

WDMoE: Wireless Distributed Large Language Models with Mixture of Experts

Nan Xue, Yaping Sun, Zhiyong Chen, Meixia Tao, Xiaodong Xu, Liang Qian, Shuguang Cui, Ping Zhang

Abstract—Large Language Models (LLMs) have achieved significant success in various natural language processing tasks, but how wireless communications can support LLMs has not been extensively studied. In this paper, we propose a wireless distributed LLMs paradigm based on Mixture of Experts (MoE), named WDMoE, deploying LLMs collaboratively across edge servers of base station (BS) and mobile devices in the wireless communications system. Specifically, we decompose the MoE layer in LLMs by deploying the gating network and the preceding neural network layer at BS, while distributing the expert networks across the devices. This arrangement leverages the parallel capabilities of expert networks on distributed devices. Moreover, to overcome the instability of wireless communications, we design an expert selection policy by taking into account both the performance of the model and the end-to-end latency, which includes both transmission delay and inference delay. Evaluations conducted across various LLMs and multiple datasets demonstrate that WDMoE not only outperforms existing models, such as Llama 2 with 70 billion parameters, but also significantly reduces end-to-end latency.

Index Terms—Distributed Large Language Models, Mixture of Experts, Expert Selection, Wireless Communications

I. INTRODUCTION

The development of large language models (LLMs), exemplified by ChatGPT [1], has garnered significant attention in recent years from academia and industry. Powerful LLMs have illuminated the concept of artificial general intelligence and demonstrated emergent capabilities. In the realm of 6G technologies, such as the metaverse, which is enhanced by projects like Sora [2], autonomous driving systems supported by LLMs [3], and intelligent networking solutions like those developed by Net Master [4], large models are increasingly critical. These models are pivotal not only in advancing technological frontiers but also in creating a wide array of applications.

N. Xue, Z. Chen, M. Tao and L. Qian are with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China. (email: {nan.xue, zhiyongchen, mxtao, lqian}@sjtu.edu.cn)

Y. Sun is with the Department of Broadband Communication, Pengcheng Laboratory, Shenzhen 518000, China. Y. Sun is also with the Future Network of Intelligent Institute (FNii), the Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China (email: sunyp@pcl.ac.cn).

X. Xu and P. Zhang are with the Beijing University of Posts and Telecommunications, Beijing 100876, China, and affiliated with the Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen 518000, China (email: xuxd@pcl.ac.cn; pzhang@bupt.edu.cn).

S. Cui is with the School of Science and Engineering (SSE) and the Future Network of Intelligent Institute (FNii), the Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China. S. Cui is also affiliated with the Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen 518000, China (email: shuguangcui@cuhk.edu.cn).

This inevitably raises a question for wireless communications systems: **How can wireless communications enable LLMs?** The answer lies in achieving seamless intelligence. There are currently approaches to this: utilizing cloud-based LLMs, and device-based LLMs [5]. Most major LLMs are deployed on cloud computing clusters. Another burgeoning area of research is device-based LLMs. Researchers are actively compressing these large models to sizes manageable for mobile devices, primarily through model quantization and distillation techniques.

However, these approaches still face significant challenges. Cloud-based LLMs meet severe privacy issues, and users are heavily dependent on their connections to cloud servers. Device-based LLMs are constrained by limited computing capabilities, memory size and energy consumption. For example, mobile devices can generate up to 20 tokens per second on the advanced mobile AI computing platform Snapdragon 8 Gen 3, under conditions optimized jointly with Llama2-7B [6]. The scaling law [7] suggests that larger models tend to perform better, implying that smaller device-based LLMs may struggle to match the capabilities of their larger counterparts.

In this paper, we try to design a wireless distributed LLMs paradigm, deploying LLMs collaboratively across edge servers of base station (BS) and mobile devices, to address the aforementioned challenges. Compared to cloud deployments, edge servers are geographically closer to users, which reduce latency. Furthermore, distributing model structure across multiple locations enhances data security, making it difficult for any single location to access original user data. In addition, the wireless distributed LLMs paradigm can utilize large models with more parameters and improve the speed of data processing compared to device-based LLMs.

