ICML
International Conference
On Machine Learning

Bridging On-Device and Cloud LLMs for Collaborative Reasoning

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Christopher G. Brinton

Motivation I: On-Device LLM

- Challenges of Cloud-based LLMs
  - **Pressure on Cloud**: substantial computational pressure on cloud infrastructure
  - **High Latency**: Real-time applications suffer from delays due to network dependency
  - **Underused Device Power**: The potential of device is underexplored
- Advantages of on-device LLMs
  - Offline Capability
  - Better Personalization
  - Enhancing Privacy

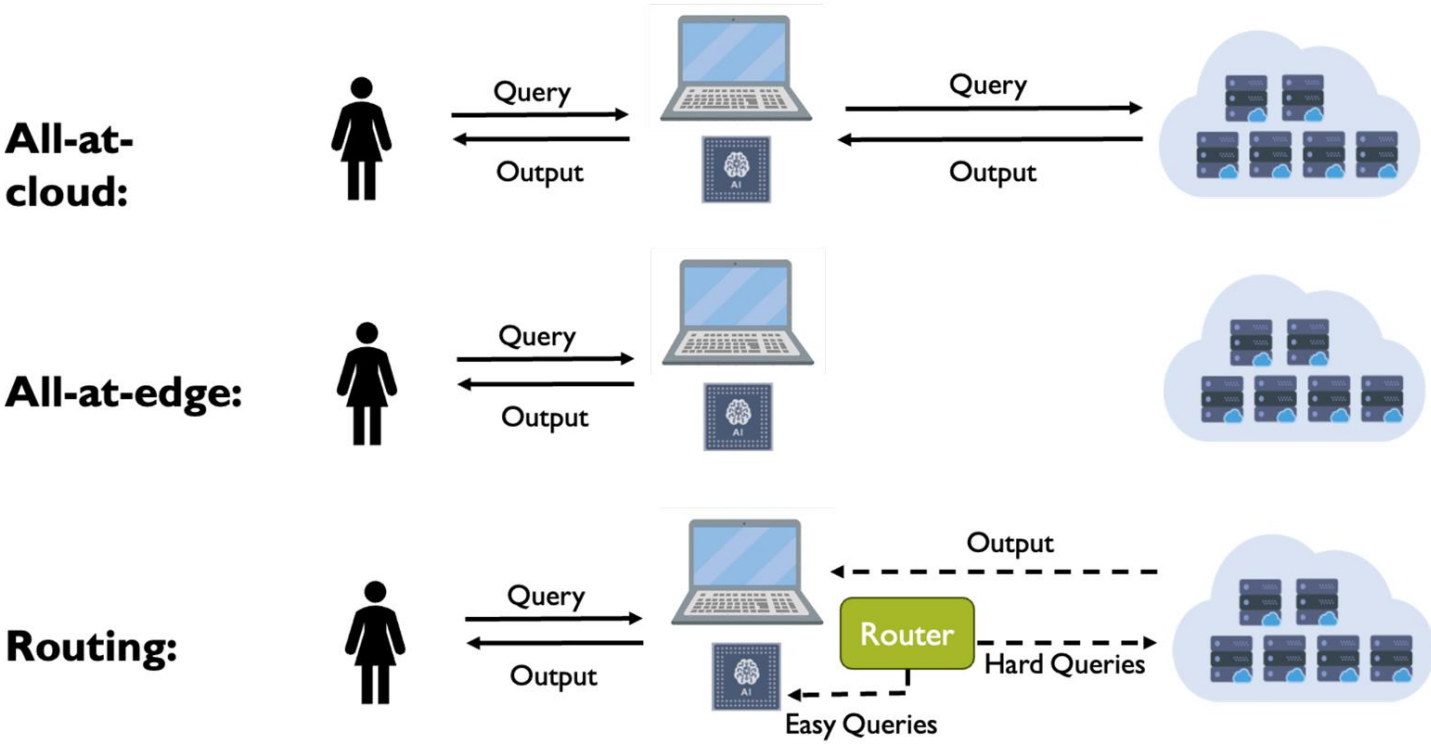
Motivation II: Performance Gap

- The trade-off
  - On-device LLMs
    - Offer privacy, low latency, and device customization
    - But **suffer from limited capacity and weaker performance**
  - Cloud LLMs
    - Deliver **superior accuracy and reasoning power**
    - But introduce latency, privacy risks, and network dependency

