Reinforcement Learning With Large Language Models (LLMs) Interaction For Network Services

Hongyang Du, Ruichen Zhang, Dusit Niyato, *Fellow, IEEE*, Jiawen Kang, Zehui Xiong, and Dong In Kim, *Fellow, IEEE*

***Abstract*—Artificial Intelligence-Generated Content (AIGC)-related network services, especially image generation-based services, have garnered notable attention due to their ability to cater to diverse user preferences, which significantly impacts the subjective Quality of Experience (QoE). Specifically, different users can perceive the same semantically informed image quite differently, leading to varying levels of satisfaction. To address this challenge and maximize network users' subjective QoE, we introduce a novel interactive artificial intelligence (IAI) approach using Reinforcement Learning With Large Language Models Interaction (RLLI). RLLI leverages Large Language Model (LLM)-empowered generative agents to simulate user interactions, thereby providing real-time feedback on QoE that encapsulates a range of user personalities. This feedback is instrumental in facilitating the selection of the most suitable AIGC network service provider for each user, ensuring an optimized, personalized experience.**

***Index Terms*—Reinforcement learning, generative artificial intelligence, large language models**

I. Introduction

The demand for artificial intelligence-generated content (AIGC) services in areas including multimedia and business [1] is propelled by advanced generative AI (GAI) models, which offer scalable and consistent output in text and imagery [2]. For instance, ChatGPT attracted more than 100 million active users within two months [3], highlighting its impact on text-based interactions [3]. In visual content generation, Stable Diffusion's capacity to create images from text prompts shows significant progress in multi-modal technologies [4]. The widespread adoption of AIGC-related network services in human societies indicates a notable transition to Interactive AI (IAI) as the next evolutionary phase of GAI [5]. This shift is redefining how humans interact with content and underscores the dynamic progression in human-AI interaction.

However, Quality of Experience (QoE) maximization in AIGC-related network services emerges as a critical challenge, due to the subjective nature of human perception, which extends beyond objective image quality metrics [6], [7]. Four images generated by different GAI models in Part B of Fig. 2 exemplify this complexity, showing four distinct images that are generated under the same prompt *"A Cat runs in the Street"*. Although each image is of high quality, these images cater to different personality profiles, affecting QoE evaluations variably. Thus, network designers aspire to develop service models and service management models tailored to individual user personalities to maximize QoE. However, the absence of a definitive mathematical model for QoE complicates this optimization process. While some studies have adopted psychological laws to approximate users' subjective QoE [8], these methods often oversimplify, failing to address the multifaceted nature of real-world applications. Another solution is leveraging Reinforcement Learning with Human Feedback (RLHF) paradigms for *management models* training, which requires continuous QoE feedback from experts. This method is costly, ethically contentious, and challenging to execute in real-time, leading to the research question:

How to obtain human-aware subjective QoE feedback efficiently and design the communication and computing resource allocation algorithm for network services?

Addressing this challenge requires the exploration of cutting-edge approaches, among which IAI stands out as a promising solution [5]. IAI focuses on designing AI models that learn and adapt through user interaction, progressively advancing AI models' performance and operational efficacy. This paradigm shift, from static to dynamic learning systems, equips IAI-enhanced networks with the ability to offer tailored responses and proactively adapt, marking a significant stride in personalized AI-based service management. We propose the Reinforcement Learning With LLM Interaction (RLLI) algorithm as one step towards IAI, incorporating Deep Reinforcement Learning (DRL) for its suitability in dynamic environments [9] and Large Language Model (LLM) for advanced knowledge understanding and generative capabilities. On the one hand, DRL regards user QoE as quantifiable rewards and circumvents the complexity of mathematically modeling subjective QoE. On the other hand, LLM-empowered generative agents can represent real AIGC users with various personalities to generate QoE feedback, minimizing human resource input and associated ethical risks. By embedding human personality traits into generative agents through prompts [10], these generative agents can simulate human interaction in the training

This research is supported by NSFC under grant No. 62102099, the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research & Development Programme, DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019 and FCP-ASTAR-TG-2022-003), Energy Research Test-Bed and Industry Partnership Funding Initiative, Energy Grid (EG) 2.0 programme, DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) programme, and MOE Tier 1 (RG87/22), the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI 20210204), the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00258639) supervised by the IITP.

