Zeroth-Order Optimization in LLM Fine-Tuning

Shaocong Ma

November 27, 2025

University of Maryland, College Park

The “Memory Wall” Challenge in LLM Fine-tuning

Motivation:

- Full-parameter fine-tuning requires storing the **computation graph** and **gradients** for backpropagation.
- A single GPU cannot handle backpropagation for large models.

Table 1: VRAM Requirements and GPU Configuration

Model Size	First-Order (Full FT)	Est. GPU Setup
OPT-1.3B	$\approx$ 27 GB	1 $\times$ A100
OPT-6.7B	$\approx$ 156 GB	2 $\times$ A100
OPT-13B	$\approx$ 356 GB	4 $\times$ A100
OPT-30B	$\approx$ 633 GB	8 $\times$ A100

Solution:

- Fine-tune LLMs without Backpropagation—**Zeroth-Order Optimization (ZOO)**.

Principles of Zeroth-Order Optimization (ZOO)

Consider the optimization problem

$$\min_{\theta \in \mathbb{R}^d} f(\theta).$$

Core Formula (Two-Point Estimator):

$$\nabla f(x) \approx \hat{\nabla} f(x) = \frac{f(x + \mu v) - f(x)}{\mu} v$$

Notation:

- $f(x)$: The loss function.
- v: A random perturbation vector (e.g., drawn from a Gaussian distribution $\mathcal{N}(0, I_d)$).
- μ: The perturbation stepsize.

Advantages and Challenges of ZOO

Core Advantage:

- Requires only the **Forward Pass**.
- Memory footprint is comparable to **Inference** only.

Advantages and Challenges of ZOO

Core Advantage:

- Requires only the **Forward Pass**.
- Memory footprint is comparable to **Inference** only.

Key Challenges (Focus of this Report):

1. **High Variance:** Gradient estimates are volatile/noisy.
2. **Biased:** Finite difference methods rely on approximations.

Advantages and Challenges of ZOO

Core Advantage:

- Requires only the **Forward Pass**.
- Memory footprint is comparable to **Inference** only.

Key Challenges (Focus of this Report):

1. **High Variance:** Gradient estimates are volatile/noisy.
2. **Biased:** Finite difference methods rely on approximations.

Roadmap: Two theoretical works improving ZOO

- Variance Reduction: Directionally Aligned Perturbation (DAP).
- Bias Elimination: A Unbiased Estimator Family.

Revisiting Zeroth-Order Optimization: Minimum-Variance Two-Point Estimators and Directionally Aligned Perturbations

Shaocong Ma, Heng Huang.

University of Maryland, College Park

ICLR 2025 Spotlight

Variance Reduction: Directionally Aligned Perturbation (DAP)

Recap: Zero-Order Gradient Estimator

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x + \mu v) - f(x)}{\mu} v$$

Variance Reduction: Directionally Aligned Perturbation (DAP)

Recap: Zero-Order Gradient Estimator

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x+\mu v) - f(x)}{\mu} v$$

Goal: Minimize Estimation Error

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V} \left\| \frac{f(x+\mu v) - f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Variance Reduction: Directionally Aligned Perturbation (DAP)

Recap: Zero-Order Gradient Estimator

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x+\mu v) - f(x)}{\mu} v$$

Goal: Minimize Estimation Error

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V} \left\| \frac{f(x+\mu v) - f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Optimization Challenges:

- **Functional space:** Optimization is taken over all probability distributions.
- **Constraints with an empty interior:** The empty interior precludes the use of Interior Point Methods.

DAPs: Directionally Aligned Perturbation

We analytically solve this functional optimization problem.

- Constant Magnitude Perturbations:
  - $\mathbb{E}_{v\sim V}[vv^\top] = I_d$.
  - $\|v\|$ is fixed (Minimum Variance).
- Directionally Aligned Perturbations (DAPs):
  - $\mathbb{E}_{v\sim V}[vv^\top] = I_d$.
  - $\nabla f(x)^\top v$ is fixed (Minimum Variance).

$\Longrightarrow$ Both estimators achieve the minimum variance, but DAPs have some nice properties.

