KAIST

NS2 Network and System Security Laboratory

KAIST WSP Lab
KAIST Web Security & Privacy Lab

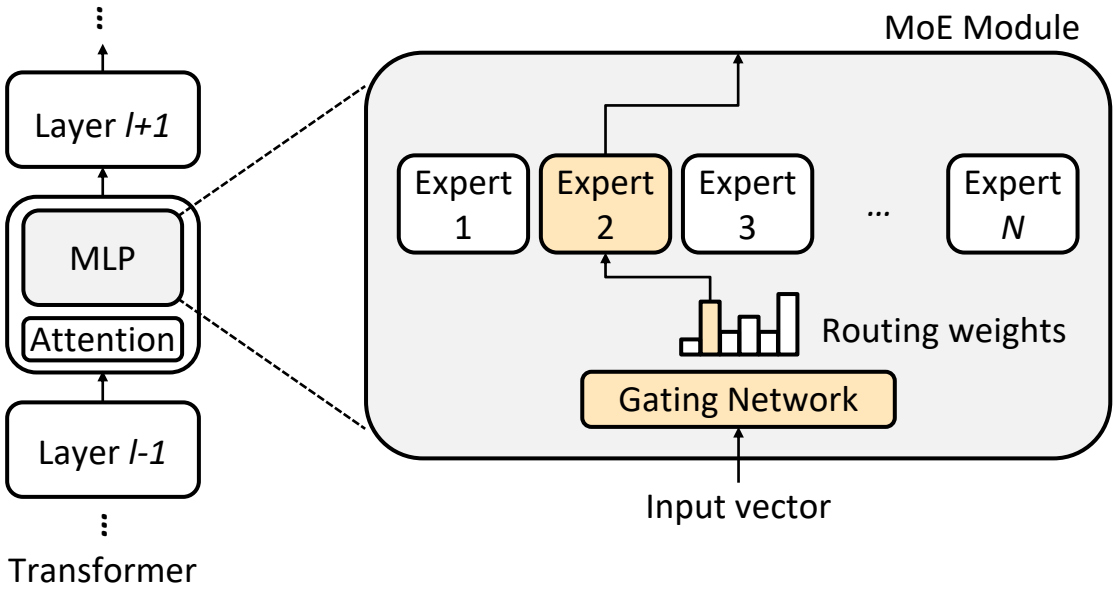
***SafeMoE*: Safe Fine-Tuning for MoE LLMs by Aligning Harmful Input Routing**

JAEHAN KIM, MINKYOO SONG, SEUNGWON SHIN, SOOEL SON

KAIST

Mixture-of-Experts (MoE)

- A highly scalable sparse neural network architecture
  - Activates a subset of *experts* by dynamically routing inputs using *gating networks*
  - Widely adopted in modern LLMs, outperforming their dense counterparts