To efficiently distribute LLMs across edge servers and devices, we employ the Mixture of Experts (MoE) as the model design strategy, which allows for scalable and dynamic allocation of computational and communications resources. MoE is a neural network model that extends Transformer architecture by replacing its feed-forward network (FFN) layer with a gating network and multiple expert networks [8]. This configuration introduces sparsity, enabling the model to scale up in size without a corresponding increase in computational burden. As shown in Fig.1(a), the parallel relationship among experts in MoE facilitates natural robustness during the inference phase in the wireless network. Existing research [9]–[11] primarily focuses on the training phase, and the inherent characteristics of MoE are often overlooked [12]. Both [13]

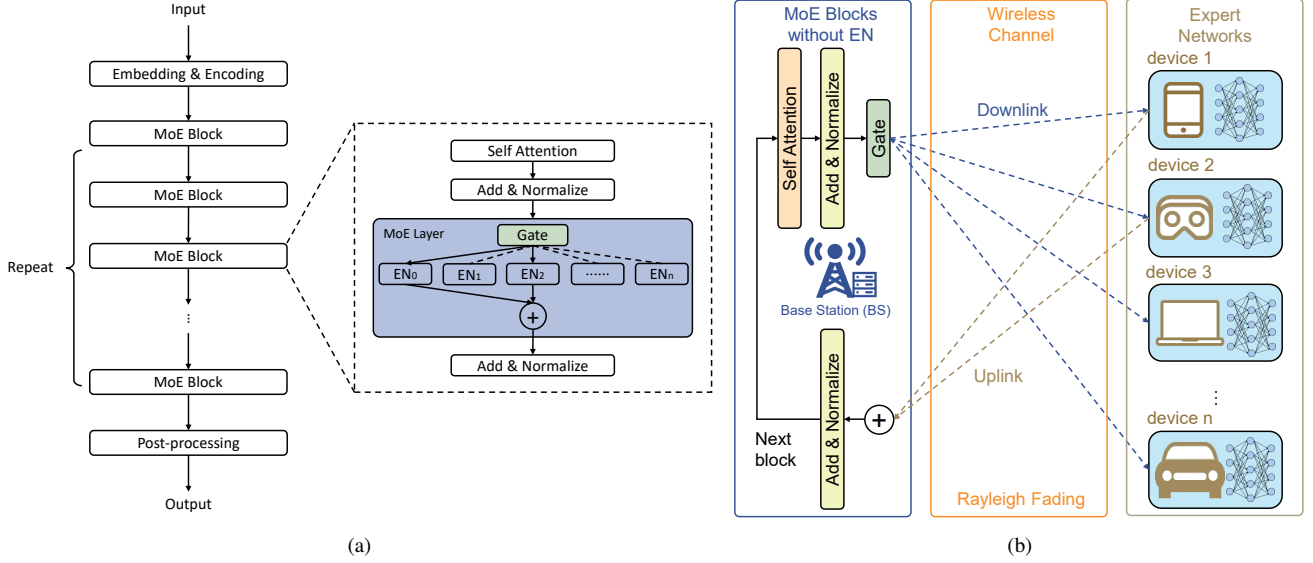


Fig. 1: (a) MoE-based LLM architecture; (b) The proposed WDMoE system model.

and [14] devise efficient routers that achieve either speed up or performance improvement. [15] develops compression techniques for expert weights and optimizes in-memory expert management by improving the compute-I/O pipeline.