Traditional Methods Cannot Identify the Important Directions

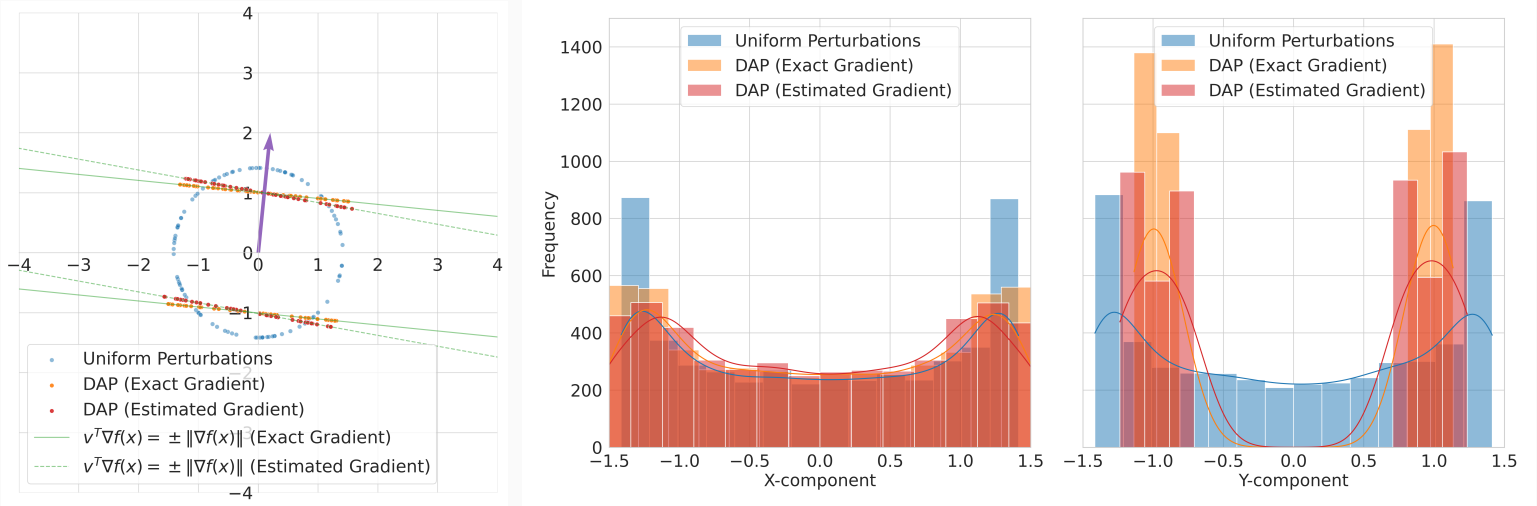


Figure 1: Illustration of the *directional alignment property* of DAP in $d = 2$ with estimating the gradient of $f(x) = x_1^2 + x_2^2$ at $x = \begin{bmatrix} 0.1 & 1 \end{bmatrix}^\top$. Traditional estimator is symmetric, but we need a non-symmetric estimator.

