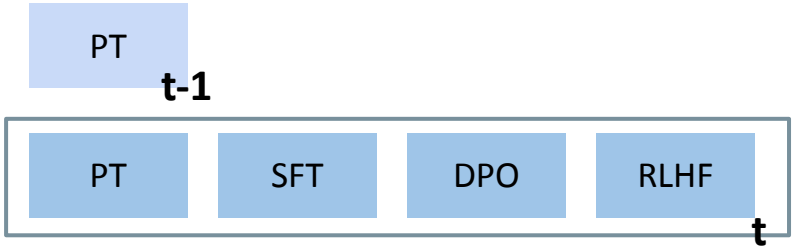
Recycling Fine-tuning across Training

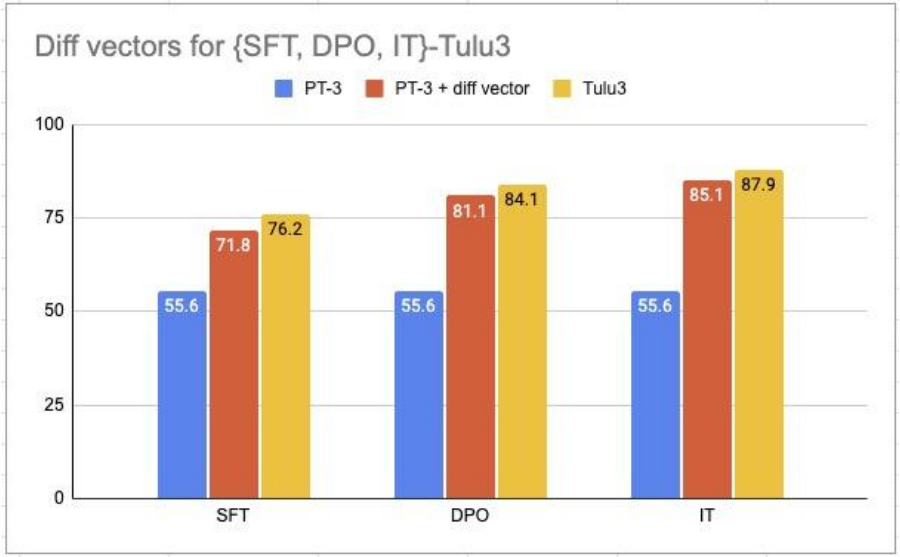
Pin-Jie Lin

Adapt fine-tuning (FT) artifacts to ever-changing Environment



transferring skill by ***diff vector***

$$PT_{t-1} + vec(FT_t - PT_t)$$



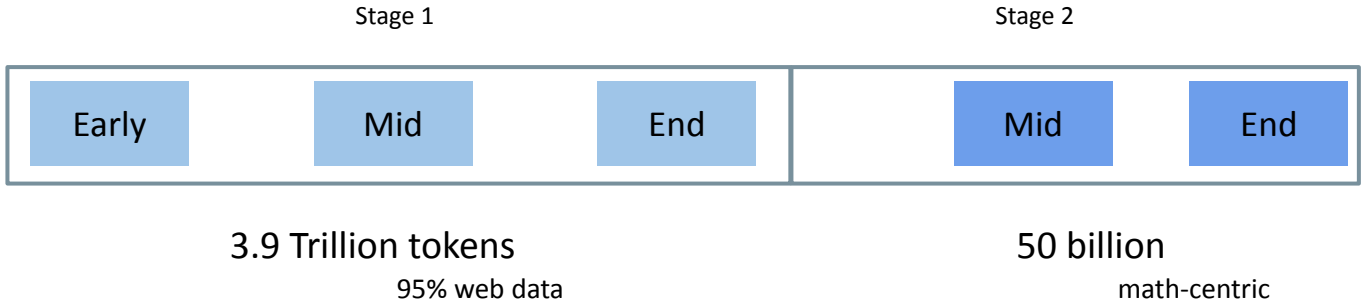
- Diff vector between {IT, DPO, SFT}-Tulu3 and IT-3.1

(1) Diff vector across fine-tuning

- **Stage 1**: Pretraining (3.9T)
- **Stage 2**: Midtraining (50B)

- Data: Tulu3 math data for 30K steps
- Diff vector:
  vec(M-FT-30K - M-PT)

OLMo2



Stage 1

- Early: **M1** (300K steps, 1.2T tokens)
- Mid: **M2** (600K, 2.5T)
- End: **M3** (928K, 3.9T)

Stage 2

- Mid: **M4** (6K, 26B)
- End: **M5** (12K, 50B)