H. Du, R. C. Zhang, and D. Niyato are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: hongyang001@e.ntu.edu.sg, ruichen.zhang@ntu.edu.sg, dniyato@ntu.edu.sg).

J. Kang is with the School of Automation, Guangdong University of Technology, China. (e-mail: kavinkang@gdut.edu.cn)

Z. Xiong is with the Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore (e-mail: zehui_xiong@sutd.edu.sg)

D. I. Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea (email:dikim@skku.ac.kr).

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

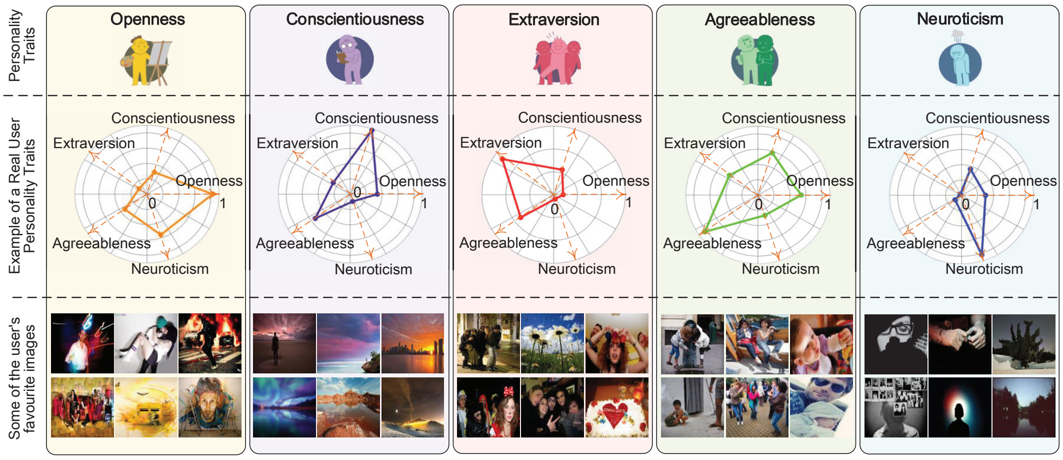


Fig. 1. Five types of personality traits in the Big-Five personality model: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. We show real user personality scores from the PsychoFlickr database, along with examples of the images that they prefer.

of DRL algorithms.

In AIGC network services, note that the difference in GAI models introduces a difference in the generated images. Therefore, the quality of generated images is impacted by the AIGC service provider selection scheme. Our edge-based IAI solution tackles a new optimization dimension introduced by this dependency, aiming to enhance the AIGC service and improve user QoE. The contributions of this paper are summarized as follows.

- We present a QoE feedback scheme by using LLM-empowered generative agents to simulate the human's different personalities. With the aid of prompts and assigning one agent per user, generative agents can mimic users of diverse subjective preferences, delivering evaluations of the quality of generated images.
- We propose an IAI algorithm, i.e., RLLI, with LLM-generated QoE at its core. Furthermore, we consider the AIGC service provider selection problem to show the effectiveness of our proposed RLLI method. The goal is to determine the optimal AIGC service provider to serve the user based on the user's personality.

II. Reinforcement Learning With LLMs Interaction

In this section, we propose the RLLI method. Specifically, we first discuss aesthetic-aware QoE modeling, and then we explain how to use the LLM-empowered generative agents as evaluators to feedback their subjective QoE values as the reward for DRL algorithms.