To this end, we propose the wireless distributed LLMs paradigm based on MoE, named WDMoE. This paradigm includes an expert network deployment strategy, along with a scheme for dynamically adjusting expert selection according to wireless channel conditions, without the need for additional training. Specifically, we decompose the MoE layer in LLMs by placing both the gating network and the neural network layer preceding it at BS, while the expert networks are distributed across the devices. Therefore, by leveraging the parallel characteristics of the expert networks, the proposed framework avoids scenarios in which deploying multiple expert networks on a single device could lead to processing halts due to disconnections. Meanwhile, it fully utilizes the computational resources of various devices. In LLMs based on MoE, the gating network is trained to optimize processing for each token. However, it cannot guarantee that each device maintains robust communications with BS in the wireless communications system. Therefore, we design an expert selection policy by taking into account both the optimal performance of the model and the end-to-end latency. Evaluations conducted across various LLMs and multiple datasets demonstrate that the proposed WDMoE not only outperforms existing models, such as Llama 2 with 70 billion parameters, but also significantly reduces end-to-end latency, including both wireless transmission delay and inference delay.

II. THE WDMoE FRAMEWORK

A. Gating Network of MoE-based LLMs

The network structure of MoE-based LLMs is depicted in Fig.1(a), where each MoE block is based on a Transformer.

In these blocks, the FFN module is replaced by an MoE layer [16]. An MoE layer consists of a gating network, also known as a router, and multiple expert networks, each of which can be any type of neural network. The gating network, which is a compact neural network, processes the input token to produce a weight vector for each expert.

The number of expert networks is denoted as n . Let $\mathbf{x} \in \mathbf{R}^{r \times m}$ be the input token set of the gating network, where r is the number of input tokens of one prompt, and m is the dimension of a token's embedding. The element $x_j^i \in \mathbf{x}$ denotes the j -th input token of the i -th MoE block. This token x_j^i serves as the input to the gating network, which then outputs weights $\mathbf{w}_j^i$ for the j -th token in the i -th block. Here, $\mathbf{w}_j^i$ is an n -dimension column vector representing weights allocated to n experts. x_j^i is also fed into the expert networks, whose outputs $\mathbf{y}_j^i$ is a n -dimension column vector. The output of an MoE layer is the weighted sum of each expert network's output as follows:

$$o_j^i = \mathbf{w}_j^i \mathbf{y}_j^i, \quad (1)$$

where o_j^i is the output of the j -th token through the i -th MoE layer.

B. Distributed Deployment of WDMoE

Experts operate in parallel and have no impact on other experts, making this approach suitable for deployment in distributed mobile edge networks. We consider a BS with edge servers that possess powerful computing capability, enabling it to process multiple data streams simultaneously for n mobile devices, as shown in Fig.1(b). In the WDMoE framework, the attention mechanism and gating network are located at BS, while only expert networks are assigned to mobile devices. Typically, the expert network consists of a simple multilayer perceptron (MLP). When a user sends a prompt,

its embedding operation can be completed either locally or at BS, depending on the choice of the user. In this paper, we focus on the communications and computing costs incurred during interactions between BS and mobile devices following the first embedding module. The proposed framework can be integrated with existing splitting methods.

C. Communication Model

Let $\mathbf{U}$ denote the device set. For the q -th device $q \in \mathbf{U}$, the bandwidth allocated by BS is denoted by B_q . The transmission rate for the q -th device is formulated as:

$$R_q^d = B_q \log_2(1 + \frac{P_q^d g_{BS,q}^2}{N_0 B_q}), \quad (2)$$

for the downlink transmission. Similarly, for the uplink transmission, we have:

$$R_q^u = B_q \log_2(1 + \frac{P_q^u g_{q,BS}^2}{N_0 B_q}), \quad (3)$$

where $g_{BS,q}$ and $g_{q,BS}$ represent the channel gain of downlink and uplink between BS and the q -th device respectively, and N_0 denotes the noise power spectral density. P_q^d and P_q^u denote the transmission power of downlink and uplink respectively.

D. Computing Model

In this paper, we assume that each device is equipped with at least one graphic processing unit (GPU). The expert network in this paper is an FFN. The number of floating point operations required by our expert network is:

$$L^{comp} = 4m \times m_h + 2m_h \times m + \eta \times m_h + m_h, \quad (4)$$

where m_h is the hidden dimension of the FFN and η denotes the FLOPs required by the activation function in the network. Note that we assume the computing latency at BS can be ignored.

