Yuxuan CHEN, Rongpeng LI, Xiaoxue YU, Zhifeng ZHAO, Honggang ZHANG, 2025. Adaptive layer splitting for wireless large language model inference in edge computing: a model-based reinforcement learning approach. *Frontiers of Information Technology & Electronic Engineering,* 26(2):278-292. https://doi.org/10.1631/FITEE.2400468

Adaptive layer splitting for wireless large language model inference in edge computing: a model-based reinforcement learning approach

Key words: Large language models (LLMs); Edge computing; Model-based reinforcement learning (MBRL); Split inference; Transformer

Corresponding author: Rongpeng LI
E-mail: lirongpeng@zju.edu.cn
ORCID: https://orcid.org/0000-0003-4297-5060

Motivation

- The field of natural language processing (NLP) has recently experienced transformative changes, driven by the rapid advancement of large language models (LLMs) such as GPT-4 and Gemini. These models are highly proficient at generating human-like text, catalyzing progress across various domains. However, LLMs' substantial computational demands often exceed the processing capacities of communication-limited user equipment (UE), which makes the integration of edge computing a complementary paradigm to enhance computational efficiency and privacy.

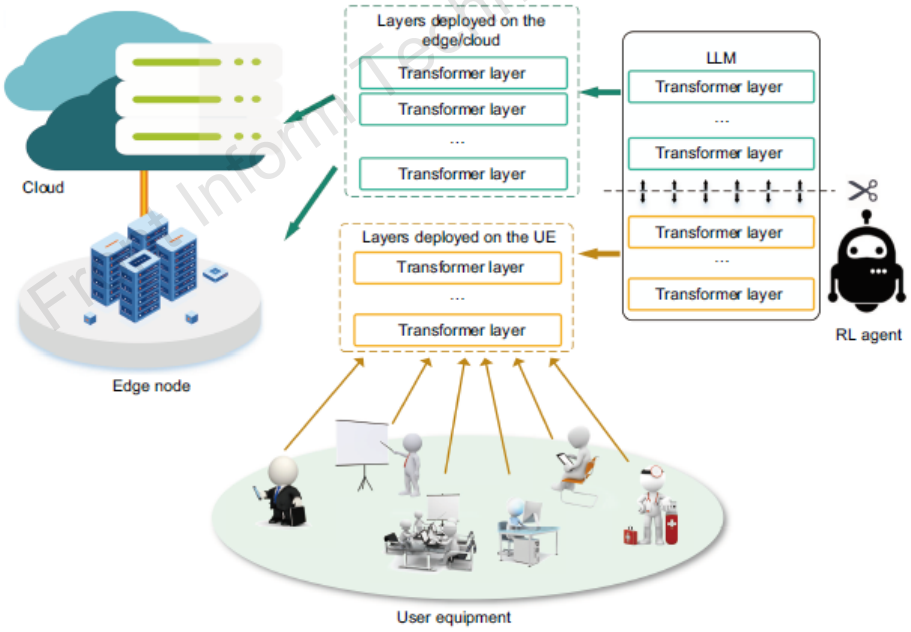

Fig. 1 A high-level architecture of the framework, depicting the distribution of the large language model (LLM) across the edge and the user equipment (UE), highlighting the role of the reinforcement learning (RL) agent in managing interactions between the LLM and wireless networks

Main idea

- To address the computational and communication challenges of deploying large language models (LLMs) in edge computing environments, this paper introduces a framework inspired by model-based reinforcement learning (MBRL). The framework focuses on determining the optimal splitting point between user equipment (UE) and edge nodes, balancing inference performance and computational load. This approach adapts to varying network conditions, minimizing the impact of unreliable wireless channels, and significantly reduces the computational cost of frequent performance evaluations.