A. Aesthetic-aware QoE Modeling

1) Big Five Personality Traits: Incorporating generative agents into the DRL training process allows for the meaningful integration of human subjective factors, specifically aesthetic preferences, a critical component affecting QoE in image-related AIGC services. Research has shown that aesthetic preferences can be influenced by an individual's personality traits [11]–[13]. The Big Five personality traits [13], also known as the Five-Factor Model (FFM), is a widely accepted framework for understanding individual differences in personality[1]. The Big Five model consists of five broad dimensions of personality traits including

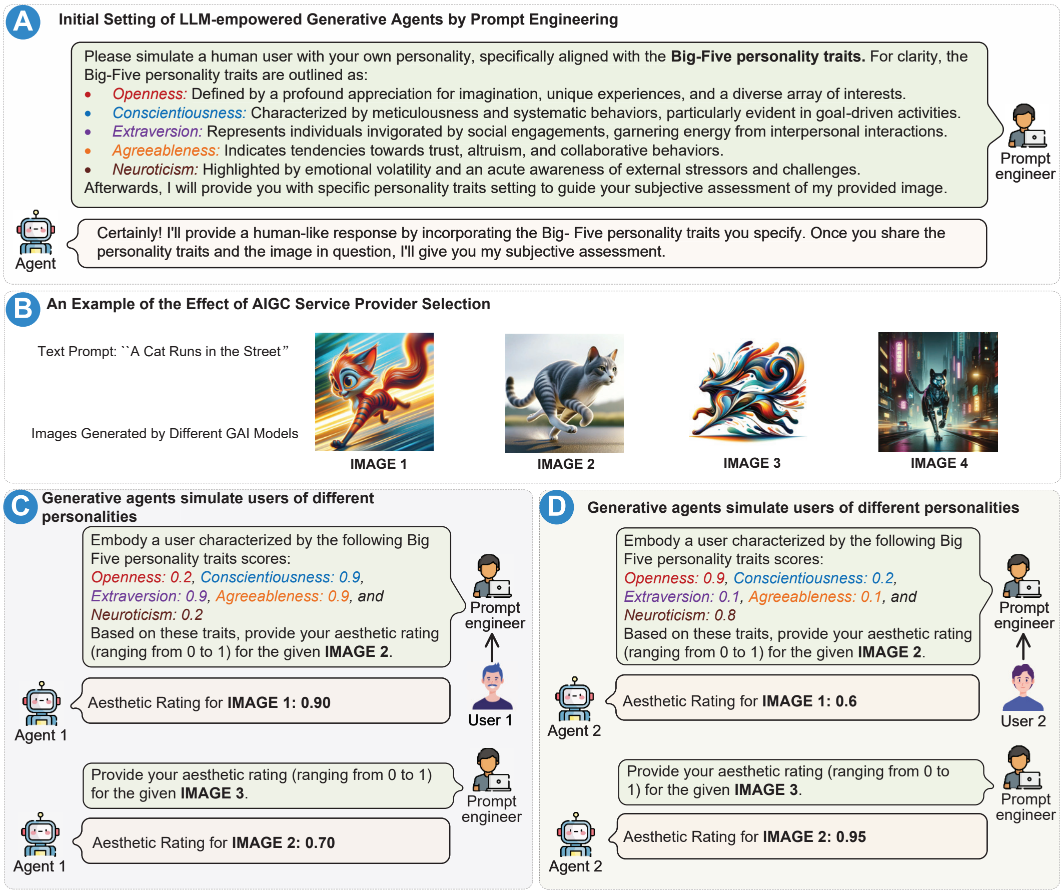
- *Openness:* A trait characterized by a deep affinity for imagination, novel experiences, and a broad spectrum of interests.
- *Conscientiousness:* Denotes meticulousness and structure, often manifesting in methodical, goal-oriented behaviors.
- *Extraversion:* Embodies individuals who thrive in social interactions, drawing energy from a company of others.
- *Agreeableness:* Reflects proclivities towards trustworthiness, altruism, and prosocial behaviors.
- *Neuroticism:* Typified by emotional fluctuations and heightened sensitivity to environmental stressors and adversities.

Each of these traits represents a continuous spectrum of personality characteristics. Considering that a recent study has demonstrated that LLMs can effectively simulate the Big Five personality traits and achieve an 82.8% alignment with human perceptions of these characteristics [14], we use the Big Five personality model as the basis for enabling LLM-empowered generative agents to mimic AIGC service users' personalities [10]. In Fig. 1, we display five exemplary user profiles from the PsychoFlickr dataset [12]. These profiles provide empirical evidence for the relationship between individual personality traits and aesthetic preferences.