E. End to End Inference Latency

Data size is calculated by:

$$L^{comm} = \epsilon \times m, \quad (5)$$

where ϵ is a coefficient determined by the quantization precision.

The total delay of the j -th token in the i -th block processed by the expert q is given by:

$$t_{j,q}^i = \frac{L^{comm}}{R_q^d} + \frac{L^{comp}}{C_q} + \frac{L^{comm}}{R_q^u}, \quad (6)$$

where C_q is the FLOPS of the GPU on the q -th device.

The number of tokens in a prompt can vary widely, ranging from a few to several hundreds or even thousands. In practical applications, WDMoE distributes tokens among multiple expert networks. The maximum end-to-end latency across these experts becomes a more appropriate measure. The overall prompt latency results from the accumulation of latencies as

all tokens pass through all blocks. We thus have the end-to-end latency as follows:

$$t_j^i = \max_{q \in \{1,2,\dots,k\}} \{t_{j,q}^i\}, \quad (7)$$

$$t^{prompt} = \sum_j^p \sum_i^b t_j^i, \quad (8)$$

where b is the number of blocks in the model, and p is the number of tokens in a prompt. b and p depend on the specified model and user input.

III. EXPERT SELECTION POLICY

Retraining an LLM to account for wireless communication conditions poses significant challenges, particularly in ensuring the model can recognize channel conditions and adjust accordingly in distributed scenarios. Moreover, retraining the gating network of an MoE layer is not advisable due to its time-consuming and computationally intensive nature. Additionally, integrating wireless communications may inadvertently misguide the gating network, transforming the original training's single objective optimization problem into a multi-objective optimization problem, which can hinder model convergence.

Observing the sparsity of LLMs activations during inference and the occasional poor performance when transferring a trained model to out-of-distribution (OOD) datasets, we acknowledge that a decrease in the number of participating experts and a suboptimal selection of experts in the trained gating network can be tolerated. To address this dilemma, we design a training-free scheme that can modify expert selection without catastrophically impacting model performance, while also significantly reducing the latency.

The WDMoE mechanism is illustrated in Fig.2. A sequence of tokens enters the gating network, which then outputs weights for each expert in the MoE layer. Due to the frozen parameters of the router, the output weights are highly correlated with the model performance. The latency for devices can be estimated at BS, where BS acquires the channel conditions of all the devices and calculates the latency according to Eq. (6). It is possible for some devices to be situated at a long distance from BS or in areas with significant coverage shadows. To address the potential issue of extended latency, we define a weight-to-latency ratio (WLR) for a token as follows:

$$\mathbf{WLR}_j^i = \mathbf{w}_j^i \oslash \mathbf{t}_j^i, \quad (9)$$

where $\mathbf{t}_j^i$ is the vector with elements $t_{j,q}^i$, for $q = 1, 2, \dots, n$. $\mathbf{WLR}_j^i$ is a vector consisting of n experts, where $WLR_{j,q}^i$ denotes the WLR of the q -th expert. The operation $\oslash$ represents element-wise division. By jointly considering weights and latency, the WLR effectively balances them. Larger weights and lower latency contribute to higher WLR. While a large weight indicates a potential for better performance, excessive latency is intolerable for devices. Consequently, a mechanism based on WLR can filter out unsatisfactory expert networks