- The Key Challenge
  - How to **balance efficiency and accuracy** by leveraging **both on-device and cloud LLMs**

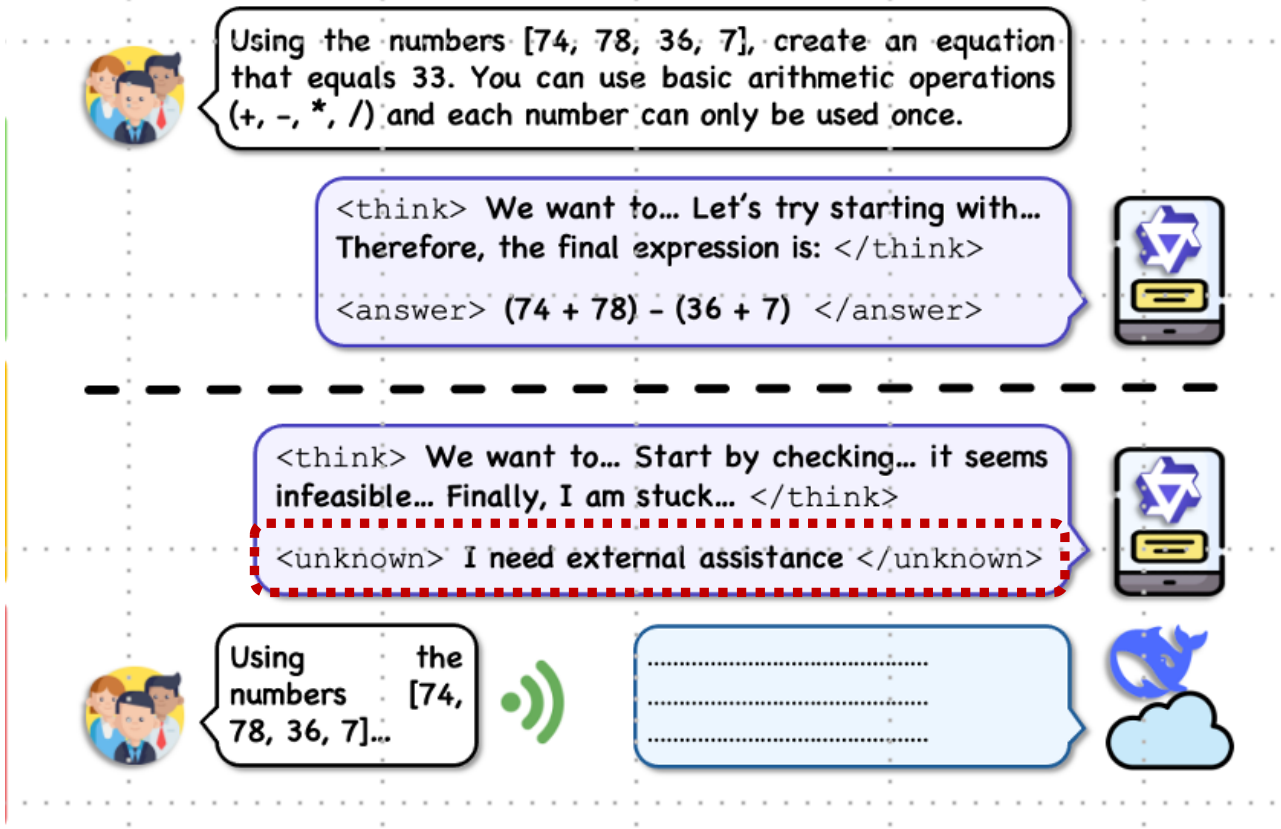
Existing Pipeline

- Task Optimization
  - ➢Prompt tuning
  - ➢SFT
  - ➢RLHF
  - ➢Etc..
- Router optimization
  - ➢Binary classifier