2) Prompt Design for LLM-empowered Generative Agents: The LLM-empowered generative agents present a powerful mechanism to feedback human-aware subjective QoE values for the generated content. A critical aspect is the initial prompts that guide the generative agents' subjective QoE assessment [10]. The initial prompts setting process is illustrated in Parts A and C of Fig. 2, including general setup and generative agent-specific settings:

- In the general setup stage, the Big-Five personality traits are introduced to all K generative agents to enhance their understanding and responsiveness to these traits.

[1] One example of open source Big Five personality traits tests: https://bigfive-test.com.



A
Initial Setting of LLM-empowered Generative Agents by Prompt Engineering
Prompt engineer
Agent

Please simulate a human user with your own personality, specifically aligned with the **Big-Five personality traits.** For clarity, the Big-Five personality traits are outlined as:

- *Openness:* Defined by a profound appreciation for imagination, unique experiences, and a diverse array of interests.
- *Conscientiousness:* Characterized by meticulousness and systematic behaviors, particularly evident in goal-driven activities.
- *Extraversion:* Represents individuals invigorated by social engagements, garnering energy from interpersonal interactions.
- *Agreeableness:* Indicates tendencies towards trust, altruism, and collaborative behaviors.
- *Neuroticism:* Highlighted by emotional volatility and an acute awareness of external stressors and challenges.

Afterwards, I will provide you with specific personality traits setting to guide your subjective assessment of my provided image.

Certainly! I'll provide a human-like response by incorporating the Big- Five personality traits you specify. Once you share the personality traits and the image in question, I'll give you my subjective assessment.

B
An Example of the Effect of AIGC Service Provider Selection
Text Prompt: ``A Cat Runs in the Street"
Images Generated by Different GAI Models
IMAGE 1
IMAGE 2
IMAGE 3
IMAGE 4
C
Generative agents simulate users of different personalities
Prompt engineer
User 1
Agent 1
Prompt engineer
Agent 1

Embody a user characterized by the following Big Five personality traits scores: *Openness: 0.2*, *Conscientiousness: 0.9*, *Extraversion: 0.9*, *Agreeableness: 0.9*, *and Neuroticism: 0.2*
Based on these traits, provide your aesthetic rating (ranging from 0 to 1) for the given **IMAGE 2**.

Aesthetic Rating for **IMAGE 1: 0.90**

Provide your aesthetic rating (ranging from 0 to 1) for the given **IMAGE 3**.

Aesthetic Rating for **IMAGE 2: 0.70**

Embody a user characterized by the following Big Five personality traits scores: *Openness: 0.9*, *Conscientiousness: 0.2*, *Extraversion: 0.1*, *Agreeableness: 0.1*, *and Neuroticism: 0.8*
Based on these traits, provide your aesthetic rating (ranging from 0 to 1) for the given **IMAGE 2**.

Aesthetic Rating for **IMAGE 1: 0.6**

Provide your aesthetic rating (ranging from 0 to 1) for the given **IMAGE 3**.

Aesthetic Rating for **IMAGE 2: 0.95**

Fig. 2. Prompts for LLM-empowered generative agents settings. **Part A** illustrates the initial setup to acquaint the generative agents with the Big-Five personality traits. **Part B** demonstrates a case where the different numbers of shared denoising steps lead to stylistic differences in the final generated images. **Part C** and **Part D** present two user personality trait configurations and the corresponding evaluated scores of generated images. Note that the *prompt engineer* is a component of our proposed RLLI, which does not require human involvement and merely acquires the personality traits of the target user for the initial setting.

- In the generative agent-specific settings stage, the users' Big-Five personality traits are individually configured for generative agents, enabling generative agents' similar subjective assessments to given images as real users.

With the Big-Five model, the personality of the k-th user, i.e., $\boldsymbol{u}_k$, can be expressed as

$$\boldsymbol{u}_k = \left[o_k, c_k, e_k, a_k, n_k\right], \tag{1}$$

where $k = 1, \ldots, K$, and each element in $\boldsymbol{u}_k$ corresponds to a score in one of the Big Five personality traits. Note that the vector $\boldsymbol{u}_k$ is significant in our RLLI framework in tailoring the subjective QoE assessment according to individual preferences. Specifically, $\boldsymbol{u}_k$ serves as the *personality traits tuning* for generative agents. Furthermore, $\boldsymbol{u}_k$ acts as the user identifier that can be used for the DRL-based resource allocation algorithm design, similar to user representation in AI-based recommendation systems where user preferences are captured and embedded to provide personalized recommendations [15].

