

RAG-based User Profiling for Precision Planning in Mixed-Precision Over-the-Air Federated Learning

Jinsheng Yuan
 jinsheng.yuan@cranfield.ac.uk

Yun Tang
 yun.tang@cranfield.ac.uk

Weisi Guo
 weisi.guo@cranfield.ac.uk

Abstract—Mixed-precision computing, a widely applied technique in AI, offers a larger trade-off space between accuracy and efficiency. The recent proposed Mixed-Precision Over-the-Air Federated Learning (MP-OTA-FL) enables clients to operate at appropriate precision levels based on their heterogeneous hardware, taking advantages of the larger trade-off space while covering the quantization overheads of the mixed-precision modulation scheme with the OTA aggregation process. A key to further exploring the potential of the MP-OTA-FL framework is the optimization of client precision levels. The choice of precision level hinges on multifaceted factors including hardware capability, potential client contribution, and user satisfaction, among which factors can be difficult to define or quantify.

In this paper, we propose a precision planning framework that integrates Retrieval-Augmented Generation (RAG) LLMs and dynamic client profiling to optimize satisfaction and contributions. This includes a hybrid interface for gathering device/user insights and an RAG database storing historical quantization decisions with feedback. Experiments show that our method boosts satisfaction, energy savings, and global model accuracy in MP-OTA-FL systems.

Index Terms—Federated Learning, Human-centred, LLM Agent

I. INTRODUCTION

Over-the-Air Federated Learning (OTA-FL) [1] represents a novel paradigm in FL dedicated to wireless networks, which leverages the inherent randomness of physical-layer channel states and electromagnetic superposition for aggregating model updates. The same property is utilized to accommodate model parameters of multiple computational precision in Mixed-Precision OTA-FL [2]. The approach enables clients with heterogeneous hardware to participate in FL with better trade-offs between performance and energy efficiency than homogeneous precision FL systems while covering the quantization overheads in mixed-precision OTA aggregation.

One of the keys to enhancing such mixed-precision OTA-FL systems is to select the optimal quantization level for each client, which hinges on multifaceted factors within two main themes, user satisfaction, and client contribution. For user satisfaction, since available quantization levels depend on hardware specification, and the quantization to different levels directly translates to the corresponding performance metrics (e.g., accuracy, delay, energy efficiency), the most satisfying

The authors are with the Faculty of Engineering and Applied Sciences, Cranfield University, United Kingdom. The work is supported by EPSRC CHEDDAR: Communications Hub for Empowering Distributed cloud computing Applications and Research (EP/X040518/1) (EP/Y037421/1).



Fig. 1. User satisfaction and client contribution potentials in federated learning vary with contextual factors such as usage patterns and operational environment.

precision level can vary largely due to different usage patterns and performance sensitivity among users, even for those with identical hardware. As for client contribution, which depends on the quality, quantity, and distribution of client data, is infeasible to quantify directly due to the opacity of the client dataset. While there exist various proxy approaches for contribution estimation [3], [4], such estimations are often limited by the impractical demand for additional computation or the invalidation of assumptions in real-world scenarios.

To plan the optimal quantization level for each client, with main considerations of user satisfaction and client contribution, it's essential to collect and assess both the intrinsic technical factors such as hardware specifications (e.g., power states, compute capacity), model performance under quantization, and the extrinsic contextual factors such as device operational environments, user-specific usage patterns and preferences (as shown in Fig.1). However, collecting and assessing such factors poses the following challenges:

- **Challenge 1: Difficulty in comprehensive enumeration of extrinsic factors.** While intrinsic factors such as hardware specifications and quantized model performance can be systematically captured, extrinsic factors, including operational environments (e.g., ambient noise levels) and usage patterns (e.g., engagement frequency, input data classes) and user satisfaction are inherently highly dynamic and multifaceted, making exhaustive enumeration challenging and necessitating nuanced approaches.
- **Challenge 2: Complexity in quantifying individual client contribution.** Direct assessment of a client's potential contribution to global model accuracy relies

on client data characteristics including quantity, quality, and distribution, which are inherently inaccessible in FL due to user privacy. Existing proxy methods often involve client exclusive testing, and hence limited by the resulted computational overhead, highlighting the need for adaptive, context-aware estimation strategies.