Unified Problem Formulation

- **Key Idea:** The **on-device model** autonomously decide whether to solve the problem independently or **invoke the cloud model**



Two possible cases:

$$y = \begin{cases} y^{\theta}, & \text{(local only)} \\ [y^{\theta}, y^{c}], & \text{(collaborative)} \end{cases}$$

Unified Problem Formulation

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[R(\boldsymbol{\theta},\boldsymbol{x})] := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{E}_{\boldsymbol{y}^\theta\sim\pi_\theta(\boldsymbol{x})}[r(\boldsymbol{x},\boldsymbol{y})] \\ \text{s.t.} \quad & \mathbb{E}\left[\mathbf{1}\left\{\boldsymbol{y}\sim(\pi_\theta,\pi_c)\right\}\right] \le \rho\mathbb{E}\left[\mathbf{1}\left\{\boldsymbol{y}\sim\pi_\theta\right\}\right]. \end{aligned}$$

Lagrangian Relaxation

Unconstrained Optimization

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{E}_{\boldsymbol{y}^\theta\sim\pi_\theta(\boldsymbol{x})}[r(\boldsymbol{x},\boldsymbol{y})] - \gamma\left(\mathbb{E}\left[\mathbf{1}\left\{\boldsymbol{y}\sim(\pi_\theta,\pi_c)\right\}\right] - \frac{\rho}{1+\rho}|\mathcal{D}|\right)$$

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{y}^\theta\sim\pi_\theta(\boldsymbol{x})}\left[r(\boldsymbol{x},\boldsymbol{y}) - \gamma\mathbf{1}\left\{\boldsymbol{y}\sim(\pi_\theta,\pi_c)\right\}\right]\right]$$

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{y}^\theta\sim\pi_\theta(\boldsymbol{x})}\left[\tilde{r}(\boldsymbol{x},\boldsymbol{y})\right]\right] \quad \tilde{r}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} r(\boldsymbol{x},\boldsymbol{y}) - \gamma, & \text{if } \boldsymbol{y}\sim(\pi_\theta,\pi_c), \\ r(\boldsymbol{x},\boldsymbol{y}), & \text{otherwise.} \end{cases}$$

