



# Community Paper Reading

## Composable Interventions for Language Models



Kyle O'Brien  
Microsoft & EleutherAI



SallyAnn DeLucia  
Arize AI



AI AGENT MASTERY

# From Architecture to Optimization

→ Six week bootcamp starting September 10

→ 12pm PST

[Register now →](#)

# LMs aren't perfect. They can be...



**Expensive:** Even inference can require many GPUs



**Incorrect:** Factual knowledge can become outdated



**Unsafe:** Models can learn potentially undesirable knowledge

# Post-Training Interventions

## Model Editing

- MEMIT
- LoRA
- Fine-Tuning

## Unlearning

- RMU
- Gradient Ascent
- Gradient Difference

## Compression

- Pruning: Wanda
- Pruning: SparseGPT
- Quantization: AWQ
- Quantization: GPTQ

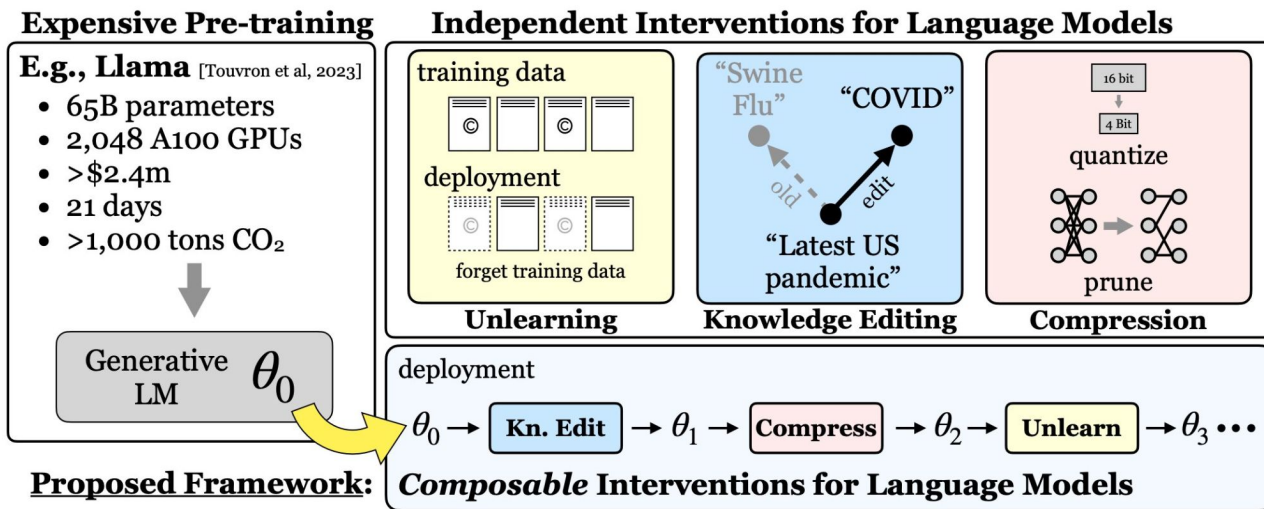
**Intervention:** A technique applied to an already trained LM to steer behavior or improve efficiency. Typically far less computationally expensive than training (pretraining, SFT, RLHF).

**Also known as: test-time, inference-time, post-training, interventions**

# Practical interventions should be

**Order Invariant:** Not sensitive to the order of application

**Not Regress Overall Perf:** Applying interventions does not limit performance



# Are popular interventions composable?

INTERVENTIONS	METHODS	DATASETS	INTERVENTION METRICS	COMPOSABILITY METRICS
<b>Knowledge Editing</b>	Finetuning, LoRA [34], MEMIT [20]	zsRE [35]	Edit Success, Edit Generalization, Strict Edit Locality, MMLU	Order-free Error (Equation 1)
<b>Model Compression</b>	<i>Pruning:</i> Wanda [36], SparseGPT [16] <i>Quantization:</i> GPTQ [15], AWQ [37]	–	MMLU	
<b>Machine Unlearning</b>	Gradient Ascent (GA) [38], Gradient Difference (GD) [39], Representation Misdirection Unlearning (RMU) [40]	WMDP [40]	WMDP, MMLU	Order Sensitivity (Equation 2)

