# Energy-Regularized Sequential Model Editing on Hyperspheres

Qingyuan Liu[1*], Jia-Chen Gu[2*], Yunzhi Yao[3], Hong Wng[4], Nanyun Peng[2]

[1]Columbia University, [2]University of California, Los Angeles,

[3]Zhejiang University, [4]University of Science and Technology of China

*ICLR 2026*

Paper

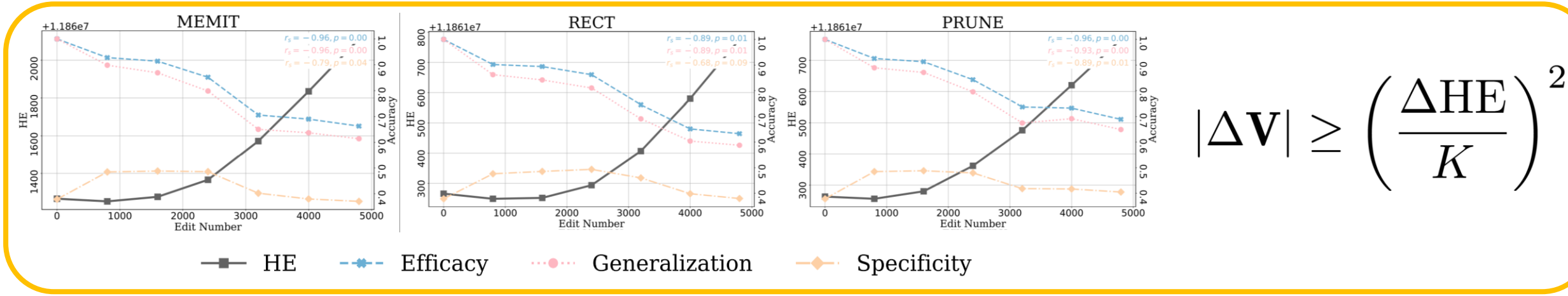
Contact

# Overview

Paper

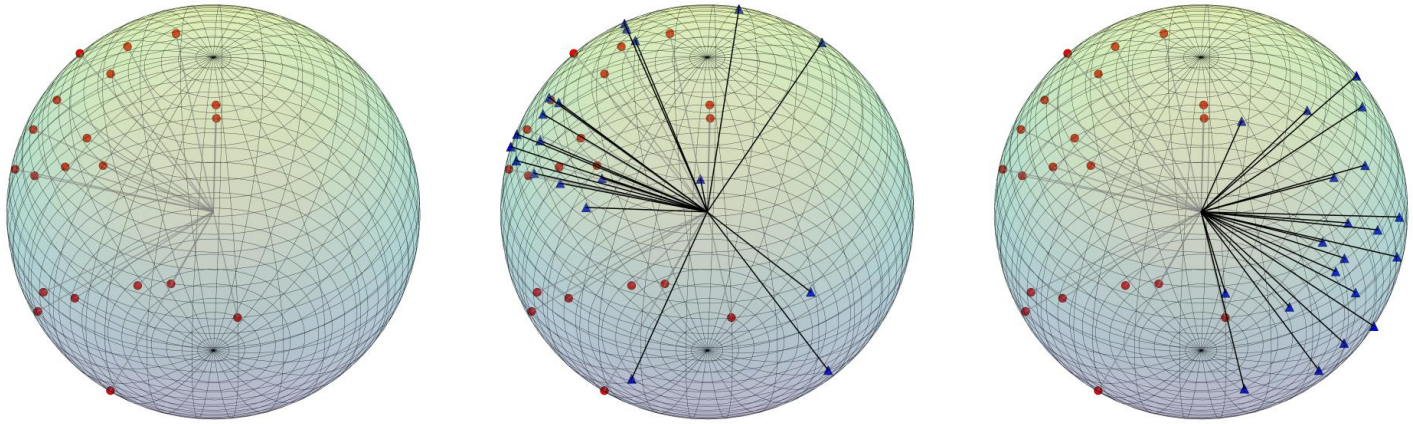
Contact

## Strong Correlation between Angular Diversity and Editing Stability




$$|\Delta \mathbf{V}| \geq \left(\frac{\Delta \mathrm{HE}}{K}\right)^2$$

Collapse!