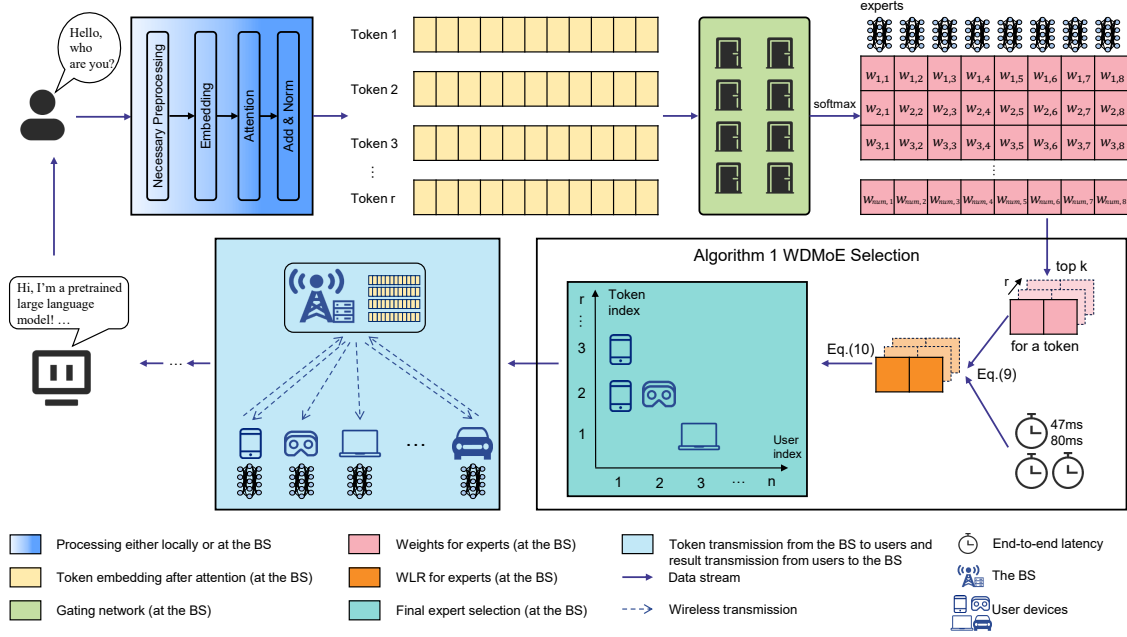


Fig. 2: The data stream and expert selection in WDMoE. When a user sends a prompt to an MoE-based LLM, it is preprocessed, embedded, and subjected to attention operations either locally or at BS, depending on user preference. Each token's embedding is then analyzed by a gating network to assign weights to each expert. WDMoE dynamically adjusts these weights and the number of experts based on gating network output and user channel conditions at the BS. The token embeddings are sent to the respective user devices for processing by expert networks. Once processed, the results are sent back to the BS, where they are weighted, summed, and then transformed from embeddings into words.

based on both model performance and latency. The proposed scheme dynamically selects any number of experts as the processing node for the current token based on WLR.

Let θ be the threshold that determines whether to drop the expert with minimal WLR. Experts are sorted from the highest weights to the lowest weights, and the top k experts are selected. Their WLRs are denoted as $\mathbf{WLR}_{j,q}^{i,k}$. Identify the minimal element $\min_{q \in \{1,2,\dots,k\}} \{WLR_{j,q}^{i,k}\}$, and the maximal element $\max_{q \in \{1,2,\dots,k\}} \{WLR_{j,q}^{i,k}\}$, within $\mathbf{WLR}_j^{i,k}$. These elements are used to calculate the decision variable κ as follows:

$$\kappa = \frac{\min_{q \in \{1,2,\dots,k\}} \{WLR_{j,q}^{i,k}\}}{\min_{q \in \{1,2,\dots,k\}} \{WLR_{j,q}^{i,k}\} + \max_{q \in \{1,2,\dots,k\}} \{WLR_{j,q}^{i,k}\}}. \quad (10)$$

When $\kappa \leq \theta$, the expert with lowest WLR will be dropped.