Gemini 3

OpenAI GPT-oss

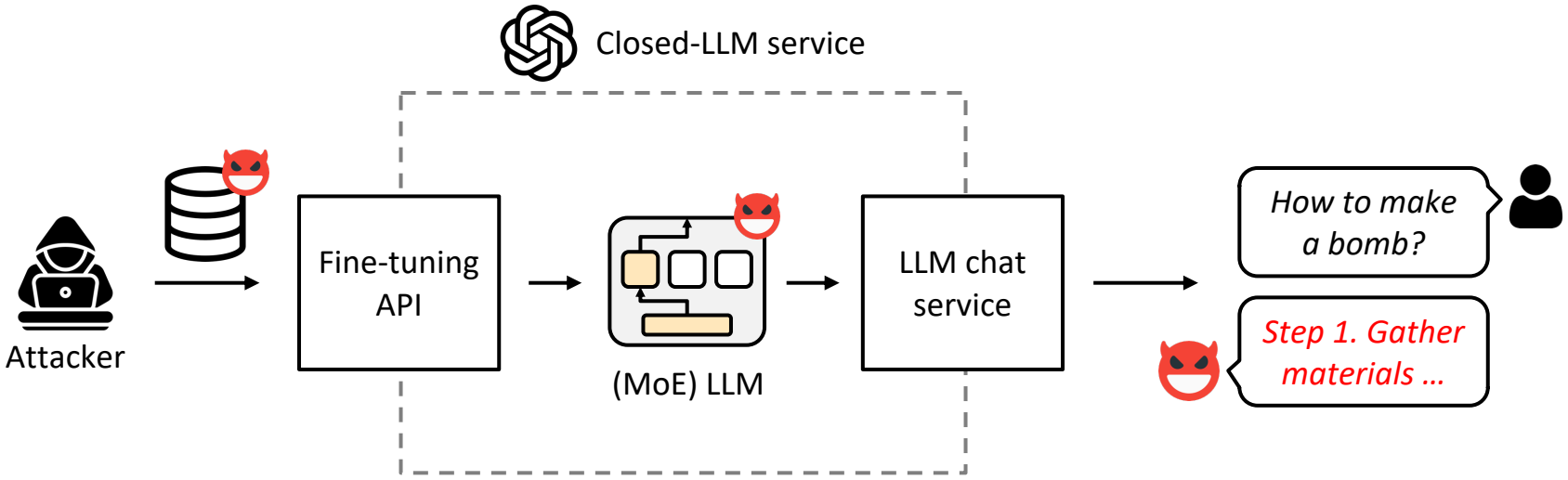
LLAMA 4

deepseek

Qwen3

Practical Attack Scenarios on LLM Services

- Harmful fine-tuning (HFT) attacks
  - Aims to elicit *unsafe* or *unethical* model responses
  - Injects a small fraction of harmful samples into the training data
  - Exploits fine-tuning APIs to compromise internal LLMs



Existing HFT Defenses

- **Alignment stage** Aligning models with perturbations
  - Enhances robustness against subsequent HFT attacks
    - Vaccine [NeurIPS'24], RepNoise [NeurIPS'24], Booster [ICLR'25], VAA [ICML'25]
- **FT stage** Directly rectifying harmful directions
  - Constrains embedding vectors to be safe during fine-tuning
    - SafeInstr [ICLR'24], Lisa [NeurIPS'24], SAFT [NeurIPS Workshop'24], SEAL [ICLR'25], SaLoRA [ICLR'25]
- **Post-FT stage** Pruning harmful parameters
  - Restores safety after HFT without additional training
    - RESTA [ACL'24], SafeLoRA [NeurIPS'24], Antidote [ICML'25], SafeDelta [ICLML'25]

No HFT defenses are tailored to MoE LLMs!

Our Approach: *SafeMoE*

The first safe fine-tuning method for MoE

Vulnerability Analysis

Systematically analyzes the safety mechanisms inherent to native MoE LLMs

Defense Design

Designs an HFT defense that directly mitigates architectural vulnerabilities in MoE

OpenAI GPT-oss

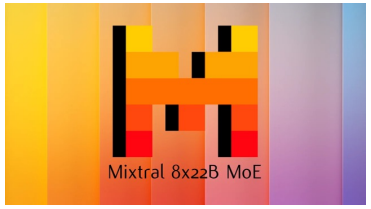
OLMoE

Qwen3

Qwen1.5-MoE

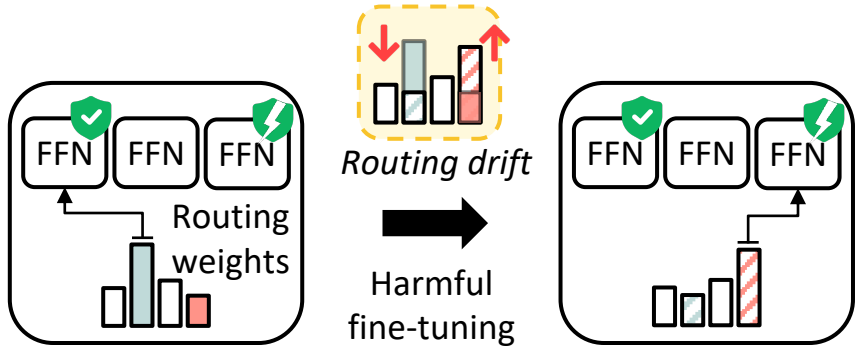
DeepSeek V2

LLAMA 4



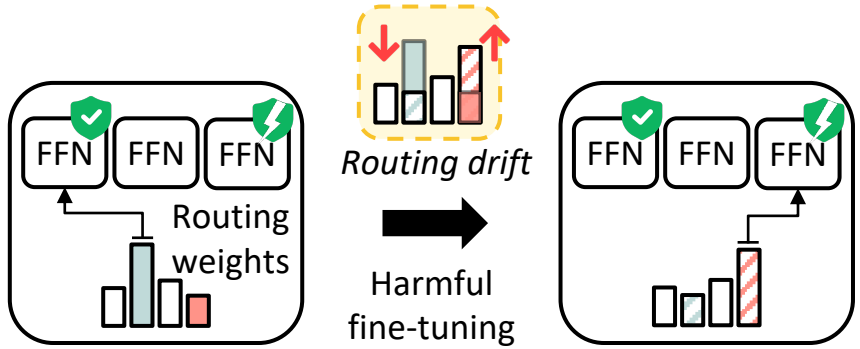
Vulnerability in MoE's Safety Mechanisms