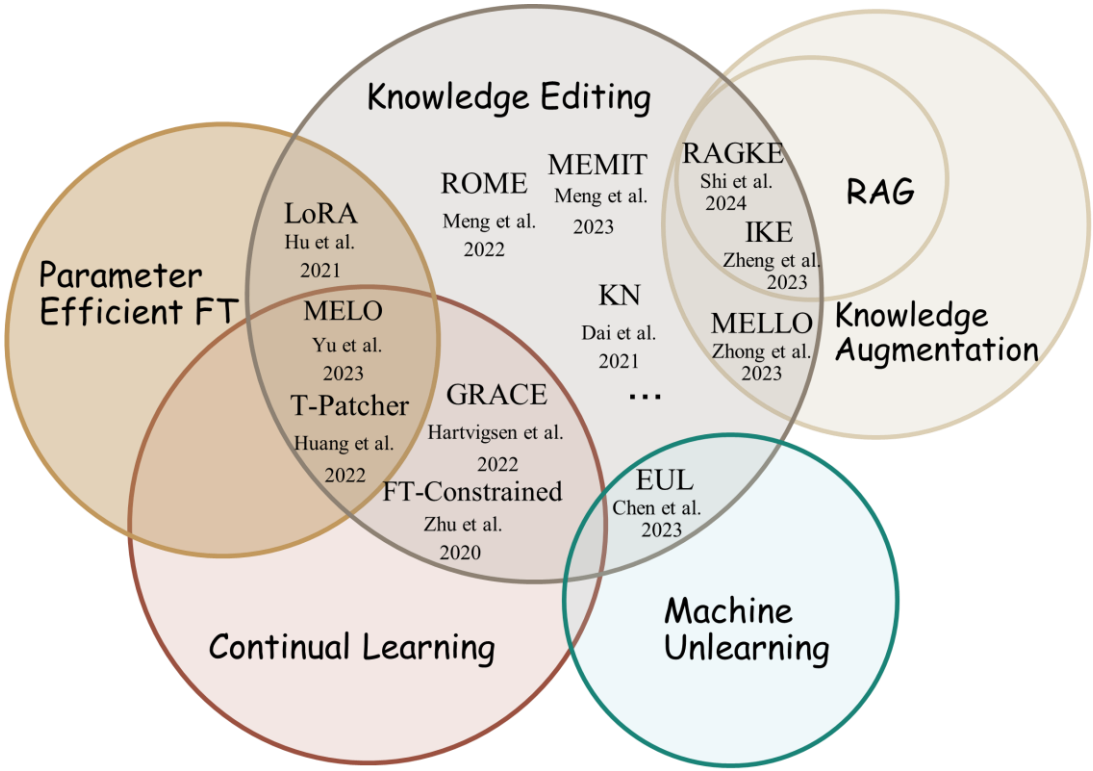
(a) Weight Neurons (b) Current Methods (c) SPHERE (Ours)

# Preliminaries: Model Editing

Paper

Contact

- Model Editing (Knowledge Editing) aims to refine a pre-trained model by applying one or more edits, where each edit replaces a factual association $(s, r, o)$ with new knowledge$(s, r, o^*)$.



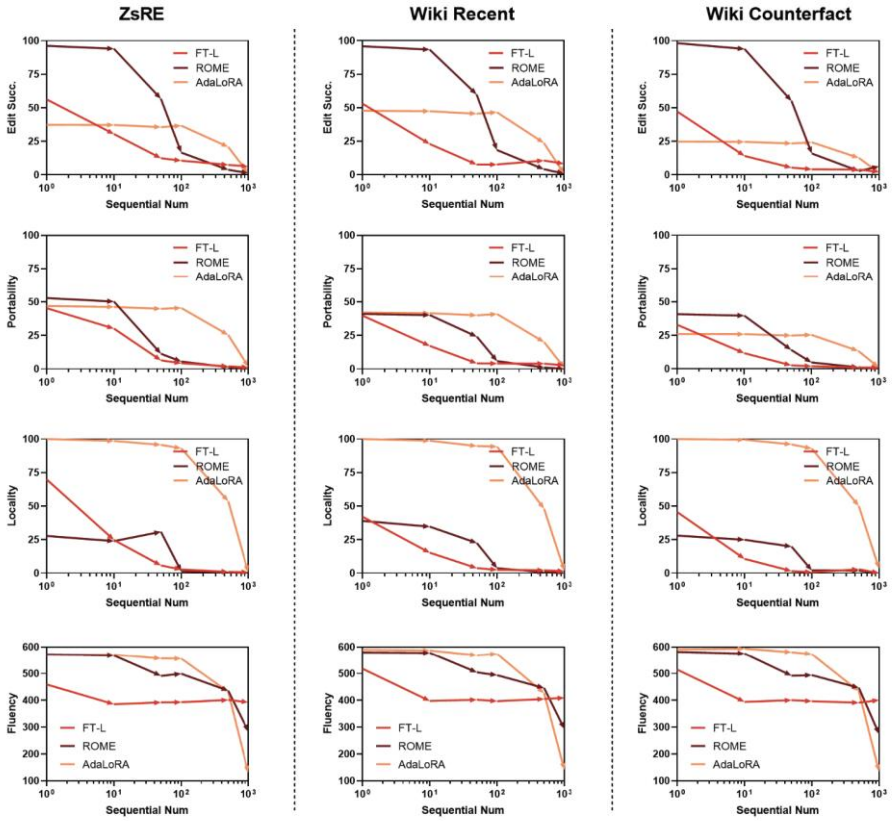

Comparison of different technologies

# Challenges

Paper

Contact

★ **Lifelong Editing (large-scale sequential edit)**: *How to maintain the efficacy of edit while preserving the **general ability** of the edited model?*




Sequential editing results in randomly selected data from WikiData$_{counterfact}$, ZsRE ~and WikiData$_{recent}$ with different numbers.