The threshold can be adjusted based on user requirements, with a higher threshold contributing to lower latency. It is important to note that a moderate increase in the threshold can reduce latency without a significant impact on model performance. However, excessively increasing the threshold can lead to a drastic drop in performance, despite achieving ultra-low latency. Additionally, this mechanism allows for the dynamic selection of any number of experts as needed and can simply repeat the process, as described in Algorithm 1. Dropping an expert involves assigning zero weight to

Algorithm 1 The WDMoE expert selection policy

Input: $\mathbf{w}_j^i, \mathbf{t}_j^i, \theta$

Output: $\mathbf{w}_j^i$

- 1: Calculate $WLR_{j,q}^i := \mathbf{w}_j^i \odot \mathbf{t}_j^i$
- 2: Rank experts in descending order of their weights
- 3: Take the top k experts' WLR as $\mathbf{WLR}_j^{i,k}$
- 4: Calculate κ according to Eq. (10)
- 5: **if** $\kappa < \theta$ **then**
- 6: Drop the expert with minimal WLR
- 7: Update the weights of experts
- 8: **end if**
- 9: **return** $\mathbf{w}_j^i := \text{Softmax}(\mathbf{w}_j^i)$

that expert, which means the corresponding device will not participate in processing.

IV. EXPERIMENT RESULTS

A. Experiment Settings

In the simulation, we consider an MEC-server deployed at BS and 8 devices. An expert network in each MoE block is deployed on one device, i.e. the expert network q in each MoE layer is deployed on the q -th device. The distance between the q -th device and BS is denoted as d_q . We consider Rayleigh fading channels with a mean $10^{-\frac{PL(d_q)}{20}}$, where the path loss is $PL(d_q)(dB) = 32.4 + 20\log_{10}(f^{carrier}) + 20\log_{10}(d_q)$. The

TABLE I: Performance of different models on various benchmarks.

Model	Active Params	MMLU	PIQA	Arc-e	Arc-c	Humaneval	GSM-8K	BoolQ
Llama 2 7B	7B	46.8%	78.3%	56.1%	40.3%	12.8%	16.7%	74.9%
Llama 2 13B	13B	55%	79.8%	71.8%	60.3%	18.9%	29.6%	82.4%
Llama 2 70B	70B	69.7%	82.5%	85.9%	78.3%	26.2%	63.5%	87.7%
Mistral 7B-v0.1	7B	64.1%	81.6%	83.6%	74.2%	22.6%	47.5%	84.1%
Mixtral 8x7B-Instruct-v0.1	13B	70.0%	83.2%	92.8%	84.8%	47.6%	70.9%	88.72%
WDMoE-0.1	7/13B	70.3%	83.7%	92.2%	88.1%	48.8%	71.0%	88.38%
WDMoE-0.2	7/13B	68.8%	84.1%	91.4%	86.8%	47.0%	72.0%	88.59%
WDMoE-0.3	7/13B	68.2%	83.7%	88.9%	80.7%	45.7%	68.6%	87.06%

carrier frequency $f^{carrier}$ is set as 3.5 GHz. The transmission power of BS and the device is 10 Watts and 0.2 Watts respectively. The total bandwidth is 100MHz, allocated evenly among all devices.

We apply the proposed WDMoE on Mixtral 8x7B model and refer to the modified version as **WDMoE-threshold** in the experiment results, with the specific numerical value for the threshold to be provided later. We leverage OpenCompass Platform [17] to conduct an in-depth and holistic evaluation of large language models. The benchmarks include MMLU [18], PIQA [19], ARC-Easy, ARC-Challenge [20], Humaneval [21], GSM-8K [22], BoolQ [23]. We implement the experiments on NVIDIA A40 GPUs.

B. Performance Evaluation

We compare WDMoE with current state-of-the-art models, including Llama 2 with 7B, 13B and 70B parameters, as well as Mistral (referred to as Mixtral 7B-v0.1) and Mixtral (referred to as Mixtral 8x7B-Instruct-v0.1), across various benchmarks. Besides, we evaluate the end-to-end latency of Mixtral and WDMoE in the wireless network.

