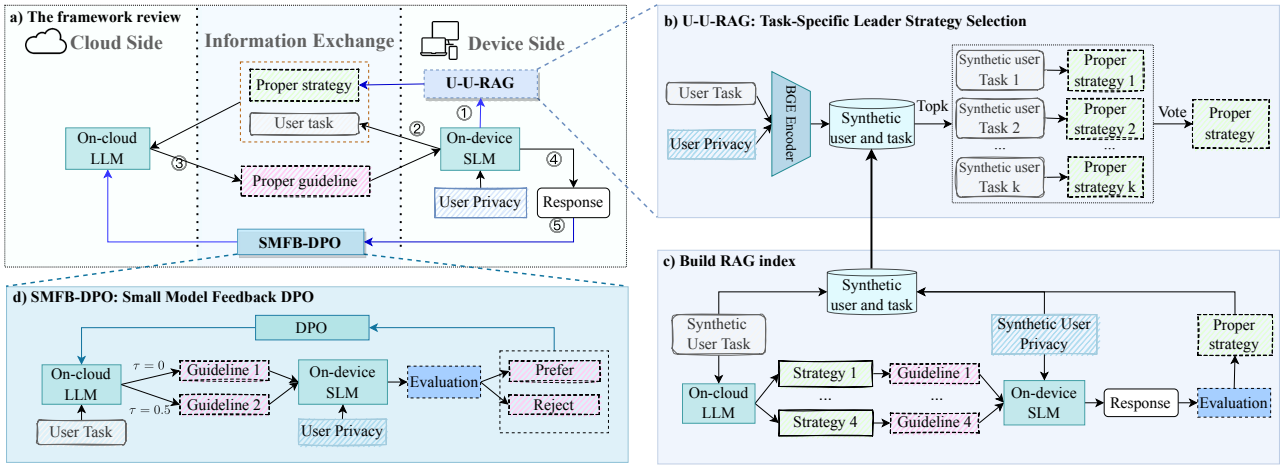

Figure 1: Main framework of LSRP.

task performance and personalization by leveraging cloud-side computation. Thus, the overall optimization problem is:

$$\max_{O_{\text{LLM}}} Q\left(\mathcal{F}_{\text{SLM}}(O_{\text{LLM}}, \mathcal{P}, \phi)\right), \quad \text{where } O_{\text{LLM}} = \mathcal{F}_{\text{LLM}}(\mathcal{T}, \theta). \quad (1)$$

To solve this, we introduce a leader–subordinate retrieval framework (LSRP), as shown in Figure 1, inspired by path-goal leadership theory. In this framework, the cloud LLM (leader) provides task-specific guidance, and the device-side SLM (subordinate) adapts it to local context and private data.

We propose a task-specific leader strategy selection module, **U-U-RAG**, to identify the most suitable leader style for each task. We define four strategies—directive, supportive, participative, and achievement-oriented—each corresponding to a prompt $\mathbb{P}_{\text{LLM}}^{\text{Lead}}$. We first build index using synthetic user-task pairs, the cloud LLM generates candidate guidelines under these strategies, and the on-device SLM evaluates their effectiveness. The best-performing strategies are stored as indexed embeddings. At inference, second, the real user-task pair $(\mathcal{P}, \mathcal{T})$ is embedded and matched against the index to retrieve similar cases. A majority vote over the top-k retrieved strategies yields the final prompt.

To further optimize cloud-side outputs, we propose **SMFB-DPO**. The cloud LLM generates two candidate guidelines per task with varied decoding temperatures. The on-device SLM executes both, evaluates them using $Q(\cdot)$, and identifies the preferred one. Using DPO, the cloud LLM parameters θ are fine-tuned to prefer these responses:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\sum \log P(O_{\text{LLM}}^{p} \mid \mathcal{T}, \theta) + \lambda \sum \log P(O_{\text{LLM}}^{r} \mid \mathcal{T}, \theta), \quad (2)$$

$$\theta^{*} = \arg\min_{\theta} \mathcal{L}_{\text{DPO}}(\theta), \quad (3)$$

where p and r denote preferred and rejected guidelines, and λ balances the two terms.