Method

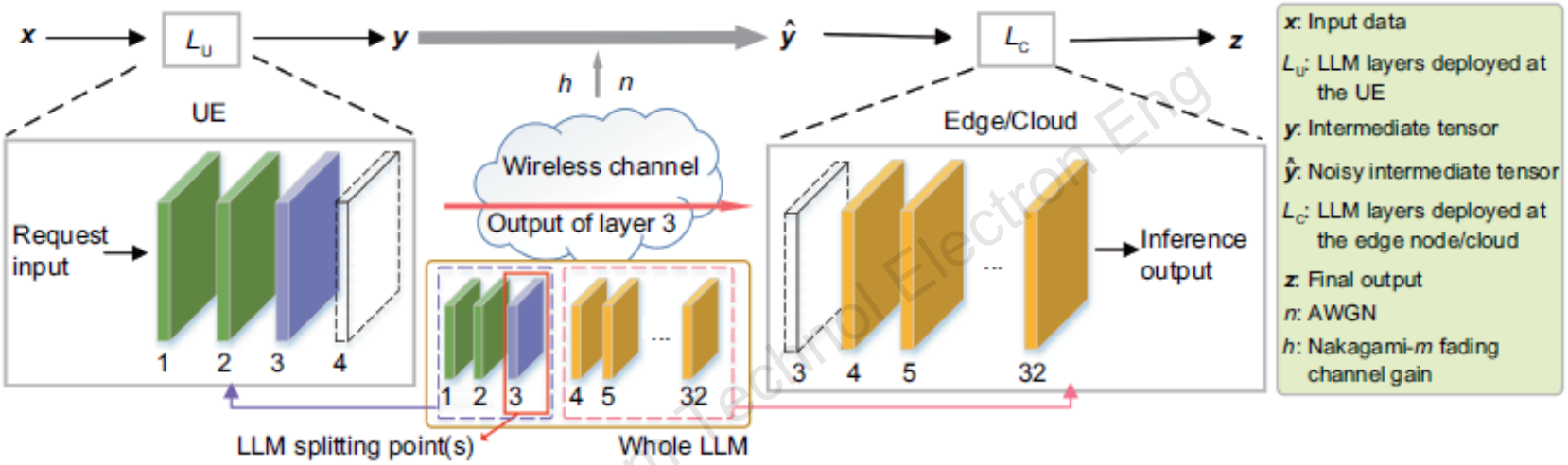

Fig. 2 Overview of the split model architecture in wireless channel, with layer 3 designated as the example splitting point. We use the 32-layer LLAMA2-7B model as an example

Optimize the splitting point p^* by minimizing the trade-off between inference performance (PPL) and computational load (C_{UE}).

$$p^* = \arg\min_p \left[\mathrm{PPL}(p;\sigma,m) + \lambda \cdot C_{\mathrm{UE}}(p)\right]$$