The rapid-advancing Large Language Models (LLMs) offer a promising approach to address these challenges. In this paper, we propose a **RAG-based User Profiling for Precision Planning Framework**, integrating retrieval-augmented LLMs with dynamic client profiling to enhance user satisfaction. Specifically, to address Challenge 1, we develop a hybrid conversational interface combining available hardware information for capturing resource constraints with an interactive LLM-driven conversational agent to identify latent user needs and operational contexts. To address Challenge 2, we establish a knowledge database using Retrieval-Augmented Generation (RAG), which maintains historical quantization planning records alongside associated user feedback, thereby creating semantic mappings between contextual factors and user factors including satisfaction and contribution potentials to global model. Following the comprehensive collection of these factors, we compute reward-penalty metrics for each client's precision levels to optimize precision selection for subsequent learning rounds. User feedback gathered during this process is continuously integrated into the knowledge database, facilitating continuous refinement in precision planning. The contributions of this paper are as follows:

- 1) A RAG knowledge database that semantically links historical quantization decisions and user feedback, enabling data-driven estimation of the effect of contextual factors on user satisfaction and global model accuracy.
- 2) A dynamic client profiling mechanism that leverages an LLM agent-powered chat interface to extract user preference and contextual factors.
- 3) An experimental demonstration of the effectiveness of the framework through a mixed-precision FL voice assistant system in terms of user satisfaction, energy consumption and accuracy.
- 4) A open-sourced framework implementation¹ for the community to use, adapt and contribute.

The rest of the paper is organized as follows. Section II reviews related works. Section III presents the proposed framework. Section IV describes the experimental setup and results. Section V concludes the paper.

II. RELATED WORKS

A. Mixed-Precision Federated Learning

Federated learning, since its introduction by McMahan et al. [5], has been widely applied in privacy-sensitive distributed computing scenarios such as healthcare, finance, and IoT. Mixed-precision computation has been widely employed in deep learning, improving efficiency in both training and inference [6]. The insight behind such design is that different types

of layers in neural networks have different sensitivity to computation precision. Generalizing from this, mixed-precision OTA FL [2], with quantization overheads covered by OTA aggregation, offers a larger trade-off area between precision and performance, especially for those clients operating at the lowest precision levels due to most limited resources.

B. RAG-LLMs

RAG-LLMs, introduced by Lewis et al. [7], are a class of large language model that generate responses based on retrieved relevant information from external knowledge sources. RAG-LLMs have been deployed and achieved impressive performance in various NLP tasks such as user profiling for recommendation systems [8].

III. METHODOLOGY

A. Framework Overview

The proposed precision planning framework adopts a full-stack architecture as illustrated in Fig. 2. The framework comprises two core components: a user-profiling frontend featuring a chat interface and a backend server hosting an LLM-powered agent.

User Profiling Frontend A LLM agent-driven conversational interface for contextual factor discovery and satisfaction feedback collection. The conversation is tasked by the backend, and the primary tasks are as follows:

- At new device initialization, the user is prompted to provide perspective usage patterns and setup contexts, e.g., device location, intended usage scenarios, and user preferences.
- At the pre-aggregation stage, the user is queried for feedback on the performance of the operation, as well as potential context change, since the last feedback collection.
- In the case of changed hardware specifications, the user is prompted to update contextual factors and preferences.

Quantization Optimization Backend: A knowledge-enhanced processing stack containing:

- A RAG Context-Quantization-Feedback Database that archives precision decision history, including usage patterns, operational contexts, and user feedback.
- A RAG Hardware-Quantization-Performance Database that archives model performance (e.g., accuracy) with the associated hardware and precision level.
- LLM interview agent that interviews user preferences (e.g., user sensitivity for accuracy, response time and energy consumption) and usage patterns (e.g., noise level, usage frequency and type of interactions).
- Hardware specification extractor that collects device hardware information based on availability and user privacy settings. The parsed hardware specs are then used to query the knowledge database to estimate quantization-performance trade-offs on similar user hardware.