- *Safety Routing Drift*
  - Deviations in routing weights for harmful inputs
  - Suppressing the activation of safety-critical experts



Vulnerability in MoE's Safety Mechanisms

- *Safety Routing Drift*
  - Deviations in routing weights for harmful inputs
  - Suppressing the activation of safety-critical experts



- Definition
  - KL divergence between initial safety-aligned and fine-tuned models' routing weights

Safety-aligned model's
routing weights

$$D_{KL}\left(\sigma\left(\Gamma\left(x|\boldsymbol{w}_{align}\right)\right) \;||\; \sigma\left(\Gamma(x|\boldsymbol{w})\right)\right)$$

Fine-tuned model's
Routing weights

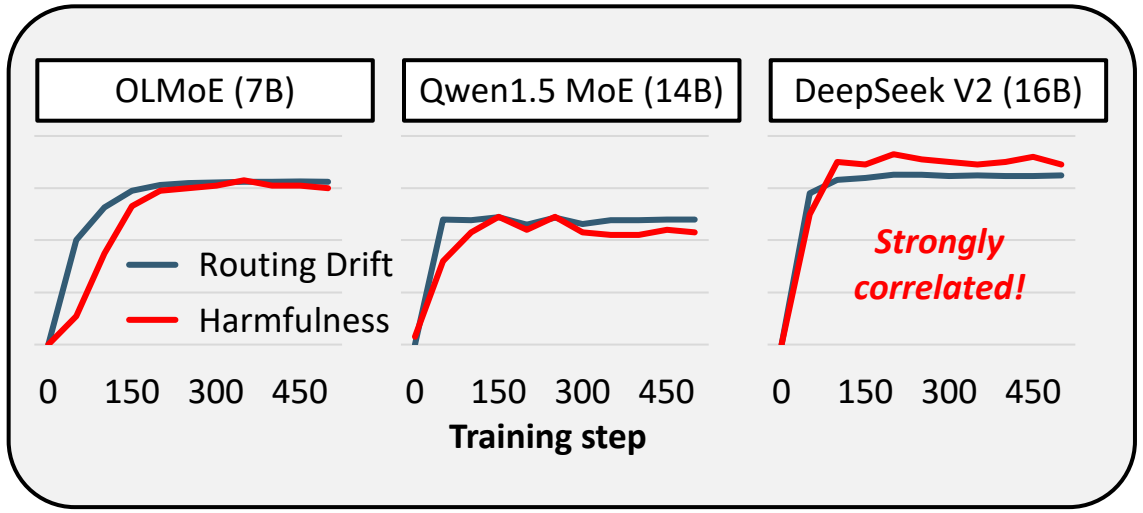
Analysis of Safety Routing Drift under HFT

Analysis 1:
Variations in safety routing drift and harmfulness during harmful fine-tuning

Analysis 2:
Fine-tuned model's harmfulness when assigning the initial model's routing weights

Analysis of Safety Routing Drift under HFT

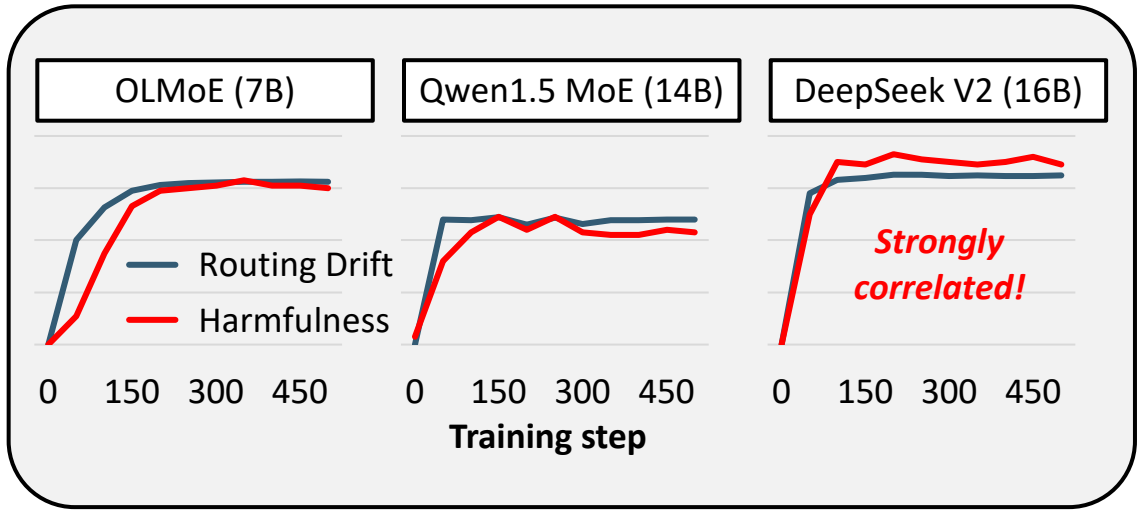
Analysis 1:
Variations in safety routing drift and harmfulness during harmful fine-tuning