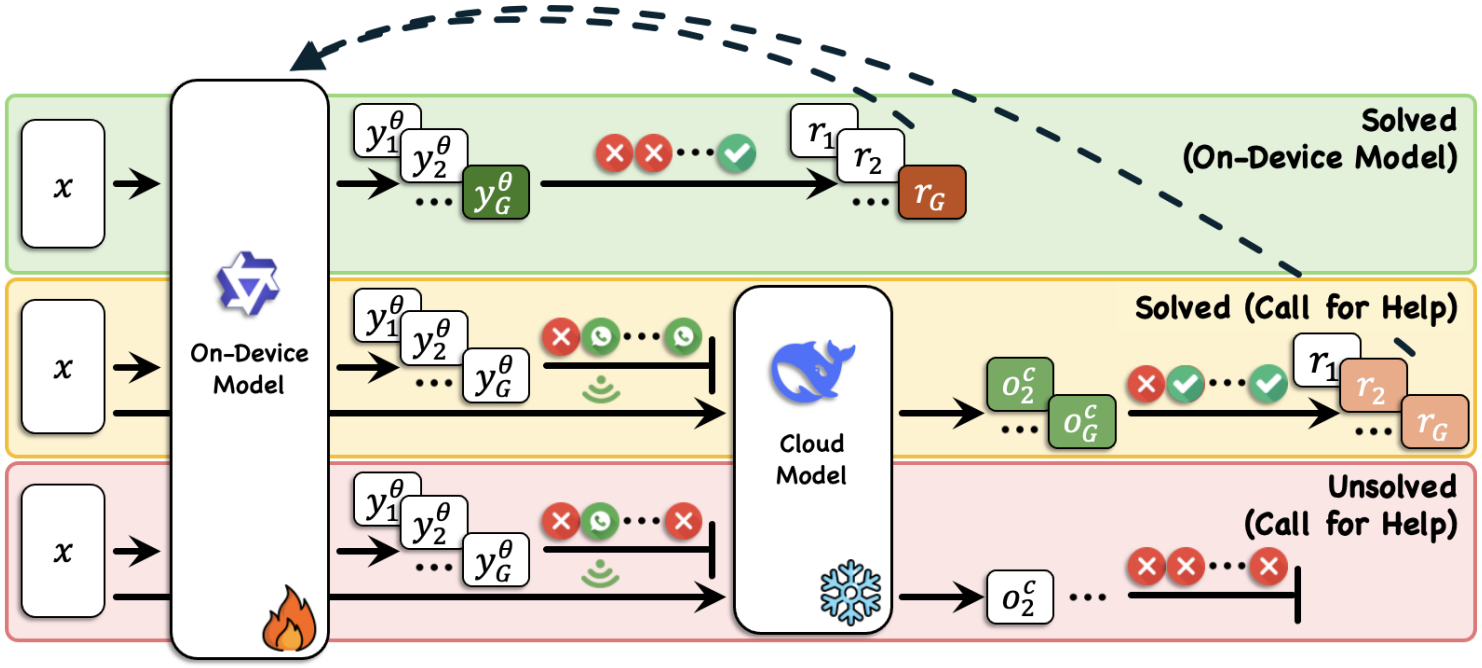
Unified Problem Formulation

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{y}^{\theta}\sim\pi_{\theta}(\boldsymbol{x})}\left[\tilde{\boldsymbol{r}}(\boldsymbol{x},\boldsymbol{y})\right]\right] \quad \tilde{r}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} r(\boldsymbol{x},\boldsymbol{y}) - \gamma, & \text{if } \boldsymbol{y} \sim (\pi_{\theta}, \pi_{c}), \\ r(\boldsymbol{x},\boldsymbol{y}), & \text{otherwise.} \end{cases}$$

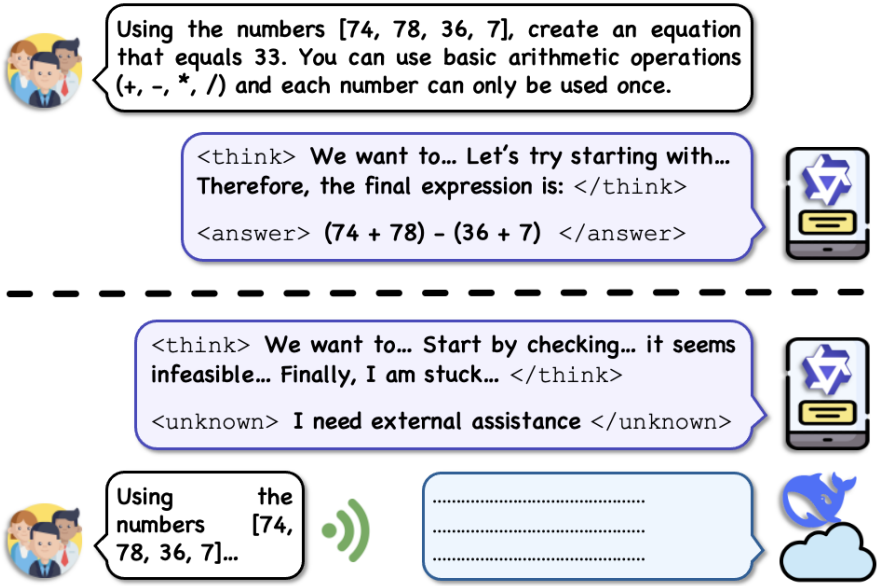
- Two possible case $y = \begin{cases} y^{\theta}, & \text{(local only)} \\ [y^{\theta}, y^{c}], & \text{(collaborative)} \end{cases}$
- **Accuracy Reward (the largest):** Rewards the correctness of the on-device model's final answer
- **Coordination Reward:** Rewards effective delegation, granted when the device invokes the cloud model and the joint output is correct