¹https://github.com/ntutangyun/user_in_the_loop_quantization_planning

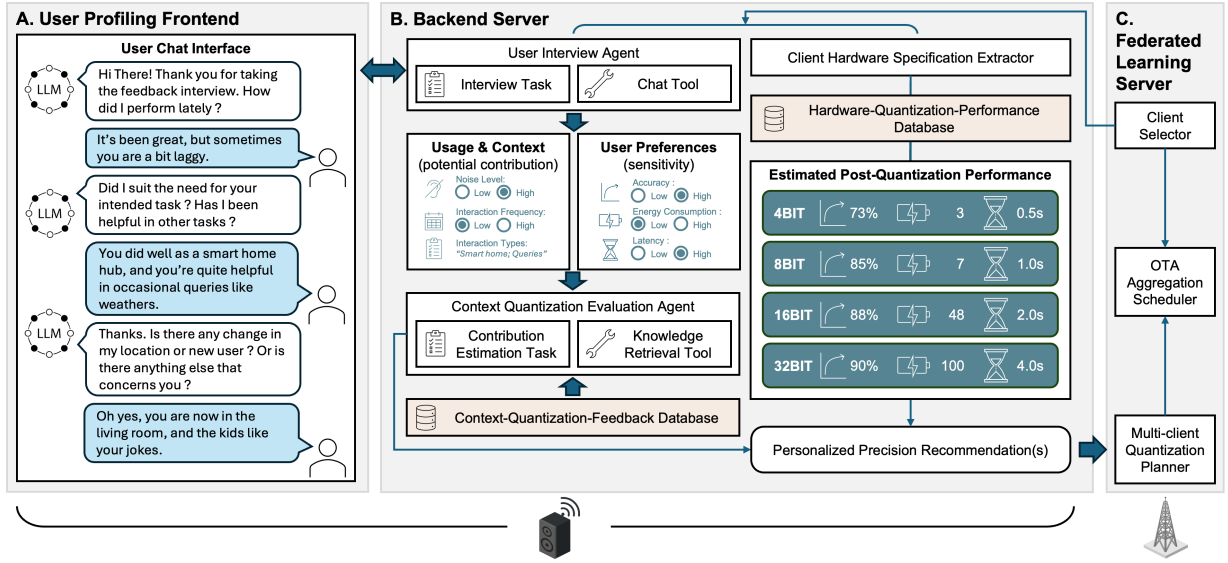


Fig. 2. User-in-the-loop Quantization Planning Framework Overview. The aim is to collect the user's feedback on the T round and select the optimal quantization level for the $T + 1$ round for the federated learning process.

- Context quantization evaluation agent that estimates the potential client contribution and user satisfaction at available precision levels based on interpreted contextual factors and retrieved hardware capabilities.

Federated learning server: The FL server mainly coordinates the following processes:

- Client selection: in the default setting, clients are scheduled to participate update and aggregation process regularly. The backend will launch the user profiling and context quantization evaluation process for selected clients.
- Multi-client quantization planning: when all selected clients have completed the profiling and evaluation process, the server will filter the clients with precision levels with similar merits, and choose the optimal precision levels that maximize communication resource utilization in mixed-precision OTA aggregation.
- Mixed-precision OTA aggregation: The FL server aggregates model updates from clients of their current precision levels. After aggregation, the server will send the updated model back to clients along with the optimal precision levels for the next round for them to quantize the received model accordingly.

B. RAG-based User Profiling

The RAG-based user profiling process collects and infers user preferences on performance and operational contexts through a user-friendly conversational interface.

1) *Contextual Factors:* The user preference for performance consists of three metrics: accuracy, energy consumption, and latency. These metrics are quantized by retrieving similar user cases from the knowledge database, based on the user's current feedback, operational contexts, and the estimated performance of their devices at the current precision levels. In comparison to conventional form-based feedback

collection, RAG-LLM can analyse the user's sensitivity in these metrics through wording nuances in their feedback, prioritize primary user concerns in performance, and hence, facilitate accurate adjustments to meet their expectations. In addition, the RAG-LLM can analyse and link these sensitivities to operational contexts, as the same user could have different expectations and sensitivities in different scenarios.

Apart from supporting performance feedback, operational contexts are also indicators of the potential contribution of the client to the global model, see Table I for examples of such contextual factors and their potential effects. Factors such as data quality, quantity and distribution, which can be inferred from these contextual factors, are essential for the client contribution estimation. FL service providers can use these inferred factors to estimate potential client contributions at different precision levels, and hence, select the optimal precision level for each client based on their learning strategies.

2) *RAG Database and LLM Integration:* To support the LLM agents, we build a Context-Quantization-Feedback database, which stores the feedback from users of different contextual factors on performance at different quantization levels. When user feedback and contexts are collected via the chat interface, the LLM agent will retrieve similar user cases from the database, and estimate the user satisfaction, preference and client contribution at different precision levels based on the retrieved cases.

3) *User Profiling Pipeline:* The user profiling pipeline consists of the following steps:

- 1) **Hardware specification extraction:** The backend extracts the hardware specification of the user device, including processor specs, RAM size, and power states.
- 2) **Hardware quantization performance trade-off retrieval:** The backend queries the knowledge database

TABLE I
EXAMPLES OF CONTEXTUAL FACTORS AND INFERABLE FACTORS

Contextual Factor	Inferable Factor	Examples
Device location	Input noise level	Bedroom → Low noise; Living room → High noise
Interaction time	Input noise level, data quantity	Daytime → High noise, High quantity; Nighttime → Low noise, Low quantity
Interaction frequency	Data quantity	High frequency → High quantity
Task Type	Data distribution	Smart home hub → Short requests

for the quantization-performance trade-off on the same or similar hardware.