Method

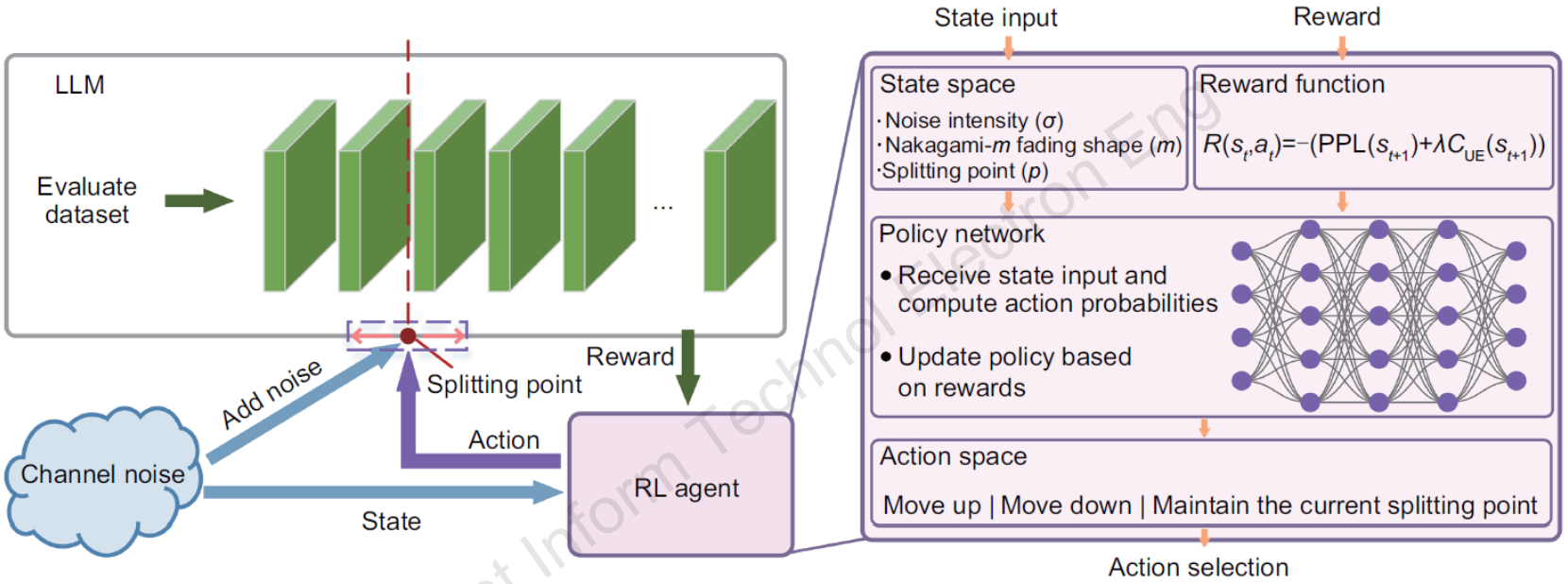

Fig. 4 Illustrations of the reinforcement learning (RL) setup, including the large language model (LLM), RL agent, and channel noise modules. The RL agent optimizes the splitting point of the LLM by receiving state inputs (noise intensity, Nakagami-m fading shape, and splitting point), computing action probabilities via the policy network, and updating the policy based on the reward function