B. Reinforcement Learning With LLMs Feedback Framework

LLMs are typically designed for language tasks, yet their ability to interpret various task instructions articulated in language has shown promise for acting as universal interfaces for general-purpose assistants [16], [17]. For our proposed RLLI framework, to leverage the inferential capabilities of LLM to simulate users with different personality traits, the LLM-empowered generative agents' instructional capacity has to be extended to encompass visual domains.

1) Visual Instruction Tuning (VIT): The VIT framework elevates the instruction tuning paradigm into the multi-modal sphere, leveraging LLMs to process both textual and visual information [18]. With the demonstrated proficiency of LLMs such as ChatGPT and GPT4 in executing complex instructions, we has seen the rise of accessible open-source LLMs like LLaMA [19], further simplifying the adoption of VIT. Specifically, VIT processes an image input to extract features that are then converted into language embedding tokens using a trainable projection matrix [18]. Training involves generating multi-turn conversational sequences from each image to form instructions and predict answer tokens using an auto-regressive objective. The model undergoes a two-stage instruction tuning: initially, it aligns features using image-text pairs to train the visual tokenizer while keeping LLM and visual encoder weights static. Subsequently, it fine-tunes both the projection matrix and LLM weights using varied datasets to improve response diversity. After VIT, an image can be processed by the LLM-powered generative agents. This advancement extends the GA's capabilities beyond merely handling text and allows for the subjective evaluation of the image's quality.

In this paper, we use the LLaMA-based LLaVA [18], i.e., an end-to-end trained large multi-modal model, to empower generative agents for our RLLI algorithm. Note that due to the generalizable nature of LLMs in understanding and generating language prompts [10], [19], our method remains applicable and effective across a range of LLMs, ensuring broad adaptability and relevance.

2) Reinforcement Learning with LLMs Interaction Framework: RL trains agents to maximize a reward function through interaction with an environment. Reinforcement Learning with Human Feedback (RLHF) enhances this process by introducing human insights into the policy optimization, often through demonstrations or comparative feedback [20], significantly improving conversational agents like ChatGPT. Nonetheless, both RL and RLHF encounter important challenges:

- **Real-Time Constraint.** Delayed feedback in RLHF hinder its applicability in scenarios demanding immediate response. Moreover, the continuous need for human expertise input raises costs.
- **Expert Availability.** Consistent expert interaction is challenging, inconsistent, and thus unreliable. Furthermore, the varying quality of human feedback affects the management model training.
- **Ethical and Privacy Risk.** Human-in-the-loop interaction system may present data confidentiality concerns in sensitive applications. For example, some AI-generated images are inappropriate for humans in all ages to view.

To address these challenges, we introduce RLLI, where real users can leverage LLM-empowered generative agents to provide feedback for DRL model training. These generative agents mimic users with varied personalities and provide immediate, context-aware feedback in the form of subjective QoE rewards. Consequently, RLLI offers a real-time, scalable, and financially efficient solution, mitigating the inherent constraints of RL and RLHF. The general algorithm for implementing RLLI is shown as **Algorithm** 1. Specifically, the management model initializes with parameters $\boldsymbol{\xi}$, while K LLM-empowered generative agents simulate diverse user feedback. Each episode e begins with state $\boldsymbol{s}$ and iterates until a terminal state is reached. Here, a terminal state is the endpoint of an episode, indicating task completion, step limit reached, or a failure event. Actions $\boldsymbol{a}$ are generated via policy $\pi_{\boldsymbol{\xi}}(\boldsymbol{s})$, with rewards aggregated from the agents' subjective QoE assessments. State transitions and experiences are stored in a Replay Buffer, facilitating policy parameter updates through experience replay. This iterative process, across E episodes, refines the model's decision-making capabilities by integrating feedback from generated agents interaction, culminating in a robustly trained management model.