# Challenges

Paper

Contact

★ **Lifelong Editing (large-scale sequential edit)**: *How to maintain the efficacy of edit while preserving the **general ability** of the edited model?*

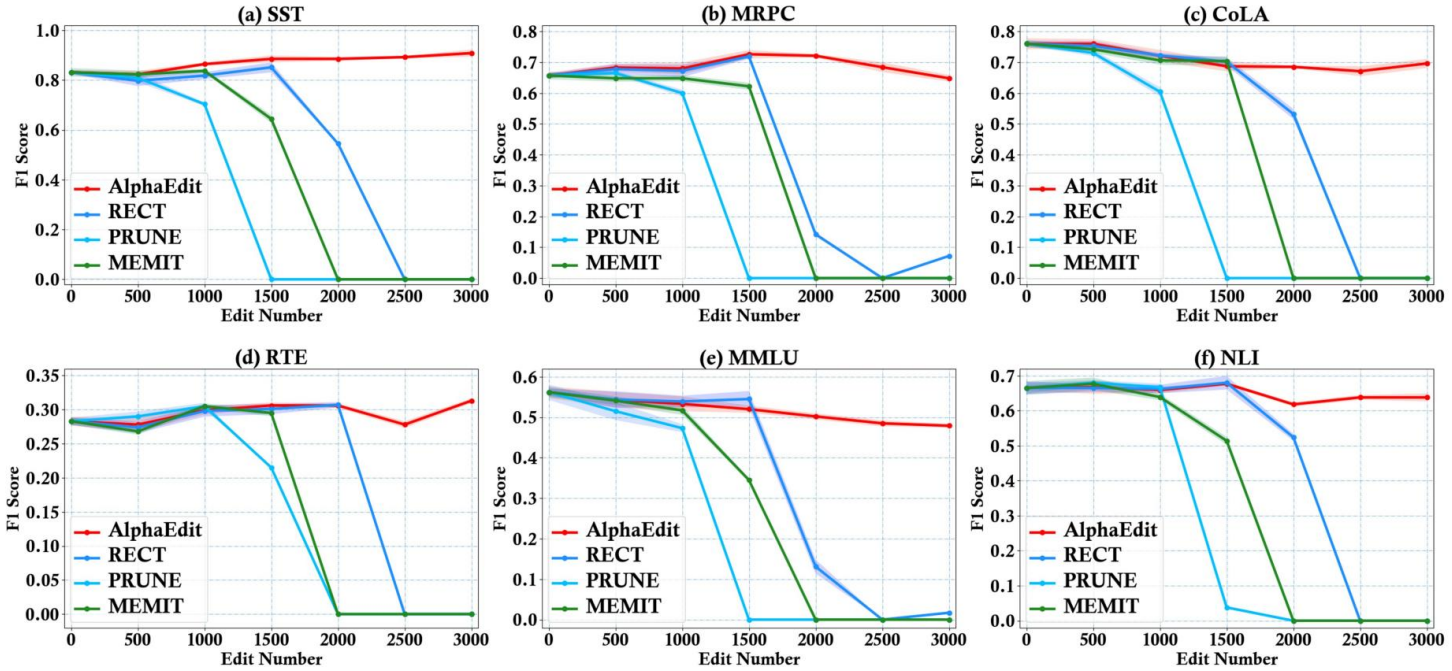



Figure 4: F1 scores of the post-edited LLaMA3 (8B) on six tasks (*i.e.*, SST, MRPC, CoLA, RTE, MMLU and NLI) used for general capability testing. Best viewed in color.

# From Angular Perspective

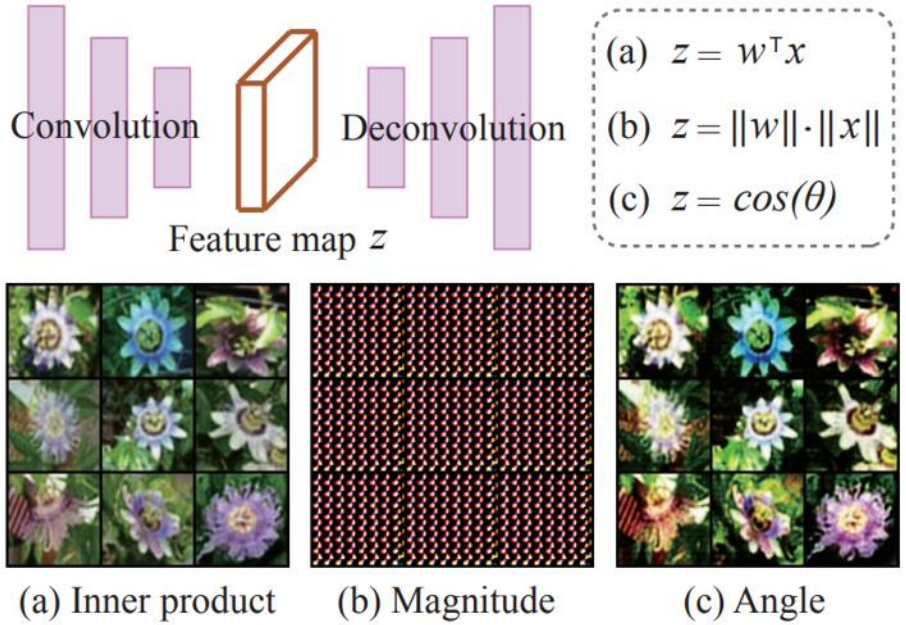
Paper

Contact

- Network generalization by alleviating redundancy through angular diversification.