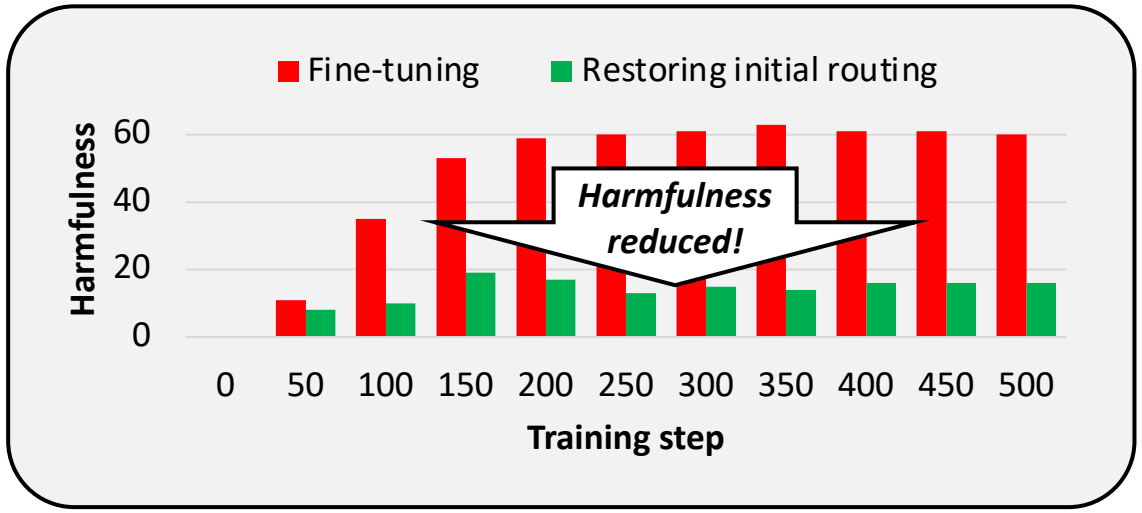
Analysis 2:
Fine-tuned model's harmfulness when assigning the initial model's routing weights

Analysis of Safety Routing Drift under HFT

Analysis 1:
Variations in safety routing drift and harmfulness during harmful fine-tuning

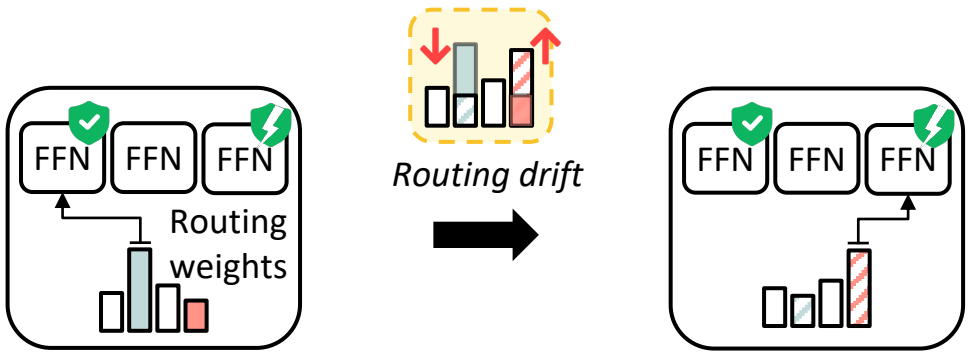


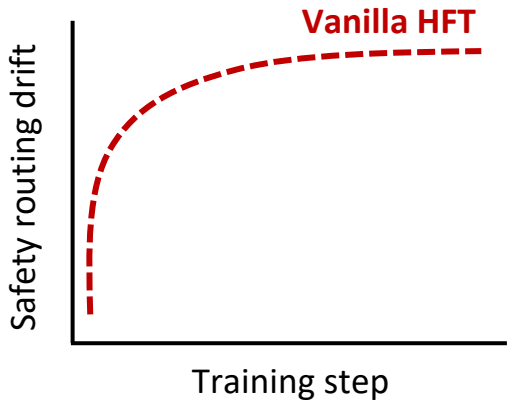
Analysis 2:
Fine-tuned model's harmfulness when assigning the initial model's routing weights



