Correlations: Benchmarks and Human–Model Alignment

	ifeval	bbh	math	gpqa	musr	mmlu	Procrustes R2	cka	RSM
ifeval	1.00	0.44	0.67	0.43	0.03	0.50	0.68	0.64	0.67
bbh	0.44	1.00	0.72	0.88	0.50	0.94	0.74	0.70	0.64
math	0.67	0.72	1.00	0.89	0.19	0.84	0.61	0.57	0.51
gpqa	0.43	0.88	0.89	1.00	0.37	0.88	0.56	0.51	0.46
musr	0.03	0.50	0.19	0.37	1.00	0.45	0.36	0.34	0.33
mmlu	0.50	0.94	0.84	0.88	0.45	1.00	0.71	0.68	0.62
Procrustes R2	0.68	0.74	0.61	0.56	0.36	0.71	1.00	0.99	0.97
cka	0.64	0.70	0.57	0.51	0.34	0.68	0.99	1.00	0.98
RSM	0.67	0.64	0.51	0.46	0.33	0.62	0.97	0.98	1.00

Pearson r