$$\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) \quad \text{s.t.} \quad 1 \leq i < j \leq m, \quad \left| \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2 \, \|\boldsymbol{w}_j\|_2} \right| \leq \tau$$

- Why Does Orthogonal Transformation Make Sense?



Paper

Contact

# Research Questions

- 1. How can we measure the angular diversity of edited weights?
- 2. Is there any correlation between angular diversity and {editing, general task} performance?
- 3. If Yes, is it possible to design an angular diversity-based regularization method to improve knowledge editing?

# From Angular Perspective

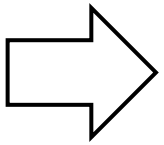
Paper

Contact

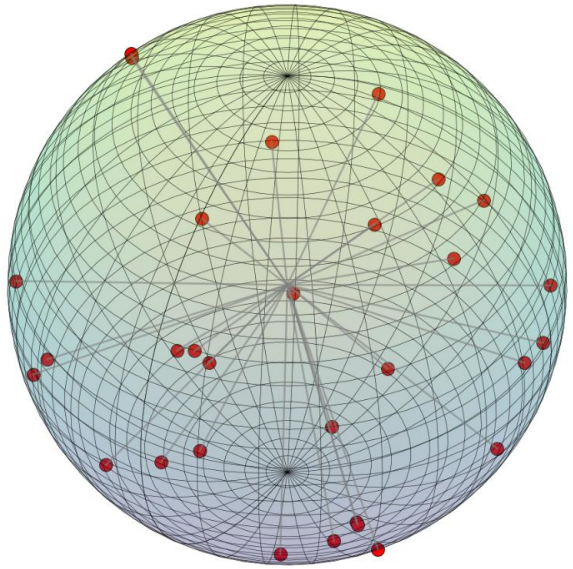
Measured by Hyperspherical Energy (HE)

$$\boldsymbol{E_{s,d}}\left(\hat{\boldsymbol{w}}_i \,|_{i=1}^{N}\right) = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} f_s\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|\right)$$

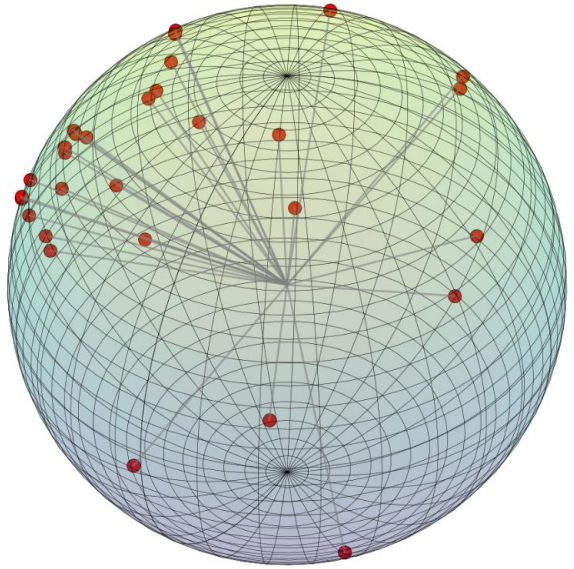
unit — decreasing — distance



Energy ↓
Distance ↑
Uniformity ↑



(a) Weight Neurons (low HE)



(b) Weight Neurons (high HE)

Paper

Contact

# Research Questions

- 2. Is there any correlation between angular diversity and {editing, general task} performance?
- 3. If Yes, is it possible to design an angular diversity-based regularization method to improve knowledge editing?

# From Angular Perspective

Paper

Contact

- Observation 1: Collapse in sequential editing is closely tied to sharp fluctuations in HE.