Traditional Methods Cannot Identify the Important Directions

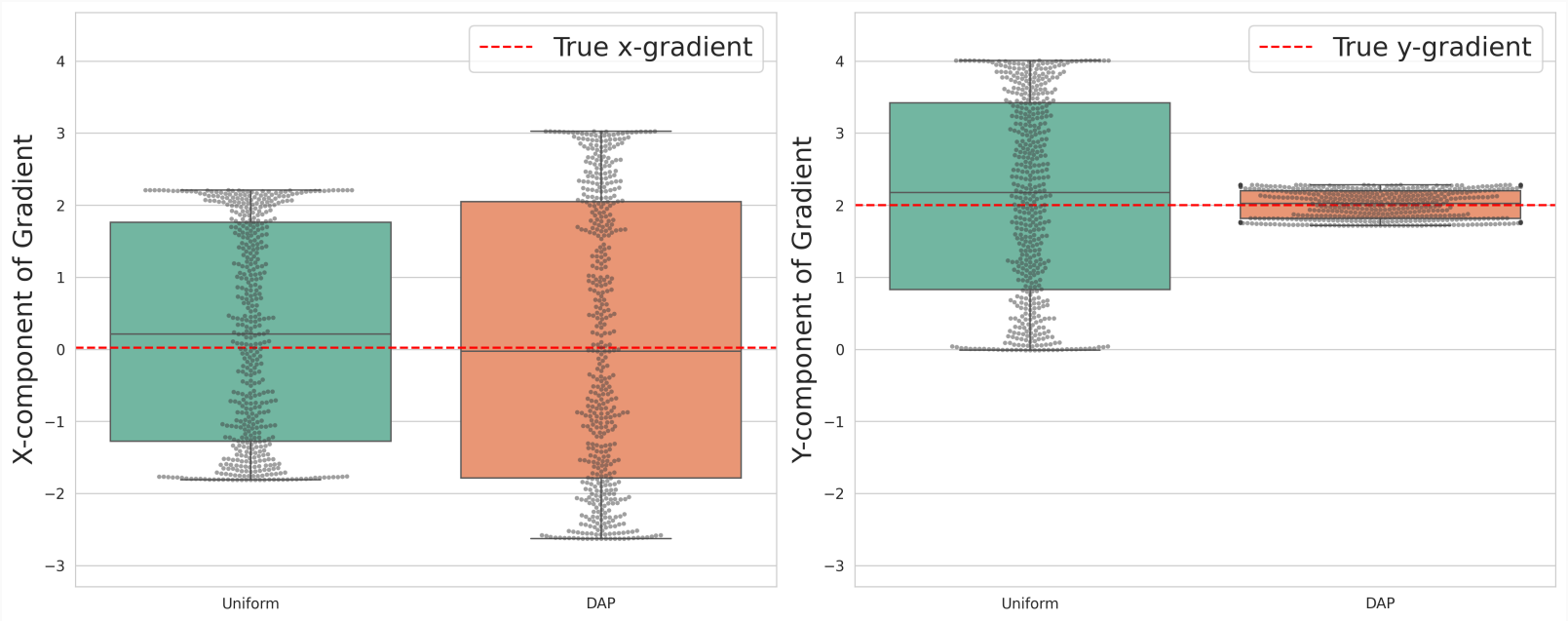
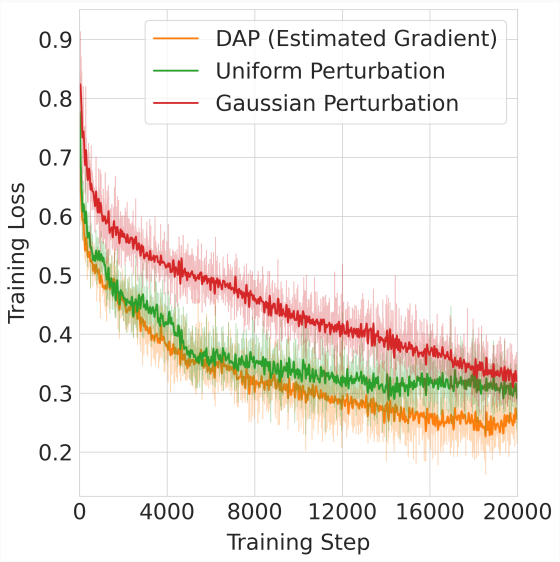


Figure 2: Comparison of gradient estimation performance with estimating the gradient of $f(x) = x_1^2 + x_2^2$ at $x = \begin{bmatrix} 0.1 & 1 \end{bmatrix}^\top$ between uniform random perturbations and DAPs. The non-symmetric estimator is more accurate in the direction with larger gradient.

