Appendix

Pile SAE transfer results

All of these results filter outlier activations.

Mistral 7B

Models: Mistral-7B, Mistral-7B Instruct
site: resid_pre layer 16
SAE widths: 131027

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE Delta	0 Abl. CE loss	Explained Variance %	MSE
Base	Base	95	98.71%	1.51	1.63	0.12	10.37	68.1%	1014
Chat	Base	72	96.82%	1.51	1.79	0.28	10.37	52.6%	1502
Chat	Chat	101	99.01%	1.70	1.78	0.08	10.37	69.2%	1054
Base	Chat	126	98.85%	1.70	1.80	0.10	10.37	60.9%	1327

Qwen 1.5 0.5B

Models: Qwen1.5 0.5B, Qwen1.5 0.5B chat
Site: resid_pre layer 13
SAE widths: 32768

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Base	Base	22	96.58 %	2.29	2.65	0.36	12.91	74.4%	184
Chat	Base	44	96.47 %	2.29	2.67	0.38	12.91	69.9%	217

Chat	Chat	27	95.75 %	2.99	3.42	0.43	12.99	76.1%	198
Base	Chat	18	97.64 %	2.99	3.23	0.24	12.99	59.2%	344

Gemma v1 2B

Models: Gemma v1 2B, Gemma- v1 2B It
Site: resid_pre layer 9
SAE widths: 32768

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Base	Base	39	95.63 %	2.11	2.56	0.45	12.45	77.6%	80
Chat	Base	6928	-832%	2.11	98.30	96.19	12.45	-266213%	855916
Chat	Chat	60	99.40 %	3.13	3.19	0.06	12.45	79.1%	302
Base	Chat	219	25.32 %	3.13	10.12	6.99	12.45	-1499 %	40534

Alpaca SAE transfer results

All of these filter outlier activations from the evals.

Mistral 7B Instruct

Model: Mistral 7B Instruct
site: resid_pre layer 16
SAE widths: 131027

Section: Rollout

SAE	Model	L0	CE Loss	Clean CE	SAE CE	CE delta	0 Abl. CE	Explained	MSE

			rec %	Loss	Loss		loss	Variance %	
Chat	Chat	168	97.67	0.16	0.46	0.30	12.92	54.4%	1860
Base	Chat	190	97.42	0.16	0.49	0.33	12.92	49.7%	2060

Section: User Prompt

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Chat	Chat	95	100.9 %	3.25	3.17	-0.08	11.63	62.6%	1411
Base	Chat	147	99.95 %	3.25	3.25	0.00	11.63	52.3%	1805

Qwen 1.5 0.5B Chat

Model: Qwen1.5 0.5B chat
Site: resid_pre layer 13
SAE widths: 32768

Section: Rollout

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Chat	Chat	43	95.87 %	0.62	1.20	0.58	14.68	66.6%	235
Base	Chat	30	95.13 %	0.62	1.31	0.69	14.68	57.0%	303

Section: User Prompt

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE

Chat	Chat	31	78.3%	4.20	6.55	2.35	15.06	67.1%	213
Base	Chat	58	104%	4.20	3.69	-0.51	15.06	55.3%	290

Gemma v1 2B it

Model: Gemma 2B it
site: resid_pre layer 9
SAE widths: 32768

Section: Rollout

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Chat	Chat	79	96.72 %	0.32	0.72	0.40	12.73	70.8%	516
Base	Chat	371	-54.10 %	0.32	19.45	19.13	12.73	-4517 %	127253

Section: User Prompt

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Chat	Chat	66	100.2 %	6.43	6.40	-0.03	20.59	74.4%	610
Base	Chat	735	-182%	6.43	46.40	39.97	20.59	-13574 %	475377

SAE fine-tuning results

All of these are on the pile, and we include all outlier activations. We did not fine-tune the Gemma v1 2B base SAE.

Mistral 7B Instruct

Model: Mistral-7B Instruct
Site: resid_pre layer 16
SAE widths: 131027
Not ignoring outliers

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Chat	Chat	101	99.01 %	1.70	1.78	0.08	10.37	69.4%	1054
Base	Chat	170	98.38 %	1.70	1.84	0.14	10.37	32.2%	724350
Fine-tuned base	Chat	86	98.75 %	1.70	1.81	0.11	10.37	65.4%	1189

Qwen 1.5 0.5B Chat

Here we fine-tune a Qwen1.5 0.5B base SAE on 2.5 million activations from the chat model.

Model: Qwen1.5 0.5B chat
Site: resid_pre layer 13
SAE widths: 32768
Not ignoring outliers

SAE	Model	L0	CE Loss rec %	Clean CE Loss	SAE CE Loss	CE delta	0 Abl. CE loss	Explained Variance %	MSE
Chat	Chat	28	95.73 %	2.99	3.42	0.43	12.99	76.2%	198
Base	Chat	46	93.08 %	2.99	3.69	0.7	12.99	-160.2 %	67934
Fine-tuned base	Chat	28	96.06 %	2.99	3.38	0.39	12.99	74.2%	215