- 3) **User interview feedback collection:** The agent prompts the user to provide feedback on the current performance and potential context changes since the last feedback collection, see Fig. 2-A for a chat example.
- 4) **Contextual factor inference:** The LLM agent infers user preferences and contexts from past conversations.
- 5) **User preference and contextual factor retrieval:** The agent retrieves similar user cases from the knowledge database with inferred user preferences and contexts.
- 6) **User satisfaction and client contribution estimation:** The agent estimates the potential client contribution and user satisfaction at available precision levels based on retrieved contextual factors and hardware capability.

C. Context-Quantization Evaluation

We define a reward-penalty model for determining the optimal quantization level for each client in a federated learning (FL) setting. The model considers multiple factors, each with an associated user-defined sensitivity weight. Assume:

- $\mathcal{F}$: Set of factors (e.g., accuracy, energy cost, latency).
- q : Quantization level assigned to a client.
- w_f : Sensitivity weight of factor $f \in \mathcal{F}$, where $\sum_{f \in \mathcal{F}} w_f = 1$.
- $R_f(q)$: Reward obtained from operating at quantization level q for factor f (e.g., improved accuracy).
- $P_f(q)$: Penalty incurred by operating at quantization level q for factor f (e.g., energy consumption).
- C_q : Contribution multiplier for potential client contribution operating at quantization level q .

Then, the total reward and total penalty for quantization level q are computed as the following weighted sums:

$$R_{\text{Total}}(q) = C_q \cdot \sum_{f \in \mathcal{F}} w_f \cdot R_f(q) \quad (1)$$

$$P_{\text{Total}}(q) = \sum_{f \in \mathcal{F}} w_f \cdot P_f(q) \quad (2)$$

$$\text{Satisfaction Score}(q) = R_{\text{Total}}(q) - P_{\text{Total}}(q) \quad (3)$$

The optimization goal is to select the quantization level q that maximizes the Satisfaction Score defined as the difference between total reward and total penalty:

$$q^* = \arg \max_q (\text{Satisfaction Score}(q)) \quad (4)$$

TABLE II
SMART VOICE ASSISTANT DATA DISTRIBUTION

Category	Entertainment	Smart Home	General Query	Personal Request
Percentage	32.7%	16.0%	31.9%	19.4%

IV. EXPERIMENTS

A. Experimental Setup

We validate our proposed RAG-based precision planning framework on a federated smart voice assistant system with the Automatic Speech Recognition (ASR) task. The federation consists of 100 simulated clients with diverse hardware capabilities and Gaussian distributed sensitivity to performance factors including accuracy, energy savings, and latency. We define the following experimental settings:

Dataset and Model: The model structure is DeepSpeech2 [9], and the federated model is trained for 100 communication rounds. For client datasets, we filter the Common Voice dataset [10] with keywords related to the four main uses of smart voice assistants, *Entertainment*, *Smart Home*, *General Query* and *Personal Request*. We define these categories and their distribution, see Table II, based on the usage statistics from the PWC research report [11].

Metrics and Comparison: To showcase the advantage of our RAG-based user profiling precision planning framework, we compared it on the same federated learning system but with a unified standard precision planner, i.e., divide users in tiers by their hardware capabilities and assign the same precision level to each tier regardless of user preference and operational contexts. We measure the following metrics:

- **User Satisfaction Score:** the user satisfaction score defined in Equation 3.
- **Relative Energy Cost:** we do not directly measure energy costs, instead, we measure the relative energy cost compared to the highest available precision level, and therefore the relative energy cost is a percentage below 100%.
- **Final Global Model Accuracy:** The final word accuracy of the global model after 100 communication rounds.

B. Results

1) *User Satisfaction versus Energy Cost:* Our RAG-based user profiling precision planning framework generates personalized standards based on user preference and their operational contexts, resulting in a more accurate satisfaction

TABLE III
WORD ACCURACY OF THE GLOBAL MODEL AFTER 100 COMMUNICATION ROUNDS BY CLASSES WITH DIFFERENT STRATEGIES.