GSM8K					
	early-stage1	mid-stage1	end-stage1	mid-stage2	end-stage2
PT	13.26	19.48	24.41	64.51	65.5
5k	30.93	41.01	44.95	67.39	70.12
10k	36.23	45.33	50.79	69.67	71.41
15k	40.56	46.62	52.99	71.56	73.84
20k	42.45	50.34	56.02	72.85	73.61
25k	43.44	49.73	59.05	73.99	74.98
30k	**45.18**	**50.79**	**60.42**	**75.73**	**75.58**

- Consistently improvement with finetung steps
- Gains diminish as M sees more math-centric data mid-training

merged model on GSM8K					
	base				
w/ vec	M1	M2	M3	M4	M5
PT	13.26	19.48	24.41	64.51	65.5
FT	**45.18**	**50.79**	**60.42**	**75.73**	75.58
early-stage1 (M1)		26.61	32.06	27.59	19.63
mid-stage1 (M2)	19.02		39.87	25.92	17.36
end-stage1 (M3)	14.32	25.01		68.61	70.35
mid-stage2 (M4)	11.82	18.04	22.66		**77.1**
end-stage2 (M5)	11.9	16.07	24.03	72.93	

M1 + vec(FT-30K-M2 - M2)

- effective to improve performance
- diff vectors from the same trajectory is more useful
  - mostly from M_j + vec(FT-M_j-1 - M_j-1)
- outperform FT
  - 77.1. from M5 + vec(FT-M4 - M4)

other tasks

ARC_C	early-stage1	mid-stage1	end-stage1	mid-stage2	end-stage2
PT	65.27	**70.05**	72.69	**77.81**	**79.43**
5k	66.04	69.62	**74.14**	77.3	78.66
10k	66.55	68.6	73.29	77.21	78.32
15k	66.55	68.17	73.63	77.64	77.98
20k	65.78	67.49	71.58	77.21	77.38
25k	65.52	68.25	72.09	76.53	77.73
30k	**66.97**	68.34	72.86	77.47	78.49

merged model on ARC_C	M1	M2	M3	M4	M5
PT	65.27	**70.05**	72.69	**77.81**	**79.43**
FT	**66.97**	68.34	72.86	77.47	78.49
early-stage1 (M1)		68.43	73.2	75.85	76.36
mid-stage1 (M2)	65.35		72.52	75.25	76.27
end-stage1 (M3)	65.1	68.6		75	76.53
mid-stage2 (M4)	64.84	69.96	**73.54**		79.09
end-stage2 (M5)	64.76	69.02	**73.54**	77.47	

DROP	early-stage1	mid-stage1	end-stage1	mid-stage2	end-stage2
PT	35.26	38.76	40.89	58.04	61
5k	38.65	41.42	42.44	59.55	**61.57**
10k	39.13	41.4	42.75	59.68	61.29
15k	39.02	41.65	43.58	57.37	59.29
20k	39.33	41.62	**43.93**	59.22	60.69
25k	**39.4**	41.62	43.46	**60**	61.21
30k	39.37	**42.03**	43.58	59.64	61.17

merged model on DROP	M1	M2	M3	M4	M5
PT	35.26	38.76	40.89	58.04	61
FT	39.37	**42.03**	**43.58**	**59.64**	**61.17**
early-stage1 (M1)		41.4	42.31	55.07	55.26
mid-stage1 (M2)	**39.4**		41.72	52.3	53.22
end-stage1 (M3)	39.06	41.57		57.58	59.01
mid-stage2 (M4)	38.97	41.43	42.44		61.03
end-stage2 (M5)	38.89	41.65	43.01	59.25	

Natural Q					
	early-stage1	mid-stage1	end-stage1	mid-stage2	end-stage2
PT	**24.87**	26.24	**28.92**	**35.32**	**35.87**
5k	24.69	26.57	28.34	33.84	34.94
10k	24.78	26.72	28.12	32.71	33.57
15k	24.38	26.23	27.91	33.36	33.31
20k	24.54	25.96	27.49	32.65	33.38
25k	24.52	26.69	27.85	32.99	33.63
30k	24.15	**27.18**	27.78	32.49	33.32

merged model on NQ					
	M1	M2	M3	M4	M5
PT	**24.87**	26.24	**28.92**	**35.32**	**35.87**
FT	24.15	**27.18**	27.78	32.49	33.32
early-stage1 (M1)		26.76	28.23	33.03	32.44
mid-stage1 (M2)	23.74		27.59	32.88	33.32
end-stage1 (M3)	23.94	25.76		32.09	33.4
mid-stage2 (M4)	24.1	26.23	28.04		33.16
end-stage2 (M5)	24.35	26.06	27.77	32.64	