Performance. The detailed results are presented in Table I. Due to the dynamics of WDMoE, the MoE's total number of active parameters is either 7B or 13B, which is lower than Llama 2 13B and Mixtral in some cases. We can see from Table I that WDMoE outperforms Llama 2 70B across all evaluated benchmarks while utilizing only around 20% parameters and is generally superior to Mixtral. WDMoE-0.1 attains the peak performance on MMLU, Arc-c, and Humaneval benchmarks. WDMoE-0.2 excels in achieving the top scores on PIQA and GSM-8K benchmarks. WDMoE-0.3 demonstrates negligible performance deterioration on diverse benchmarks. Taking the Arc-c benchmark as an example, WDMoE demonstrates significantly higher accuracy compared to Llama 2 70B and Mixtral, with scores of 88.1% vs. 78.3% and vs. 84.8%, respectively.

Latency. The end-to-end latency of Mixtral and WDMoE is evaluated in the wireless network. We deploy Mixtral and WDMoE on distributed devices, respectively and calculate the transmission latency and inference latency. The average end-to-end latency for one prompt, along with the corresponding model evaluation score results, is shown in Fig.3. The left axis represents model accuracy, and the right axis represents the end-to-end latency. The thresholds are set as 0.1, 0.2, and 0.3, respectively. As the threshold increases, the mechanism becomes more aggressive in dropping experts, and the latency

is statistically lower. WDMoE-0.1 and WDMoE-0.2 have been demonstrated to outperform the current state-of-the-art on a majority of the evaluated benchmarks. WDMoE-0.2 guarantees 1.5% gain in performance while achieving an average latency reduction of 30.21% on GSM-8K benchmark. WDMoE-0.3 has acquired a higher performance score and is $1.65\times$ faster than Mixtral on PIQA benchmark.

Within a threshold range varying from 0 to 0.2, the model's performance remains relatively stable, and the latency is significantly reduced, by approximately 25.59% on MMLU, 20.08% on PIQA, 25.33% on ARC-Easy, 16.2% on ARC-Challenge, 37.54% on Humaneval, 30.21% on GSM-8K, and 21.86% on BoolQ. When the threshold is increased to 0.3, the model's performance begins to deteriorate. Nevertheless, the slight decline in performance is offset by a notable reduction in latency, offering a more flexible choice for users. From the experiment results, we can conclude that WDMoE can significantly reduce the latency in wireless scenarios without severe performance deterioration.

V. CONCLUSION

In this paper, we propose a wireless distributed LLMs paradigm based on MoE, named WDMoE, which deploys LLMs collaboratively across edge servers at BS and mobile devices in the wireless communications system. The deployment framework leverages the parallel characteristics of expert networks. We also devise an expert selection policy that considers both the performance of the model and the end-to-end latency. Extensive experiments across various LLMs and on multiple benchmarks demonstrate that WDMoE ensures high performance and significantly reduces latency. This enables the feasibility of distributed LLMs in wireless scenarios and showcases the promising future of cooperative edge-device large models.

REFERENCES

- [1] OpenAI. (2022) Introducing chatgpt. [Online]. Available: <https://www.openai.com/blog/chatgpt/>
- [2] —. (2024) Video generation models as world simulators. [Online]. Available: <https://openai.com/research/video-generation-models-as-world-simulators>
- [3] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 910–919.
- [4] Huawei. (2024) Huawei introduces industry's first network large model. [Online]. Available: <https://e.huawei.com/en/news/2024/solutions/enterprise-network/first-network-large-model>