**Model:** Llama-3 8B **Compute:** A100 80GB

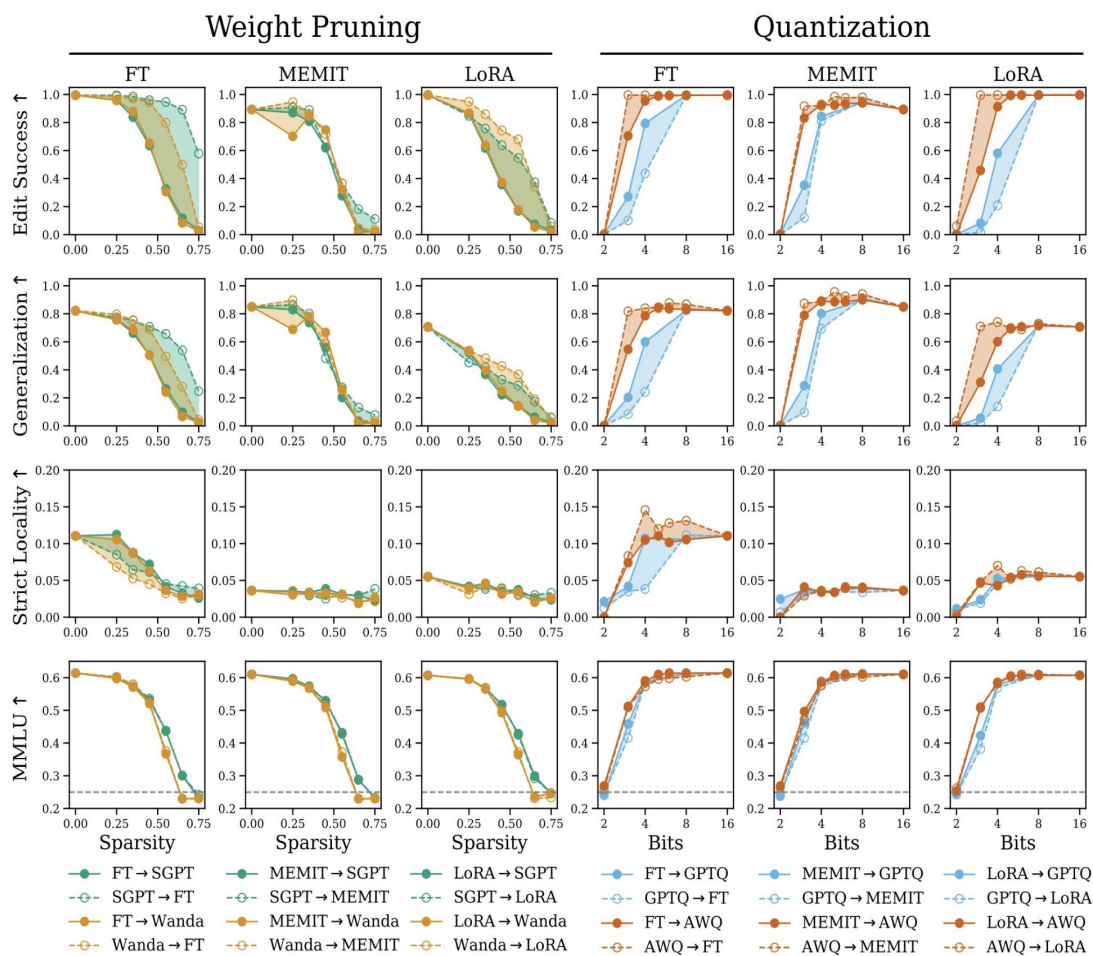
# Editing ↔ Compression

1 — Model compression degrades editing performance

2 — Editing performance hinges on the order of interventions.

3 — Composability can vary within the same intervention category.

4 — Overall utility evaluations fail to measure composability.