III. Case Study

In this section, we consider the AIGC service provider selection problem and show the effectiveness of RLLI.

A. User-centric QoE Maximization Problem

As shown in Fig. 3, we consider the AIGC-as-a-service concept highlighting the capability of networks to support AIGC services by deploying GAI models on edge servers. The selection of an AIGC service provider is crucial due to user preferences and the diversity in image styles generated by different GAI models, influenced by their unique training datasets. The objective is to maximize the aggregated QoE of users. In our model, we consider K users and L GAI models, which correspond to L different AIGC Service Providers (ASPs). The optimization problem can be formulated as follows:

$$\max \sum_{i=1}^{K} QoE_i(\mathrm{ASP}_j) \tag{2}$$

Here, $QoE_i(\mathrm{ASP}_j)$ represents the QoE of the i-th user when served by j-th ASP. The optimization seeks to allocate each user to an ASP in a manner that maximizes the total QoE across all users.

Algorithm 1 Reinforcement Learning with Large Language Model Interaction (RLLI)

Initialize: The management model with parameters $\boldsymbol{\xi}$, LLM-empowered generative agents $\boldsymbol{K}$ to simulate K users

Output: The trained management model $\boldsymbol{\xi}$

```
Input prompts to $K$ generative agents, letting them to simulate users with different personalities
for each episode $e = 1, 2, \dots, E$ do
    Initialize state $s$
    while $s$ is not terminal do
        Generate action $a$ using policy $\pi_{\xi}(s)$
        Obtain reward $r = \sum_{k=1}^{K} \mathrm{Agent}_k(s, a)$
        Transition to new state $s'$
        Store transition $(s, a, r, s')$ in Replay Buffer
        Sample a random minibatch of transitions from Replay Buffer
        Update policy parameters $\xi$
        $s \leftarrow s'$
```

B. PPO With LLMs Interaction for ASP Selection

To address the ASP Selection challenge, we propose an innovative approach utilizing Proximal Policy Optimization (PPO) integrated with LLMs. PPO, a cutting-edge RL algorithm, is distinguished for its stability and efficacy in complex environments. Its integration with LLMs enables context-aware decision-making. Moreover, the action space, state space, and reward function are designed as follows:

1) **Actions:** The action space $\mathcal{A}$ is defined as the current user task to one of the available ASPs, where the cardinal number of the action space $\mathcal{A}$ is L.

2) **States:** In defining the state space, we aim to incorporate as much relevant environmental information as possible for the considered problem. In the considered system, the state space is composed of the current information of the selected action vector $\boldsymbol{a}$, and the instant reward $\boldsymbol{r}$. As a result, the state space is given by

$$\mathcal{S} = \{\{\mathbf{a}_i\}, \{r_i\}\}, \tag{3}$$

where the cardinal number of the action space $\mathcal{S}$ is $L + 1$.

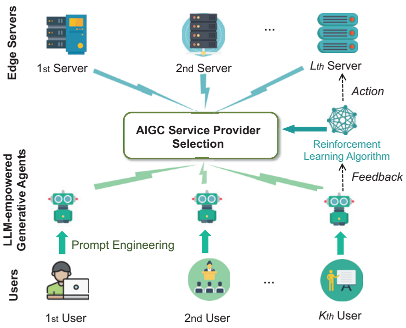


Fig. 3. System model for the AIGC service provider selection problem.

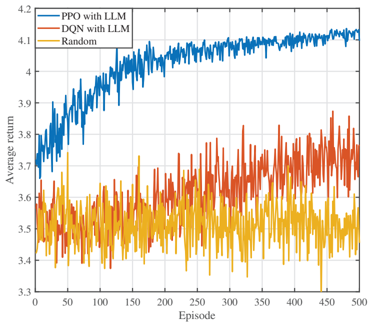


Fig. 4. The average return versus the number of iterations.

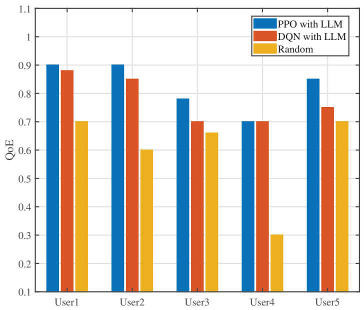


Fig. 5. The QoE feedback for different users.