Results

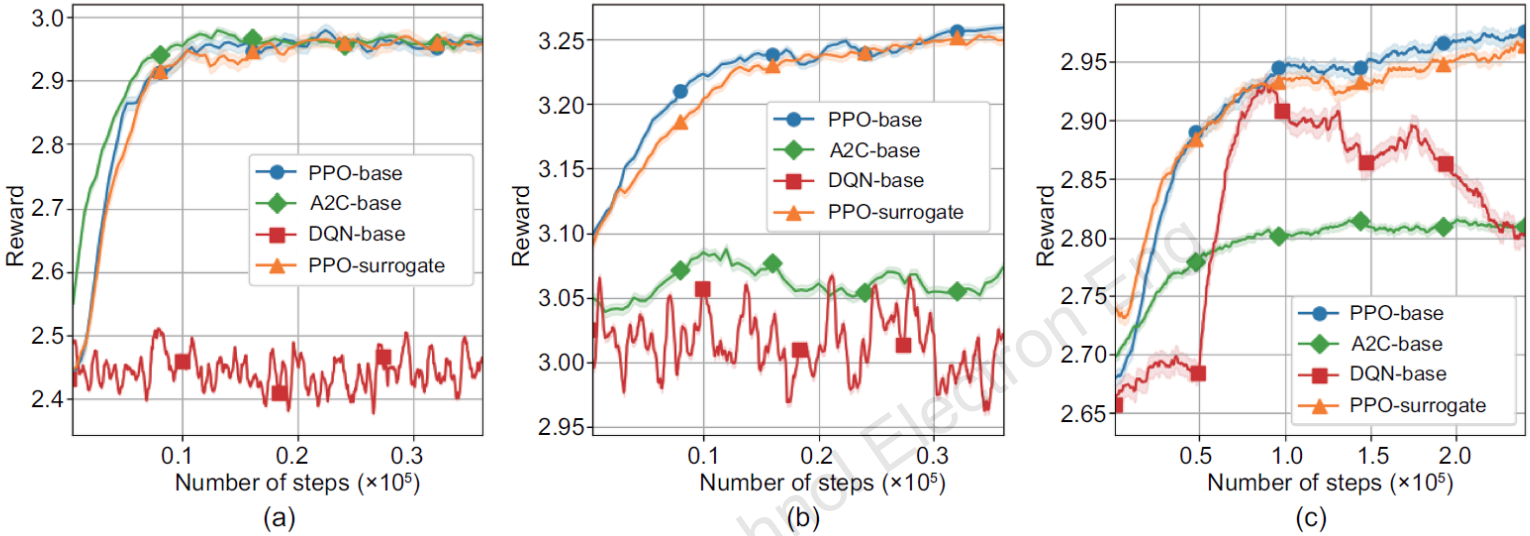

Fig. 5 Comparison of training performances for different reinforcement learning (RL) approaches under case *L* (a), case *H* (b), and case *A* (c)

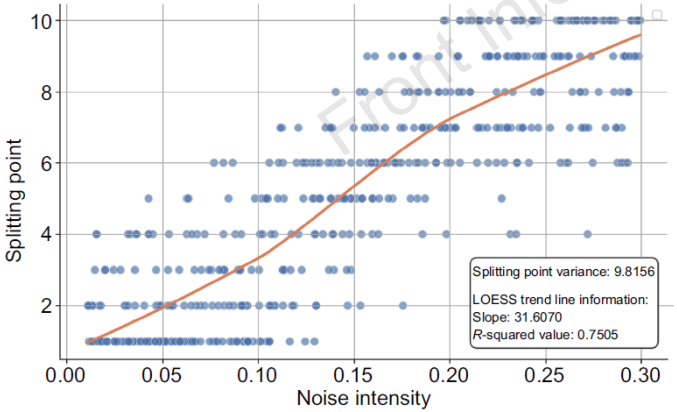

Fig. 7 Splitting points determined by the trained PPO agent across different noise intensities, along with a LOESS trend line

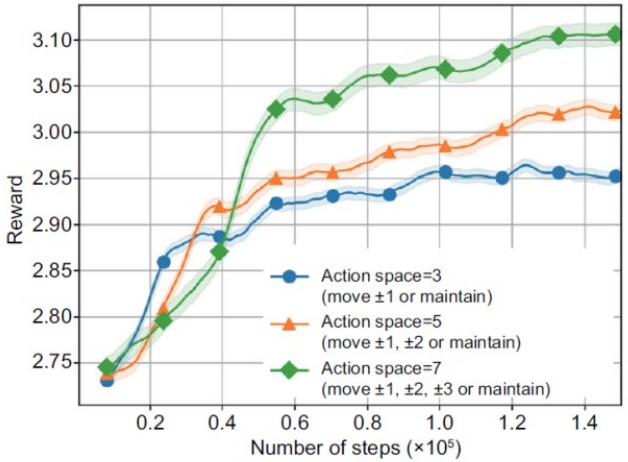

Fig. 8 Comparison of training performances across different action spaces (action space=3, 5, or 7) for varying movement ranges

Conclusions

This paper presents an MBRL-based framework for dynamically optimizing the splitting point of large language models (LLMs) across edge devices and cloud resources. The proposed method effectively balances inference performance and computational load, adapting to fluctuating network conditions. Our results demonstrate that this approach offers a robust solution for deploying LLMs in decentralized environments, improving both efficiency and flexibility.

Yuxuan CHEN is a PhD candidate in Zhejiang University, Hangzhou, China. His research interests currently focus on large language models in communication.

Rongpeng LI is currently an associate professor with the College of Information Science and Electronic Engineering, Zhejiang University. His research interest currently focuses on networked intelligence for communications evolving.

Xiaoxue YU is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Her research interests include communications in distributed learning and multiagent reinforcement learning.

Zhifeng ZHAO is currently with Zhejiang Lab, Hangzhou, as the Chief Engineering Officer, as well as Zhejiang University as an Adjunct Professor. His research areas include software-defined networks, wireless networks in 6G, computing networks, and collective intelligence.

Honggang ZHANG is a professor with the Faculty of Data Science, City University of Macau, Macau, China. His research interests include cognitive radio networks, semantic communications, green communications, machine learning, artificial intelligence, intelligent computing, and Internet of Intelligence. He is the Associate Editor-in-Chief of *China Commun*.