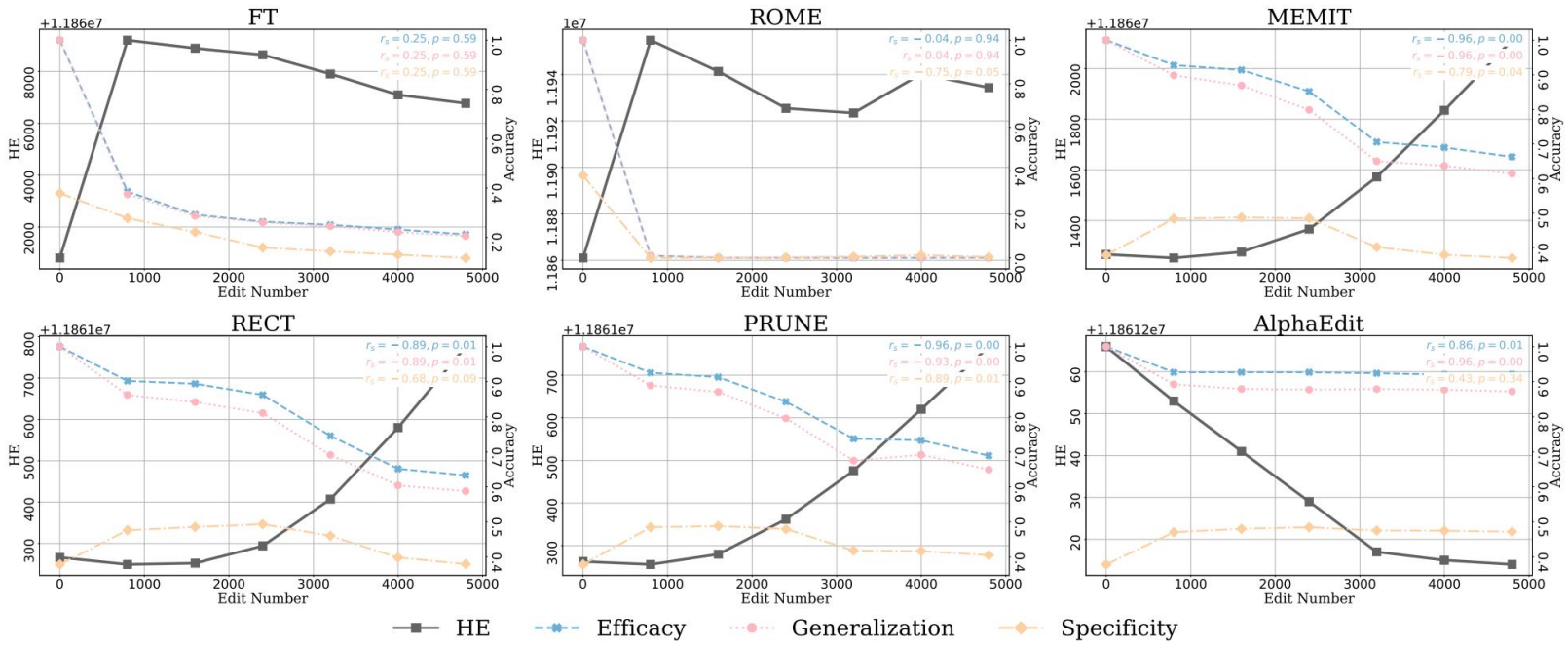
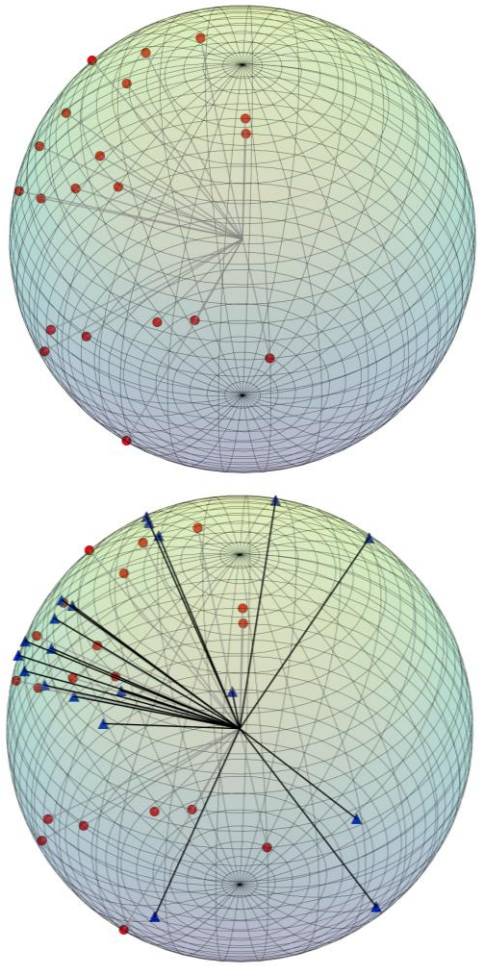



Figure 2: Trends of HE and editing performance throughout sequential editing. The Spearman correlation scores between HE and each editing metric displayed at the end of each curve.



● original weight

▲ update weight

- 5,000 sequential edits with batch size of 100 on LLaMA3-8B.

# From Angular Perspective

Paper

Contact

- Observation 2: Advanced editing methods suppress HE fluctuations effectively.