Collaborative Unified Training Framework

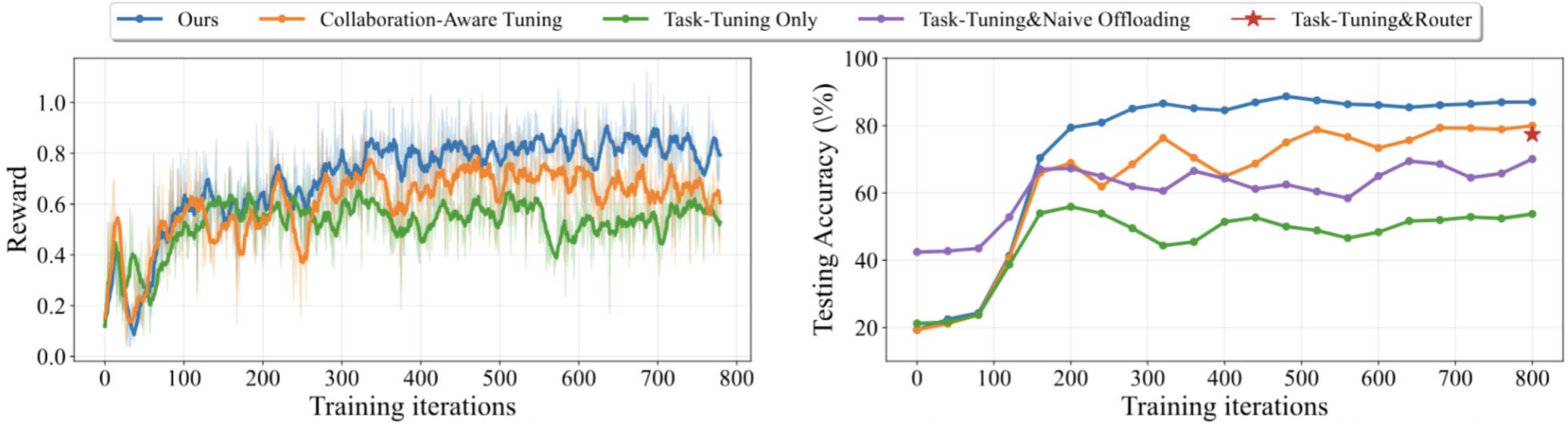
- Our methodology



(a) Our Training Methodology



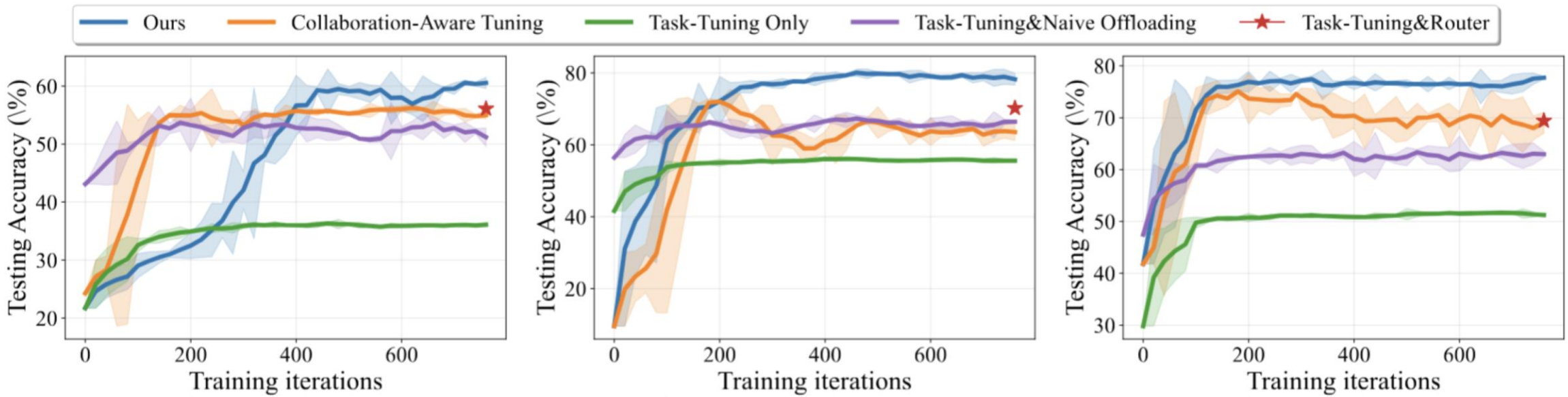
(b) Collaborative Inference



(a) Reward versus training iterations

(b) Testing Accuracy versus training iterations

Qwen2.5-3B-Instruct



(a) Llama-3.2-1B-Instruct

(b) Qwen2.5-1.5B-Instruct

(c) Llama-3.2-3B-Instruct

Model	Metric	Method	MATH-lighteval	MATH-500	AMC23	MinervaMath	AGI-Eval-Math	Avg.
		Cloud LLM	98.4	97.3	97.5	80.9	94.7	93.8
Qwen2.5-1.5B	Cond. Acc.	Task-Tuning Only	56.1	54.8	35.0	20.6	51.8	43.7
		Task-Tuning&Naive Offloading	56.3	55.4	32.1	21.0	51.6	43.3
		Collaboration-Aware Tuning	54.1	61.2	28.6	13.6	55.3	42.6
		Task-Tuning&Router	61.8	64.9	39.3	21.8	58.7	49.3
		Ours	**72.6**	**75.1**	**42.9**	**24.1**	**64.6**	**55.8**
	Overall Acc.	Task-Tuning Only	56.1	54.8	35.0	20.6	51.8	43.7