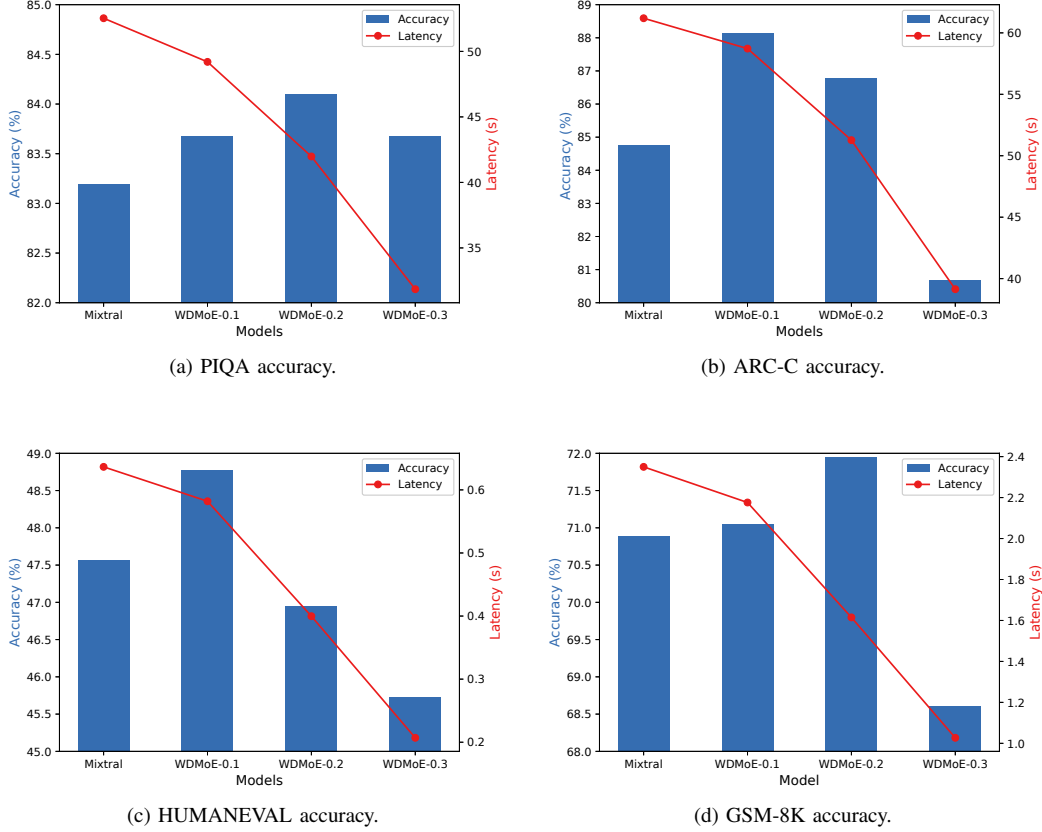


Fig. 3: Performance and latency of WDMoE on various benchmarks.

- [5] L. Li, S. Qian, J. Lu, L. Yuan, R. Wang, and Q. Xie, "Transformer-lite: High-efficiency deployment of large language models on mobile phone gpus," *arXiv preprint arXiv:2403.20041*, 2024.
- [6] Qualcomm. (2023) Snapdragon 8 gen 3 mobile platform product brief. [Online]. Available: <https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform>
- [7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
- [8] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006.16668*, 2020.
- [9] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint arXiv:1812.00564*, 2018.
- [10] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8485–8493.
- [11] W. Wu, M. Li, K. Qu, C. Zhou, X. Shen, W. Zhuang, X. Li, and W. Shi, "Split learning over wireless networks: Parallel design and resource management," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 1051–1066, 2023.
- [12] C. Hu and B. Li, "Distributed inference with deep learning models across heterogeneous edge devices," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 330–339.
- [13] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon *et al.*, "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.
- [14] Z. Zeng, Q. Guo, Z. Fei, Z. Yin, Y. Zhou, L. Li, T. Sun, H. Yan, D. Lin, and X. Qiu, "Turn waste into worth: Rectifying top- k router of moe," *arXiv preprint arXiv:2402.12399*, 2024.
- [15] R. Yi, L. Guo, S. Wei, A. Zhou, S. Wang, and M. Xu, "Edgemoe: Fast on-device inference of moe-based large language models," *arXiv preprint arXiv:2308.14352*, 2023.
- [16] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.
- [17] O. Contributors, "Opencompass: A universal evaluation platform for foundation models," <https://github.com/open-compass/opencompass>, 2023.
- [18] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.
- [19] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.
- [20] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.
- [21] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.
- [22] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.
- [23] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *arXiv preprint arXiv:1905.10044*, 2019.