Applications in LLM Fine-Tuning and Scientific Optimization



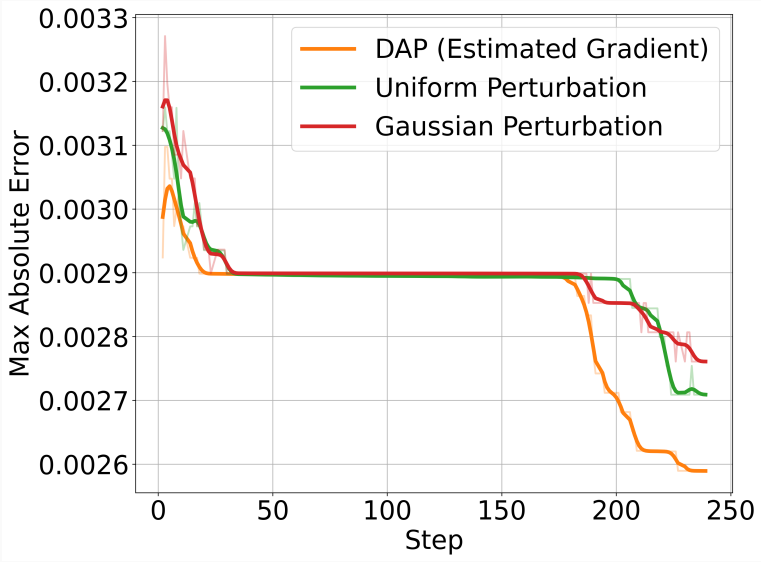


Figure 3: Comparison of training loss curves among different random perturbations on Large Language Model Fine Tuning and Mesh Optimization for the Physical Numerical Solver.

On the Optimal Construction of Unbiased Gradient Estimators for Zeroth-Order Optimization

Shaocong Ma, Heng Huang.

University of Maryland, College Park

NeurIPS 2025 Spotlight

Inherent Bias of Two-Point Estimator

Recap: Zero-Order Gradient Estimator

$$\hat{\nabla} f(x) := \frac{f(x + \mu v) - f(x)}{\mu} v,$$

$$\mathbb{E}[\hat{\nabla} f(x)] \neq \nabla f(x).$$

Inherent Bias of Two-Point Estimator

Recap: Zero-Order Gradient Estimator

$$\hat{\nabla} f(x) := \frac{f(x + \mu v) - f(x)}{\mu} v,$$

$$\mathbb{E}[\hat{\nabla} f(x)] \neq \nabla f(x).$$

The μ Dilemma

We cannot simply take $\mu \to 0$ to eliminate bias:

- **Too Large ($\mu \gg 0$):** Bias $\mathcal{O}(\mu^2)$ dominates, yielding inaccurate gradient estimation.
- **Too Small ($\mu \to 0$):** Triggers **numerical instability**, causing gradient collapse due to floating-point errors.

Unbiased Estimator based on Infinite Series

Q: Is it possible to design an unbiased zeroth-order gradient estimator using only function evaluations?

- *Step 1.* Directional derivative along the direction v.

$$\nabla_v f(x) = \lim_{\mu \to 0} \frac{f(x + \mu v) - f(x)}{\mu}.$$

- *Step 2.* Telescoping series. Let $\mu_n \to 0$.

$$\begin{aligned}\nabla_v f(x) =& \frac{f(x + \mu_1 v) - f(x)}{\mu_1} \\ &+ \sum_{n=1}^{\infty} \left[\frac{f(x + \mu_{n+1} v) - f(x)}{\mu_{n+1}} - \frac{f(x + \mu_n v) - f(x)}{\mu_n} \right].\end{aligned}$$

Unbiased Estimator based on Infinite Series

- *Step 3.* Expectation representation.
  Let $\sum_n p_n = 1$ and $0 < p_n < 1$.

$$\nabla_v f(x) = \sum_{n=1}^{\infty} p_n \Bigg[\frac{f(x+\mu_1 v) - f(x)}{\mu_1} + \frac{1}{p_n} \left(\frac{f(x+\mu_{n+1} v) - f(x)}{\mu_{n+1}} - \frac{f(x+\mu_n v) - f(x)}{\mu_n} \right) \Bigg].$$

Then $\nabla_v f(x)$ can be represented as

$$\mathbb{E}\Bigg[\frac{f(x+\mu_1 v) - f(x)}{\mu_1} + \frac{1}{p_n} \left(\frac{f(x+\mu_{n+1} v) - f(x)}{\mu_{n+1}} - \frac{f(x+\mu_n v) - f(x)}{\mu_n} \right) \Bigg].$$