		Task-Tuning&Naive Offloading	67.2	67.4	50.0	38.2	68.3	58.2
		Collaboration-Aware Tuning	61.5	61.2	42.5	33.5	66.9	53.1
		Task-Tuning&Router	70.9	72.2	55.0	36.8	69.2	60.8
		Ours	**80.4**	**81.6**	**57.5**	**40.8**	**73.4**	**66.7**
Llama-3.2-3B	Cond. Acc.	Task-Tuning Only	51.2	43.0	27.5	19.1	45.5	37.3
		Task-Tuning&Naive Offloading	51.6	41.8	27.5	19.4	45.7	37.2
		Collaboration-Aware Tuning	60.2	43.7	21.4	13.4	43.9	36.5
		Task-Tuning&Router	64.9	45.7	25.0	20.2	35.1	38.2
		Ours	**72.2**	**56.6**	**35.7**	**27.7**	**59.7**	**50.4**
	Overall Acc.	Task-Tuning Only	51.2	43.0	27.5	19.1	45.5	37.3
		Task-Tuning&Naive Offloading	65.1	59.0	45.0	37.1	58.7	53.0
		Collaboration-Aware Tuning	66.8	59.6	42.5	36.8	58.9	52.9
		Task-Tuning&Router	69.4	62.0	46.0	39.1	53.8	54.1
		Ours	**79.5**	**68.6**	**52.5**	**43.4**	**64.5**	**61.7**

The proposed method generalize well to the out-of-distribution data