FL Strategy	Entertainment	Smart Home	General Query	Personal Request
a. Sample Equal (Baseline)	71.2	68.5	78.6	69.4
b. Class Equal	68.7 (-2.5%)	77.3 (+8.8%)	76.4 (-2.2%)	74.8 (+5.4%)
c. Majority Centric	75.5 (+4.3%)	66.2 (-2.3%)	81.9 (+3.3%)	67.8 (+1.6%)

Compared to the default strategy, with b) class equal strategy, biased towards minority classes, our framework trades 2% accuracy of the majorities for 5% of that of the minorities; while with c) majority centric strategy, our framework extended the accuracies of majority classes by 4% with 3% lower accuracies for minority classes.

estimate compared to the FL system that plans precision levels with unified standards, and the average satisfaction score is 10% higher (0.66 compared to 0.60) while saving about 20% energy. We also tested that when energy savings is the top priority of the mixed-precision FL system, our framework can trade 22% average satisfaction score ((0.47 compared to 0.60)) for a total of 28% energy savings.

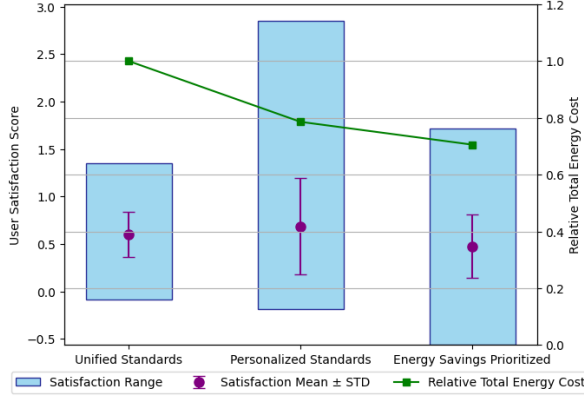


Fig. 3. Distribution of User Satisfaction Scores and Relative Energy Cost. Compared to planning precision levels with unified standards, personalized standards can achieve 10% higher average satisfaction score, and 20% of energy cost. When prioritise the federated system towards energy savings, 22% satisfaction score can be traded for a total of 28% energy saving.

2) *Global Model Performance*: Estimation of potential client contribution to the global model depends on the training strategy. We experimented with three different strategies with our framework: a) default FedAvg [5] i.e. treat every sample equally; b) class equal strategy, attempts higher precision levels to samples of minority classes; c) majority centric strategy, attempts higher precision levels to samples of majority classes. Our RAG-based framework can estimate data distribution via contextual factors without breaching user privacy. Refer to actual data distribution in Table II, see Table. III compared to accuracies of FedAvg [5], our framework improved the accuracies of minority classes (smart home and personal request) and majority classes (entertainment and general query) with the corresponding biased strategies.

V. CONCLUSION

In this paper, we proposed a novel precision planning framework for Mixed-Precision Over-the-Air Federated Learning (MP-OTA-FL) systems, utilizing Retrieval-Augmented Generation (RAG)-powered Large Language

Models (LLMs) for dynamic client profiling and quantization optimization. The proposed framework addresses key challenges in quantization-level selection and produces personalized precision planning through a conversational user profiling interface and dynamic RAG database utilization. Experimental evaluations demonstrated significant improvements in user satisfaction, energy savings, and global model accuracy compared to traditional quantization approaches with unified standards. Furthermore, our implementation is open-sourced to foster community-driven innovation in human-centred federated learning.

REFERENCES

- [1] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.
- [2] J. Yuan, Z. Wei, and W. Guo, "Mixed-precision federated learning via multi-precision over-the-air aggregation," 2024. [Online]. Available: <https://arxiv.org/abs/2406.03402>
- [3] R. Jia *et al.*, "Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification?" in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8235–8243.
- [4] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1167–1176.
- [5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54, 2017, pp. 1273–1282.
- [6] P. Micikevicius *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.
- [7] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
- [8] Y. Deldjoo *et al.*, "A review of modern recommender systems using generative models (gen-recsys)," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 6448–6458.
- [9] D. Amodei *et al.*, "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 173–182.
- [10] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.
- [11] PwC, "The impact of voice assistants on consumer behavior," 2025, accessed: 2025-03-14. [Online]. Available: <https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html>



DSPACE

<https://dspace.org/>

RAG-based user profiling for precision planning in mixed-precision over-the-air federated learning

Yuan, Jinsheng; Tang, Yun; Guo, Weisi

2025-10-19

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/vtc2025-fall65116.2025.11310436>

Yuan J, Tang Y, Guo W. (2025) RAG-based user profiling for precision planning in mixed-precision over-the-air federated learning. In: Proceeding of the 2025 IEEE 102nd Vehicular Technology Conference (VTC2025-Fall), 19-22 Oct 2025, Chengdu, China