$\Longrightarrow$ Unbiased Estimator Family

Unbiased Estimator Leads to Better Accuracy

The quadratic loss $f_{\text{reg}} : \mathbb{R}^d \to \mathbb{R}$ and the logistic loss $f_{\text{cls}} : \mathbb{R}^d \to \mathbb{R}$:

$$f_{\text{reg}}(x) = x^\top A^\top A x, \qquad f_{\text{cls}}(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i \cdot (a_i^\top \cdot x))).$$

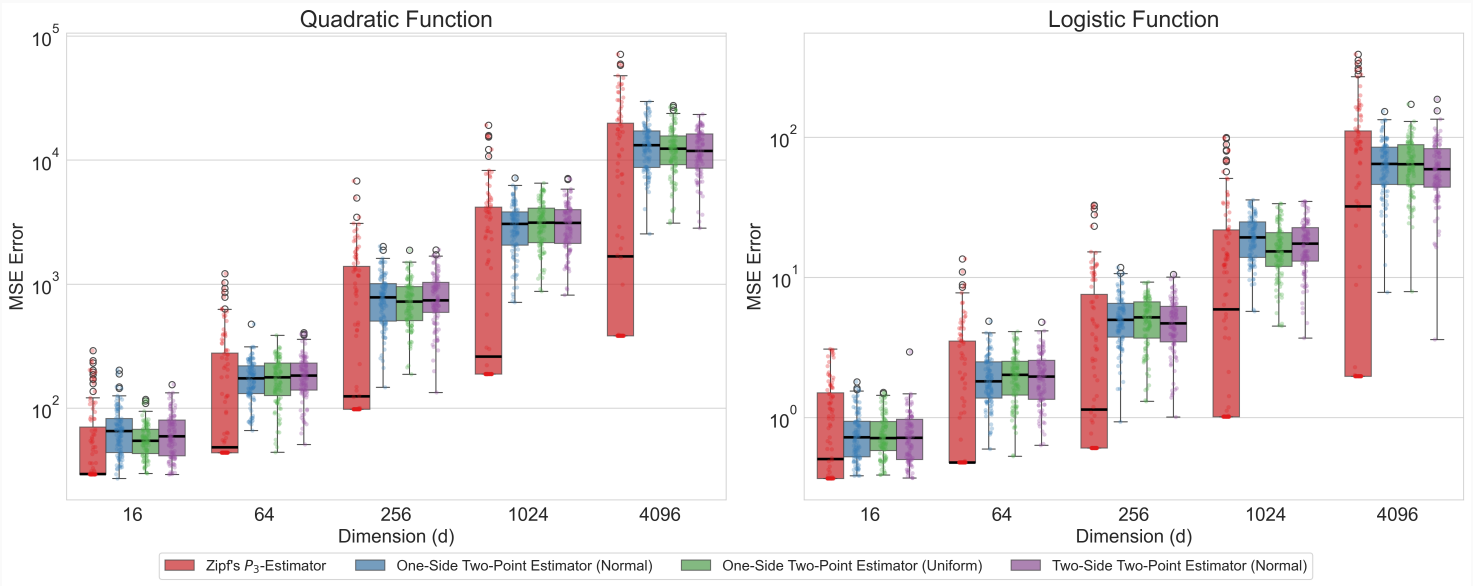


Figure 4: The MSE error (Left: f_{reg}, Right: f_{cls}) of different estimators.