MATH					
	early-stage1	mid-stage1	end-stage1	mid-stage2	end-stage2
PT	3.45	4.77	5.76	16.73	19.04
5k	5.21	6.74	7.78	18.83	**19.44**
10k	6.08	5.46	8.02	17.54	17.47
15k	7.56	5.54	10.85	18	18.78
20k	8.22	7.11	11.83	18.45	18.93
25k	7.55	6.29	11.98	18.37	18.16
30k	**8.26**	**7.42**	**13.01**	**19.44**	18.94

(2) Merged model as initialization

2-1: Applying Newer Vectors to Previous Models

- from *i* = 1 to 4:
  M_j ← M_j + vec(FT-M_j-i - M_j)

- Examples
  M5 + vec(FT-M4 - M4)
  M4 + vec(FT-M3 - M3)...

- Total: 10 merged models in total
- Training

M_j + vec(FT-M_j-1 - M_j-1)

GSM8k	M2 + vec(M1-FT - M1)	M3 + vec(M2-FT - M2)	M4 + vec(M3-FT - M3)	M5 + vec(M4-FT - M4)
PT	19.48	24.41	64.51	65.5
FT	50.79	60.42	75.73	75.58
Merged model	26.61	39.87	68.61	77.1
CT-5k	44.65	53.82	70.58	73.23
CT-10k	50.03	55.52	75.81	73.23
CT-15k	51.09	56.93	75.66	76.57
CT-20k	52.01	58.98	75.43	76.42
CT-25k	54.05	61.25	75.43	**77.17**
CT-30k	**56.93**	**62.69**	**77.55**	76.95

- Continually adapting merged model > FT

M_j + vec(FT-M_j-1 - M_j-1) & M_j + vec(FT-M_j-2 - M_j-2)

GSM8k	M2 + vec(M1-FT - M1)	M3 + vec(M2-FT - M2)	M4 + vec(M3-FT - M3)	M5 + vec(M4-FT - M4)
PT	19.48	24.41	64.51	65.5
FT	50.79	60.42	75.73	75.58
Merged model	26.61	39.87	68.61	77.1
CT-5k	44.65	53.82	70.58	73.23
CT-10k	50.03	55.52	75.81	73.23
CT-15k	51.09	56.93	75.66	76.57
CT-20k	52.01	58.98	75.43	76.42
CT-25k	54.05	61.25	75.43	**77.17**
CT-30k	**56.93**	**62.69**	**77.55**	76.95

GSM8k	M3 + vec(M1-FT - M1)	M4 + vec(M2-FT - M2)	M5 + vec(M3-FT - M3)
PT	24.41	64.51	65.5
FT	60.42	75.73	75.58
Merged model	32.06	25.92	70.35
CT-5k	49.88	72.63	73
CT-10k	54.66	72.25	75.43
CT-15k	55.49	74.45	77.33
CT-20k	56.55	75.43	76.42
CT-25k	59.13	76.8	77.1
CT-30k	**62.77**	**78.62**	**78.77**

all results

continually training of	merged model				
	base				
w/ vec	M1	M2	M3	M4	M5
PT	13.26	19.48	24.41	64.51	65.5
FT	45.18	50.79	60.42	75.73	75.58
early-stage1 (M1)		**56.93**	**62.77**	77.78	78.62
mid-stage1 (M2)			62.69	**78.62**	78.69
end-stage1 (M3)				77.55	**78.77**
mid-stage2 (M4)					77.17
end-stage2 (M5)					

training M5 + vec(FT-M1 - M1)

training M5 + vec(FT-M4 - M4)

M + vec(FT-M_j-1 - M_j-1)

ARC_C	M2 + vec(M1-FT - M1)	M3 + vec(M2-FT - M2)	M4 + vec(M3-FT - M3)	M5 + vec(M4-FT - M4)
PT	**70.05**	72.69	**77.81**	**79.43**
FT	68.34	**72.86**	77.47	78.49
Merged model	68.43	72.52	75	79.09
CT-5k	66.89	72.52	75.93	78.49
CT-10k	66.29	72.35	76.36	77.73
CT-15k	65.78	71.33	76.36	78.41
CT-20k	66.8	70.3	75.59	77.3
CT-25k	66.46	70.22	75.93	78.32
CT-30k	67.15	71.92	77.55	78.07