SafeMoE: Safe Fine-Tuning for MoE

- Safety routing drift regularization
  - ***Goal*** Aims to preserve initial safety-aligned routing for harmful inputs
  - ***How?*** Penalizes safety routing drift during fine-tuning





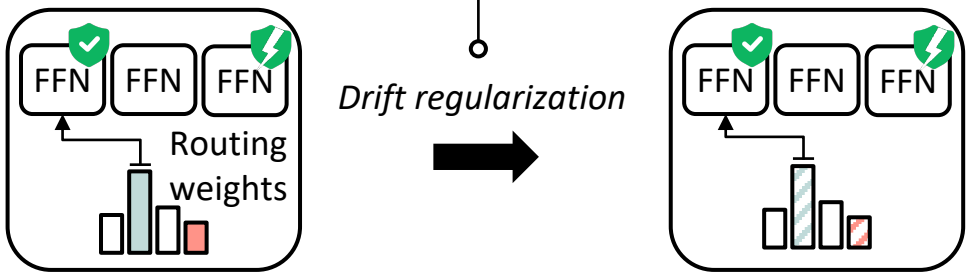
SafeMoE: Safe Fine-Tuning for MoE

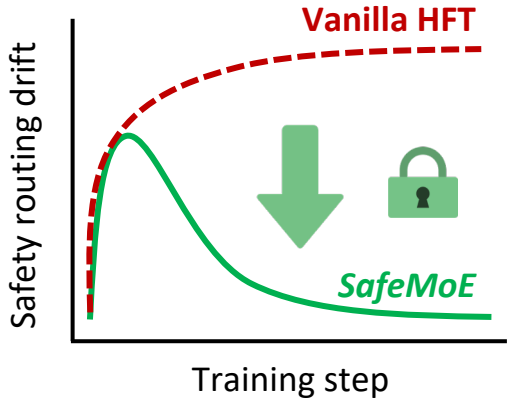
- Safety routing drift regularization
  - ***Goal*** Aims to preserve initial safety-aligned routing for harmful inputs
  - ***How?*** Penalizes safety routing drift during fine-tuning

Safety-aligned model's routing weights

$$\mathcal{L}_{reg}(\boldsymbol{w}) = D_{KL}\left(\sigma\left(\Gamma\left(x|\boldsymbol{w}_{align}\right)\right) \;||\; \sigma\left(\Gamma(x|\boldsymbol{w})\right)\right)$$