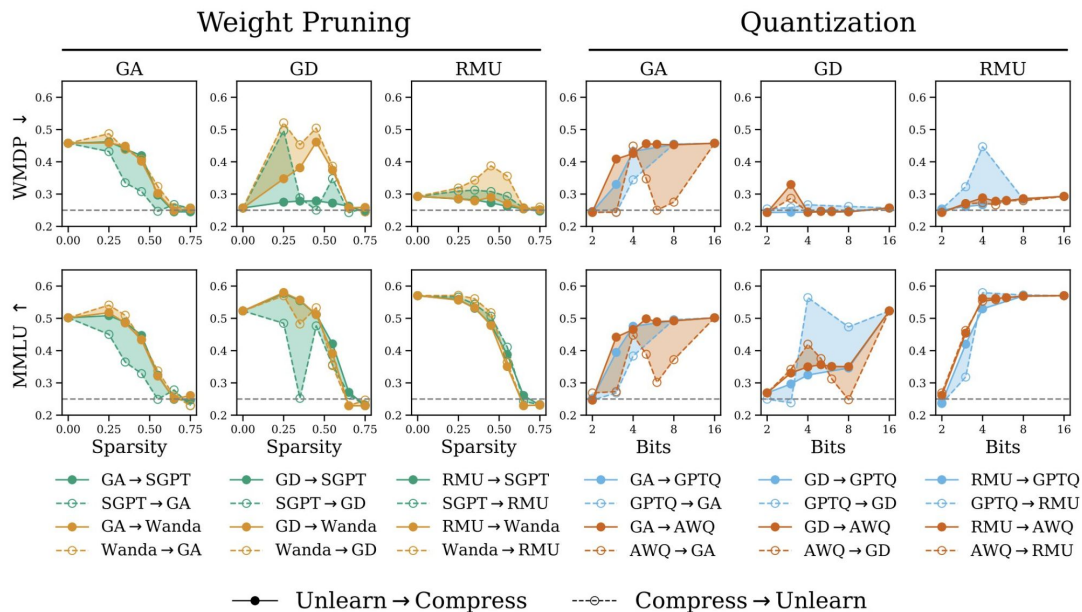
—●— Edit → Compress

--○-- Compress → Edit

# Unlearning $\leftrightarrow$ Compression

5 — Compression hinders unlearning.

6 — Order Sensitivity can determine overall composability



# Editing ↔ Unlearning

## 7 — Editing and unlearning are highly composable for some unlearning methods

Method	Edit Success						Edit Generalization						# Wins
	Order-free Error (↓)			Order Sensitivity (↓)			Order-free Error (↓)			Order Sensitivity (↓)			
	FT	MEMIT	LoRA	FT	MEMIT	LoRA	FT	MEMIT	LoRA	FT	MEMIT	LoRA	
<b>GA</b>	.93	.52	.00	.07	.48	1.0	.96	.59	.22	.04	.41	.78	1
<b>GD</b>	.01	.07	.00	.67	.40	.56	.19	.11	.29	.56	.41	.48	0
<b>RMU</b>	.00	.03	.00	.01	.01	.00	.18	.07	.29	.03	.04	.04	<b>10</b>
<i># Wins</i>	0	0	<b>3</b>	1	1	1	0	<b>2</b>	1	<b>2</b>	1	0	

Method	WMDP (Unlearning)						MMLU						# Wins
	Order-free Error (↓)			Order Sensitivity (↓)			Order-free Error (↓)			Order Sensitivity (↓)			
	FT	MEMIT	LoRA	FT	MEMIT	LoRA	FT	MEMIT	LoRA	FT	MEMIT	LoRA	
<b>GA</b>	.47	.40	.28	.00	.05	.07	.47	.51	.64	.01	.04	.07	1
<b>GD</b>	.29	.26	.30	.00	.02	.24	.41	.42	.41	.18	.22	.14	4
<b>RMU</b>	.28	.29	.29	.04	.01	.00	.43	.44	.44	.01	.00	.04	<b>5</b>
<i># Wins</i>	1	1	1	<b>2</b>	0	1	<b>2</b>	0	0	1	1	1	

# Takeaways

**Aggressive compression struggles to compose well**

**Editing and Unlearning are (generally) composable**

**Overall performance does not measure composability**

- New metrics are needed

# Future Work

Studying More Interventions

Studying Across LM Families

Composability Scaling Laws

Complicated Compositions

**1: Interventions are a promising way to make LMs better**

**2: More work is needed to make interventions practical**

**3: It would be awesome *IF* we get this right**

# Collaborators



**Arinbjörn Kolbeinsson**

Visiting Scholar  
University of Virginia



**Tianjin Huang**

Assistant Professor  
University of Exeter



**Tom Hartvigsen**

Assistant Professor  
University of Virginia

**Shanghua Gao** (Harvard), **Shiwei Liu** (Oxford), **Jonathan Richard Schwarz** (Harvard), **Anurag Vaidya** (Harvard), **Faisal Mahmood** (Harvard), **Marinka Zitnik** (Harvard), **Tianlong Chen** (UNC),

**@KyleDevinOBrien — kyobrien.io**

---

**Thank You**