DROP	M2 + vec(M1-FT - M1)	M3 + vec(M2-FT - M2)	M4 + vec(M3-FT - M3)	M5 + vec(M4-FT - M4)
PT	38.76	40.89	58.04	61
FT	42.03	43.58	**59.64**	**61.17**
Merged model	41.4	41.72	57.58	61.03
CT-5k	41.59	42.95	52.35	59.26
CT-10k	41.5	43.01	53.06	57.6
CT-15k	41.53	43.08	52.29	56.98
CT-20k	41.78	43.84	53.27	59.17
CT-25k	42.06	43.55	53.43	59.33
CT-30k	**42.21**	**43.62**	54.49	58.58

M + vec(FT-M_j-1 - M_j-1)

NQ	M2 + vec(M1-FT - M1)	M3 + vec(M2-FT - M2)	M4 + vec(M3-FT - M3)	M5 + vec(M4-FT - M4)
PT	26.24	28.92	35.32	35.87
FT	**27.18**	**27.78**	**32.49**	33.32
Merged model	26.76	27.58	32.09	33.16
CT-5k	26.7	26.38	31.13	33.31
CT-10k	26.97	26.78	30.69	33.06
CT-15k	26.81	27.15	31.24	32.52
CT-20k	26.62	25.76	31.65	**33.35**
CT-25k	26.62	25.72	31.87	32.9
CT-30k	27.09	26.36	31.59	32.64

merged model on GSM8K					
	base				
w/ vec	M1	M2	M3	M4	M5
PT	13.26	19.48	24.41	64.51	65.5
FT	**45.18**	**50.79**	**60.42**	**75.73**	75.58
early-stage1 (M1)		26.61	32.06	27.59	19.63
mid-stage1 (M2)	19.02		39.87	25.92	17.36
end-stage1 (M3)	14.32	25.01		68.61	70.35
mid-stage2 (M4)	11.82	18.04	22.66		**77.1**
end-stage2 (M5)	11.9	16.07	24.03	72.93	

M1 - (1.2T)

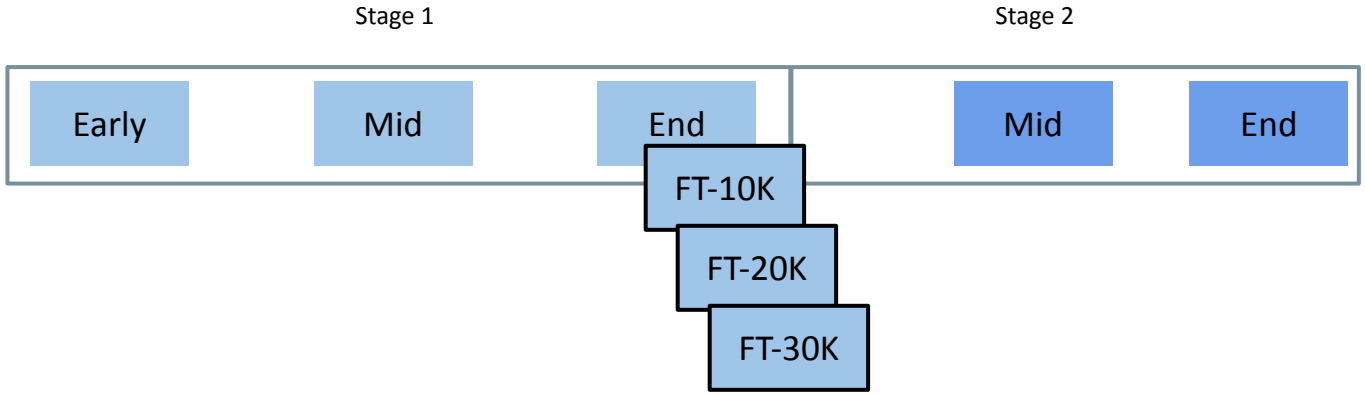
| - 1.3T

M2 - (2.5T)

| - 1.4T

M3 - (3.9T) — 26B — M4 - (3.9T + 26B) — 24B — M4 - (3.9T + 50B)

OLMo2


Stage 1
Stage 2
Early
Mid
End
Mid
End
FT-10K
FT-20K
FT-30K


Early
FT-30K
End

Recurrent merging fine-tuned vector

Online setup

FT_M_1 ← FT(M_1)

M_2’ ← M_2 + vec(FT_M1 - M_1)

FT_M_2’ ← FT(M_2’)

M_3’ ← M3 + vec(FT(M_2 + vec(FT_M1 - M_1)) - M2)

M_3’ ← M3 + vec(FT_M_2’ - M2)

One step

FT_M_1 ← FT(M_1)

M_2’ ← M_2 + vec(FT_M_1 - M_1)

M_3’ ← M3 + vec(FT(M_2) - M_2)

M_4’ ← M4 + vec(FT_M_3 - M_3)

...