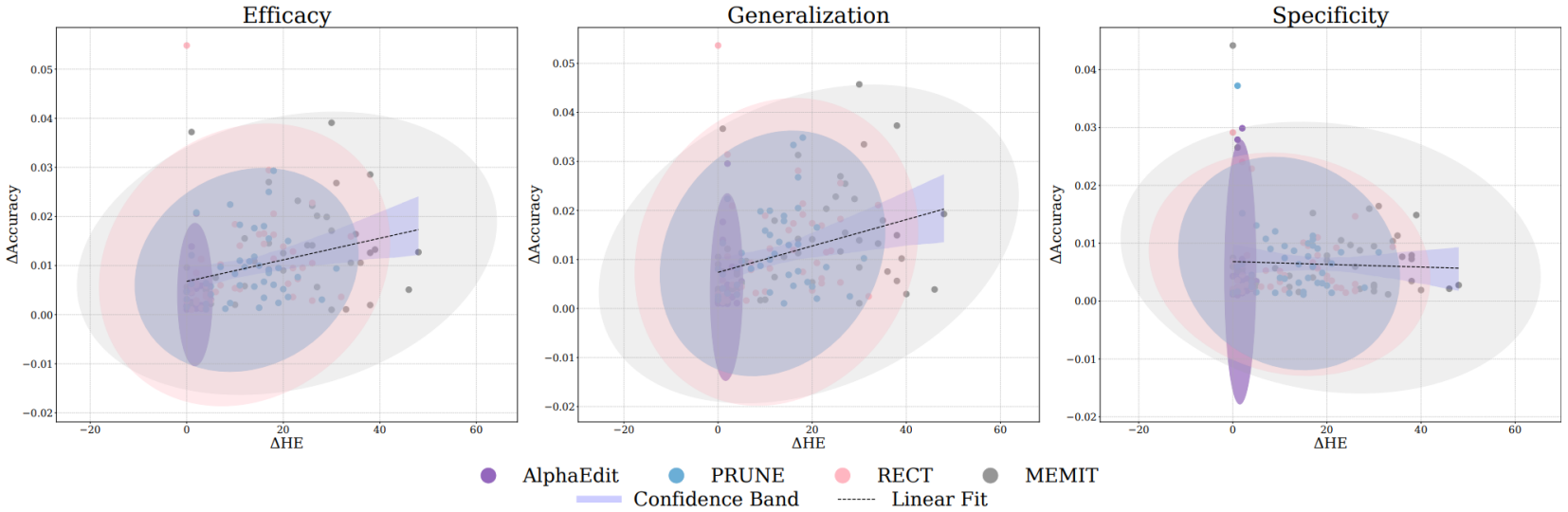



Figure 3: Correlation between changes in HE and editing performance across consecutive edited weights. Each point corresponds to a ΔHE–ΔAcc. pair for one method over five thousand sequential edits. Confidence ellipses and regression lines illustrate overall trends.

- Point: (ΔHE, ΔAcc.)

# From Angular Perspective

Paper Contact

- Theoretically: $|\Delta HE|$ is the lower-bound constraint for $\Delta V$ (disruption caused by editing).

- **Theorem 1 (Lower Bound on Output Perturbation).** Under the assumptions of orthonormal inputs and small perturbations, the output perturbation $\Delta V$ is lower-bounded by squared change in $HE$:

$$|\Delta \mathbf{V}| \geq \left(\frac{\Delta \mathrm{HE}}{K}\right)^2, \qquad K = 4\left(\sum_{k=1}^{p}\left(\sum_{j\neq k}\|\mathbf{w}_k - \mathbf{w}_j\|^{-3}\right)^2\right)^{1/2}$$

where K is a constant dependent on the original weight matrix geometry.

Paper

Contact

# Research Questions

- 3. If Yes, is it possible to design an angular diversity-based regularization method to improve knowledge editing?

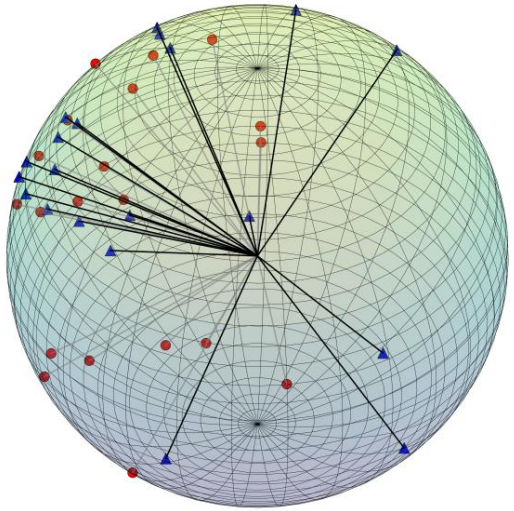
# SPHERE

Paper

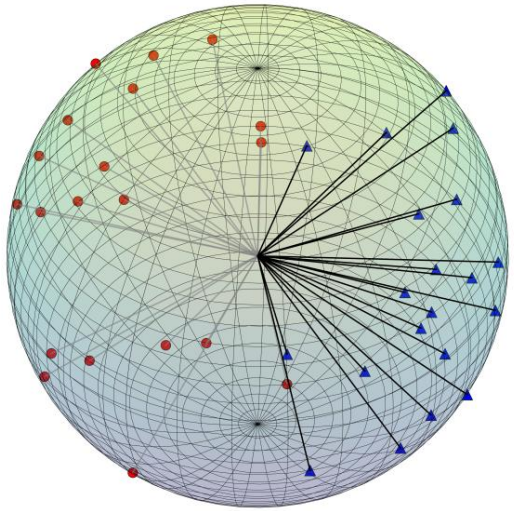
Contact

- We introduce SPHERE (Sparse Projection for Hyperspherical Energy Regularized Editing)