3) **Rewards:** The reward function takes into account the objective function, i.e., sum QoE, from LLMs. It is designed as the instant reward term indicating the sum QoE of all users. As a result, the reward is given by $r = \sum_{i=1}^{K} QoE_i$.

C. Experiments Results

Without loss of generality, we consider $K = 5$ and $L = 4$ in the training process. Fig. 4 shows the cumulative return value obtained with the increase of the number of iterations using the different RL with LLMs interaction, where the corresponding curves are smoothed via a sliding window to provide a clearer overall trend of the raw results. In the testing stage, Fig. 5 further shows that, regardless of the types of RL with LLMs, the RL-based strategy can converge with an increment of iterations. It also shows that the proposed PPO with LLMs interaction is superior to the benchmark, e.g., random policy and Deep Q-Network (DQN)-based DRL algorithm, where the corresponding performance gains are notable. This superiority can be attributed to the unique mechanism of PPO, which balances exploration and exploitation efficiently. PPO's clipped objective function prevents drastic policy updates, ensuring stable and consistent learning.

IV. Conclusion

We proposed a novel approach to enhancing user QoE in the AIGC network service, focusing on image generation services. Our primary contribution is the RLLI method, designed to address the subjective nature of user QoE. By employing LLM-empowered generative agents, RLLI provided real-time, personalized feedback on QoE, reflecting a spectrum of user preferences. This approach enabled a more informed selection of service providers, optimizing the user QoE by aligning it with individual perceptual and interpretative preferences. Our methodology demonstrates the potential of using advanced AI techniques to enhance user engagement and satisfaction in network services, paving the way for more user-centric content generation models.

References

[1] X. Guo and L. Zhao, "A systematic survey on deep generative models for graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5370–5390, May 2022.

[2] H. Du, Z. Li, D. Niyato, J. Kang, Z. Xiong, D. I. Kim *et al.*, "Enabling AI-generated content (AIGC) services in wireless edge networks," *IEEE Wireless Mag.*, to appear, 2023.

[3] B. D. Lund and T. Wang, "Chatting about ChatGPT: how may AI and GPT impact academia and libraries?" *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, Mar. 2023.

[4] S. AI, "Stable diffusion," https://stability.ai/.

[5] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for human-AI interaction," in *Proc. 2019 Chi Conf. Human Fact. Comput. Syst.*, 2019, pp. 1–13.

[6] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[7] P. Juluri, V. Tamarapalli, and D. Medhi, "Measurement of quality of experience of video-on-demand services: A survey," *IEEE Commun. Surv. Tut.*, vol. 18, no. 1, pp. 401–418, Jan. 2015.

[8] H. Du, J. Liu, D. Niyato, J. Kang, Z. Xiong, J. Zhang, and D. I. Kim, "Attention-aware resource allocation and QoE analysis for metaverse xURLLC services," *IEEE J. Selec. Areas Commun.*, vol. 41, no. 7, Jul. 2023.

[9] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surv. Tut.*, vol. 21, no. 4, pp. 3133–3174, Apr. 2019.

[10] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

[11] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3898–3910, Jan. 2020.

[12] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 213–222.

[13] G. Saucier and L. R. Goldberg, "What is beyond the Big Five?" *J. Pers.*, vol. 66, pp. 495–524, 1998.

[14] X. Wang, Y. Fei, Z. Leng, and C. Li, "Does role-playing chatbots capture the character personalities? Assessing personality traits for role-playing chatbots," *arXiv preprint arXiv:2310.17976*, 2023.

[15] H. Ko, S. Lee, Y. Park, and A. Choi, "A survey of recommendation systems: Recommendation models, techniques, and application fields," *Electronics*, vol. 11, no. 1, p. 141, Jan. 2022.

[16] X. Ma, G. Fang, and X. Wang, "LLM-Pruner: On the structural pruning of large language models," arXiv preprint arXiv:2305.11627, 2023.

[17] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?" arXiv preprint arXiv:2305.03514, 2023.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," arXiv preprint arXiv:2304.08485, 2023.

[19] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

[20] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.