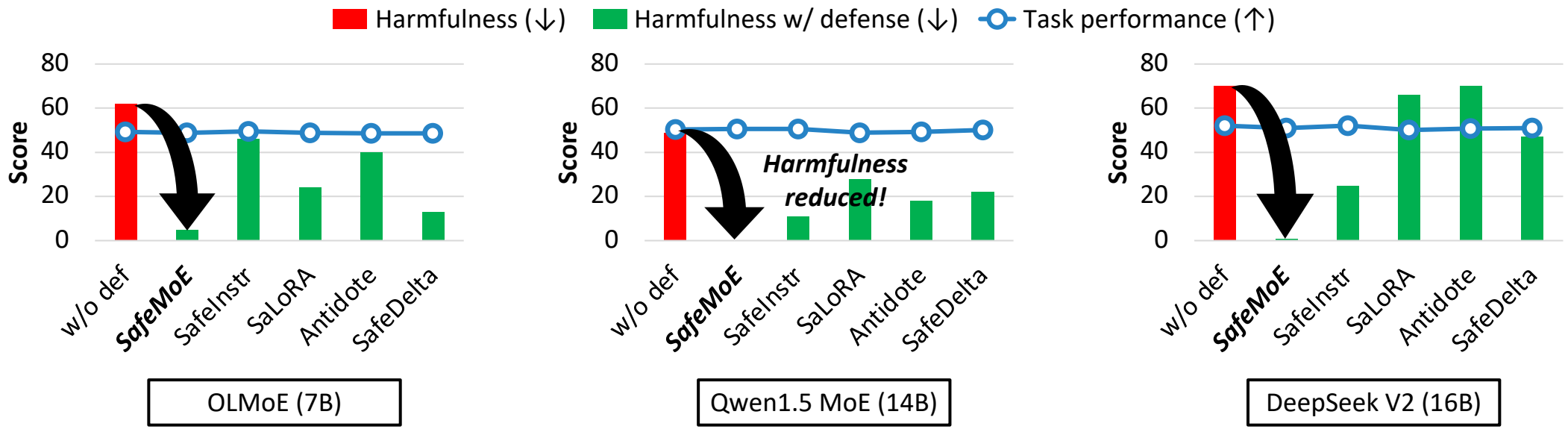
Fine-tuned model's Routing weights





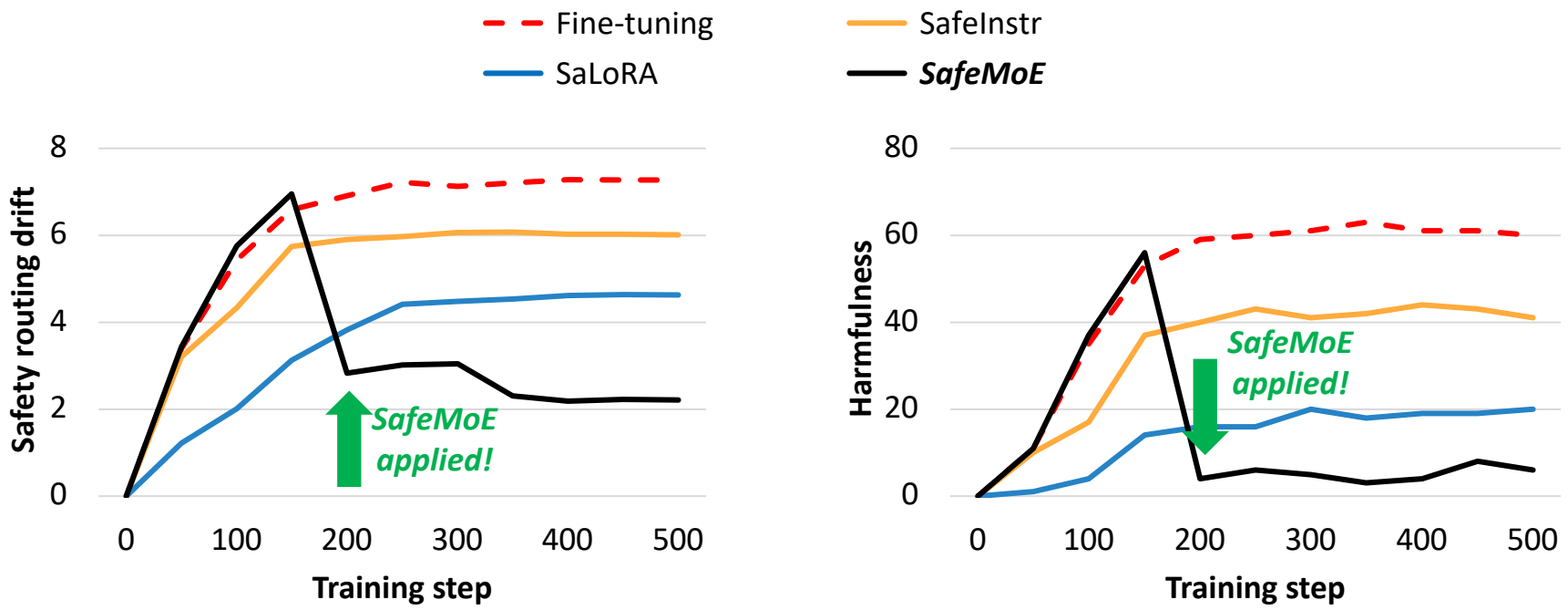
Defense Performance against HFT Attacks

Target size	Type	Fine-tuning data
7-16B	Practical HFT attack	5k task samples + 500 harmful samples
20-140B	Strong HFT attack	500 harmful samples



- Effectively mitigates HFT attacks across diverse MoE LLMs, outperforming SOTA defenses

Safety Routing Drift under Defense



- Significantly prevents safety routing drift during fine-tuning where baselines fail
- Empirically demonstrates the validity of our safety drift regularization method

Thank You !

Paper:

Contact: jaehan@kaist.ac.kr