● original weight ▲ update weight



(b) Current Methods



(c) SPHERE (Ours)

Project onto a sparse space which is complementary to principal directions

- Mitigate ▲'s interfere with ●
  -> preserve weight geometry
- Enable new knowledge incorporation

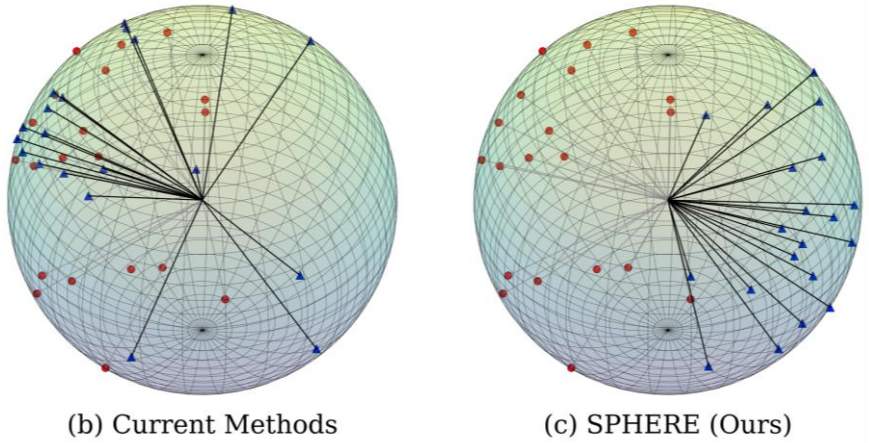
# SPHERE

Paper

Contact

original weight

update weight



(b) Current Methods (c) SPHERE (Ours)

Project onto a sparse space which is complementary to principal directions

- Mitigate ▲'s interfere with ●
  -> preserve weight geometry
- Enable new knowledge incorporation

## 1 Principal Space Estimation

$$\boldsymbol{v} = \arg\max_{\|\hat{\boldsymbol{v}}\|=1} \left(\frac{1}{n}\|\boldsymbol{W}\hat{\boldsymbol{v}}\|^2\right)$$

Top-*r*: $\boldsymbol{U} = [\boldsymbol{v}_{d-r+1}, \ldots, \boldsymbol{v}_d] \in \mathbb{R}^{d\times r}$

## 2 Sparse Space Definition

suppression strength

$\Longleftrightarrow$ orthogonal

$$\boldsymbol{P}_\perp = I - \alpha \boldsymbol{U}\boldsymbol{U}^\top \in \mathbb{R}^{d\times d}$$

## 3 Sparse Space Projection

by any editing method

$$\hat{\boldsymbol{W}} = \boldsymbol{W} + \Delta\boldsymbol{W}\boldsymbol{P}_\perp$$

determined by

Paper

Contact

# Research Questions

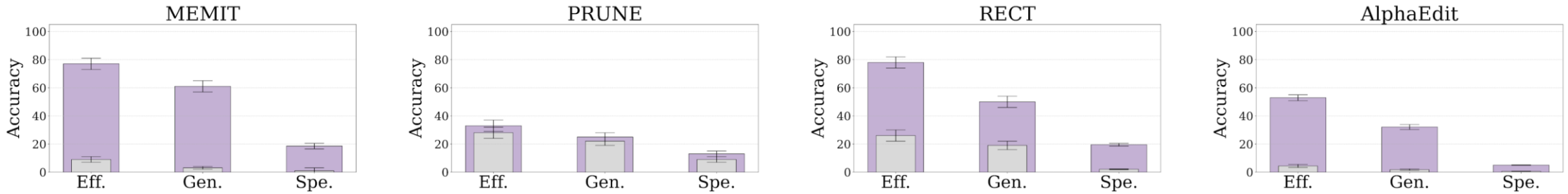
- **RQ1**: Can baseline methods be significantly improved with plug-and-play SPHERE?
- **RQ2**: Can SPHERE effectively preserve the hyperspherical uniformity of edited weights?
- **RQ3**: How does SPHERE-edited LLMs perform on general ability evaluations?

# Experiment

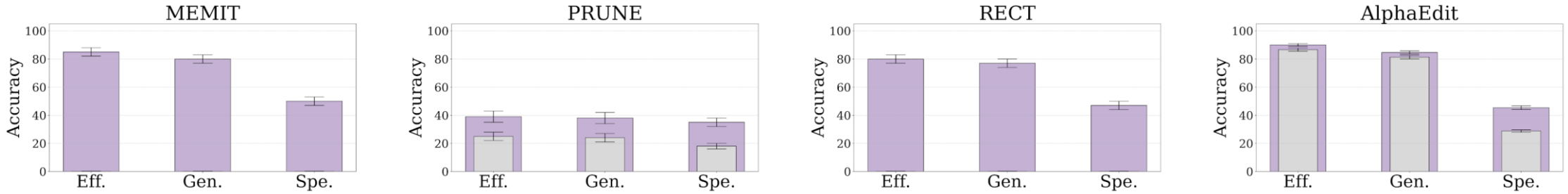
Paper

Contact

- **RQ1**: Can baseline methods be significantly improved with plug-and-play SPHERE?