Applications in LLM Fine-Tuning

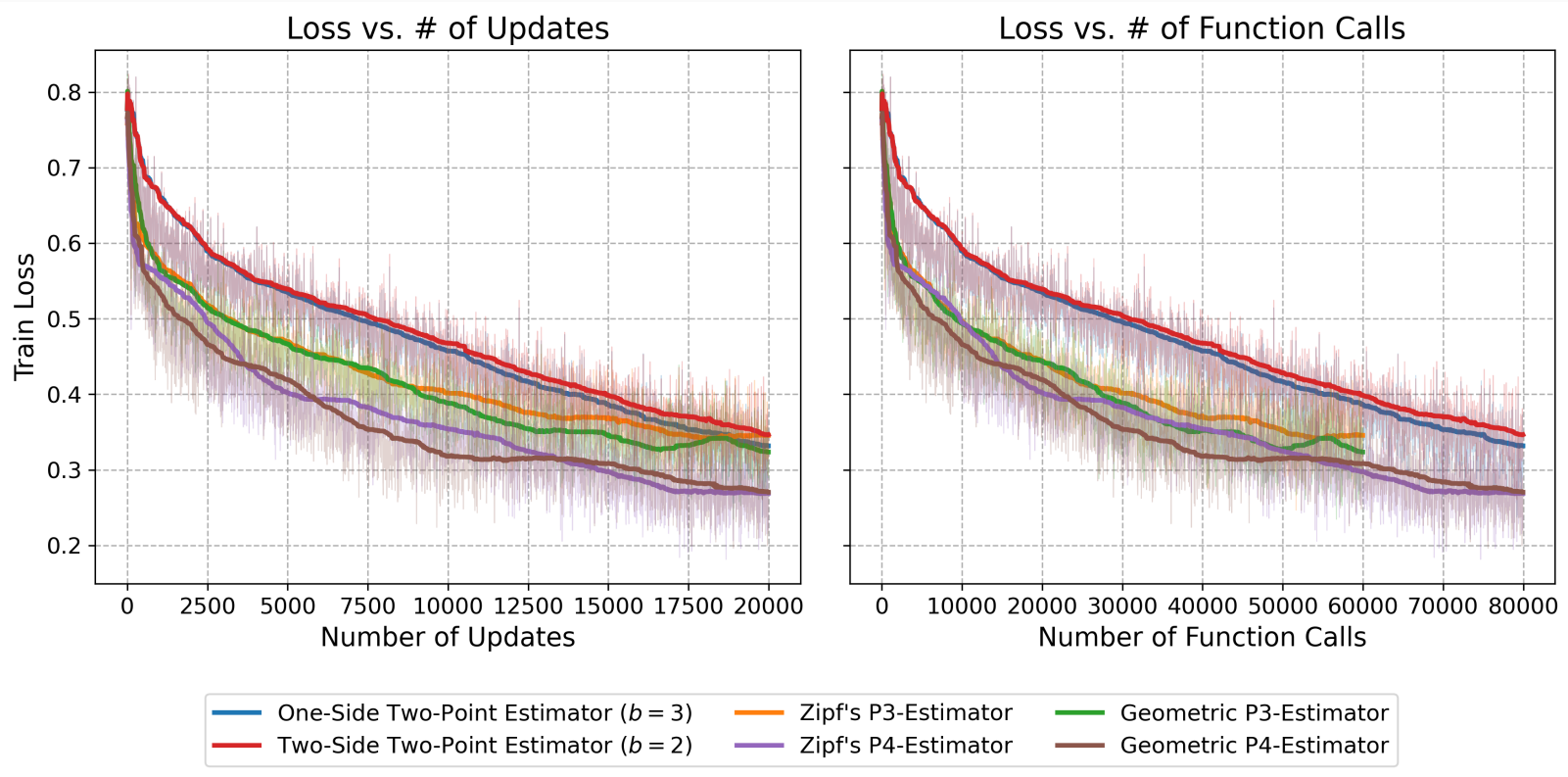


Figure 5: Training loss for fine-tuning the OPT-1.3B model on SST-2 using different gradient estimators.

Future Research Directions

1. Efficient Scaling: 70B Models on Edge

- Enable fine-tuning of large-scale models (e.g., LLaMA-70B) on a **single consumer-grade GPU**.

Future Research Directions

1. **Efficient Scaling: 70B Models on Edge**
   - Enable fine-tuning of large-scale models (e.g., LLaMA-70B) on a **single consumer-grade GPU**.
2. **Black-Box Optimization for Agentic Systems**
   - Automated **prompt engineering** and searching for the optimal **topological structure** of agent interaction graphs.

Future Research Directions

1. **Efficient Scaling: 70B Models on Edge**
   - Enable fine-tuning of large-scale models (e.g., LLaMA-70B) on a **single consumer-grade GPU**.
2. **Black-Box Optimization for Agentic Systems**
   - Automated **prompt engineering** and searching for the optimal **topological structure** of agent interaction graphs.
3. **LLM-Driven Scientific Discovery**
   - Jointly optimizing external parameters within **scientific simulators** where gradients are unavailable.

Thank You!