We define the evaluation metric Q as a weighted combination of Q-A relevance, user data reference rate (UDRR), and perplexity (PPL):

$$Q(O_{\text{SLM}}) = w_1 \cdot \text{Q-A Rel.} + w_2 \cdot \text{UDRR} - w_3 \cdot \text{PPL}. \quad (4)$$

Table 1: This table presents the performance comparison of our method, LSRP, against baseline approaches. "*" indicates the statistically significant improvements.

	LLaMa Series		
CoGen Dataset	**Q-A Rel.↑**	**Persona.↑**	**FRE↑**
On-device SLM	8.9383±4.0649	7.5166±1.5889	36.0624±5.3881
Distillation	8.5541±6.3387	7.5291±1.6525	34.7733±4.0880
CoGenesis	6.1583±2.8082	5.8250±2.4821	29.7497±6.3766 1
LLM guide SLM (ours)	9.0125±2.4790	7.4500±2.6391	34.6195±4.8505
LSRP	**9.2458±1.2950***	**7.5458±1.5213***	**37.6269±1.9864***
On-device LLM†	9.9791±0.0203	8.8166±0.8008	38.7600±25.0150
Movie Explain Dataset	**Q-A Rel.↑**	**Persona↑**	**FRE↑**
On-device SLM	7.6333±6.8822	6.8750±2.2915	54.7389±3.5898
Distillation	8.2666±2.3738	7.4416±1.2130	55.3887±2.1095
CoGenesis	6.2291±4.0099	5.6166±3.2124	54.7784±2.9047
LLM guide SLM (ours)	8.5500±3.7475	7.7500±1.8062	54.3418±2.5286
LSRP	**8.6708±2.5124***	**7.9750±1.4431***	**56.5052±2.2602***
On-device LLM†	9.8958±0.1016	9.2416±0.4999	58.5266±2.1488

Note: On-device LLM† represents deploying the LLM directly on the device. However, this is often infeasible in real.

Weights (w_1, w_2, w_3) are determined by NSGA-II multi-objective optimization, with the knee point on the Pareto front used for final trade-off selection. This enables LSRP to generate personalized, privacy-preserving, and high-quality responses.

3 Experimental Summary

We evaluate our proposed LSRP framework on CoGenisis to evaluate its effectiveness, component contributions, and ability to leverage both cloud and device-side advantages. Experiments are conducted using LLaMa models under real-world privacy constraints.

Results on Table 1) show that LSRP significantly outperforms baselines in Q-A relevance, personalization, and readability. Notably, LSRP achieves a Q-A Rel. of 9.31 on CoGenisis and 8.67 on Movie Explain, validating its adaptability across diverse and single-task settings. Compared to sketch-based or distillation methods, even the basic "LLM Guide SLM" shows strong performance, while full LSRP brings further improvements via strategy selection and feedback optimization.

References

[1] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* 27, 4 (2024), 42.

[2] Robert J House. 1971. A path goal theory of leader effectiveness. *Administrative science quarterly* (1971), 321–339.

[3] Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. OpenELM: An Efficient Language Model Family with Open Training and Inference Framework. *arXiv.org* (April 2024). https://arxiv.org/abs/2404.14619v1

[4] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.

[5] Yehui Tang, Fangcheng Liu, Yunsheng Ni, Yuchuan Tian, Zheyuan Bai, Yi-Qi Hu, Sichao Liu, Shangling Jui, Kai Han, and Yunhe Wang. 2024. Rethinking optimization and architecture for tiny language models. *arXiv preprint arXiv:2402.02791* (2024).

[6] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[7] Bin Wu, Zhengyan Shi, Hossein A Rahmani, Varsha Ramineni, and Emine Yilmaz. 2024. Understanding the Role of User Profile in the Personalization of Large Language Models. *arXiv preprint arXiv:2406.17803* (2024).

[8] Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. 2024. Cogenesis: A framework collaborating large and small language models for secure context-aware instruction following. *arXiv preprint arXiv:2403.03129* (2024).