**(a) Performance Improvement on Counterfact Dataset**




**(b) Performance Improvement on ZsRE Dataset**

Relative improvements of 43.27% (Eff.), 36.20% (Gen.), and 20.96% (Spe.)

# Experiment

Paper

Contact

- **RQ2**: Can SPHERE effectively **preserve the hyperspherical uniformity** of edited weights?

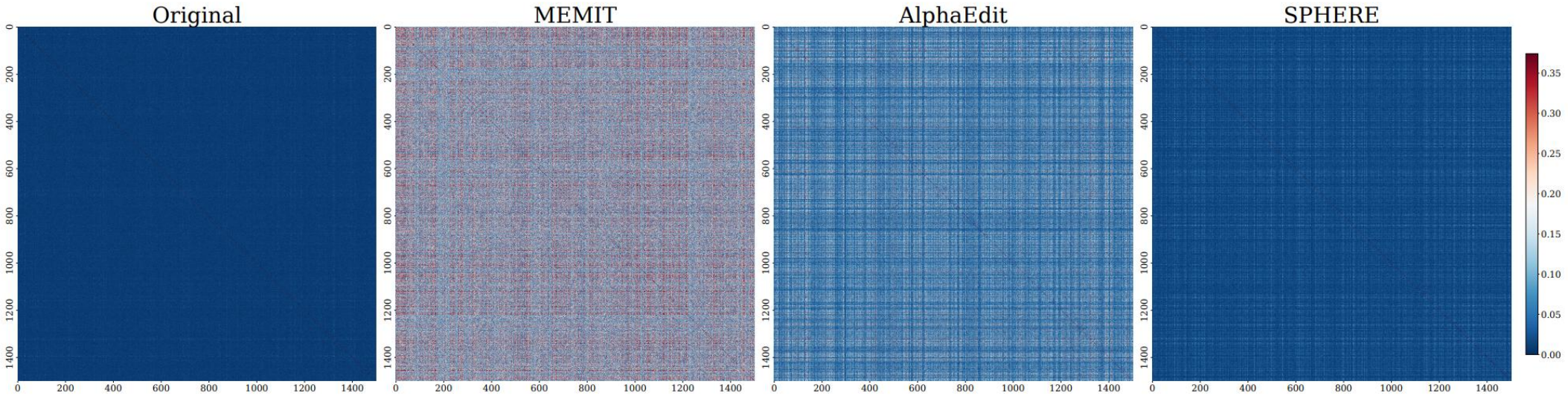



Figure 4: Cosine similarity between neurons in the updated weight matrix after 15,000 edits. Darker colors indicate lower similarity, reflecting better hyperspherical and orthogonal uniformity. SPHERE effectively preserves the weight structure, demonstrating the most stable hyperspherical uniformity.

# Experiment

Paper

Contact

- **RQ3**: How does SPHERE-edited LLMs perform on **general ability** evaluations?

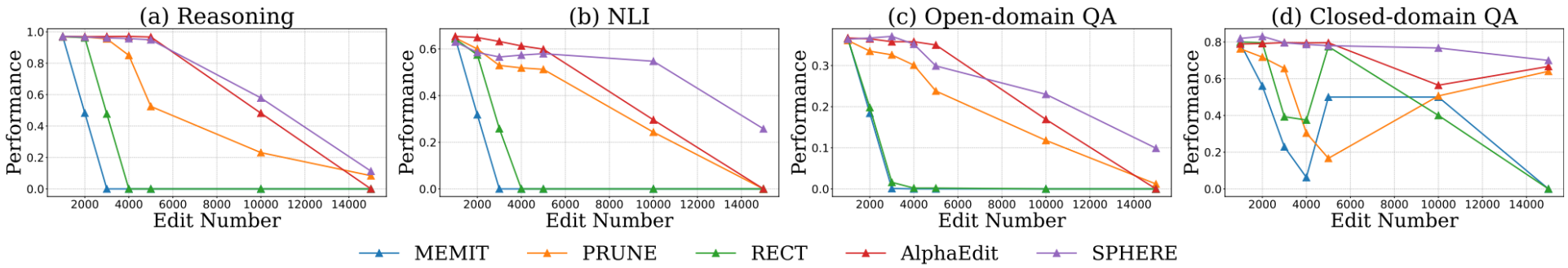



Figure 6: General ability testing of post-edited LLaMA3 (8B) on four tasks.

Effectively preserve the general abilities of post-edited LLMs

# Thank you for listening!

Paper

Contact