	MMLU	SAD	facts_human_defaults	facts_llms	facts_which_llm	facts_names	influence	introspection_count_tokens	introspection_predict_tokens	introspection_rules	stages_full	stages_oversight	self_recognition_who	self_recognition_groups	id_leverage_entity_name	id_leverage_multihop	output_control	do_not_imitate
MMLU	1	0.76	0.69	0.85	0.68	0.46	0.79	0.21	0.15	0.39	0.86	0.47	0.01	0.55	0.38	0.42	0.06	0.33
SAD	0.76	1	0.95	0.74	0.7	0.71	0.9	0.24	0.26	0.12	0.81	0.44	0.38	0.62	0.77	0.85	-0.23	-0.07
facts_human_defaults	0.69	0.95	1	0.64	0.55	0.68	0.92	0.17	0.29	0.07	0.79	0.47	0.4	0.59	0.67	0.76	-0.2	-0.06
facts_llms	0.85	0.74	0.64	1	0.81	0.51	0.7	0.49	-0.08	0.13	0.65	0.52	0.15	0.28	0.62	0.6	-0.16	-0.14
facts_which_llm	0.68	0.7	0.55	0.81	1	0.56	0.53	0.29	-0.09	0.24	0.66	0.38	0	0.37	0.75	0.67	-0.47	-0.16
facts_names	0.46	0.71	0.68	0.51	0.56	1	0.5	0.19	-0.13	-0.11	0.5	0.07	0.67	0.53	0.63	0.76	-0.33	-0.23
influence	0.79	0.9	0.92	0.7	0.53	0.5	1	0.27	0.33	0.23	0.75	0.62	0.29	0.55	0.52	0.64	-0.09	0
introspection_count_tokens	0.21	0.24	0.17	0.49	0.29	0.19	0.27	1	-0.31	0.1	0.05	0.54	0.29	-0.08	0.21	0.28	-0.12	-0.4
introspection_predict_tokens	0.15	0.26	0.29	-0.08	-0.09	-0.13	0.33	-0.31	1	0.48	0.36	0.29	-0.24	0.28	-0.16	-0.11	0.26	0.48
introspection_rules	0.39	0.12	0.07	0.13	0.24	-0.11	0.23	0.1	0.48	1	0.44	0.54	-0.36	0.23	-0.25	-0.23	-0.03	0.47
stages_full	0.86	0.81	0.79	0.65	0.66	0.5	0.75	0.05	0.36	0.44	1	0.44	-0.04	0.7	0.48	0.48	-0.09	0.36
stages_oversight	0.47	0.44	0.47	0.52	0.38	0.07	0.62	0.54	0.29	0.54	0.44	1	-0.14	0.13	0.14	0.17	-0.07	-0.09
self_recognition_who	0.01	0.38	0.4	0.15	0	0.67	0.29	0.29	-0.24	-0.36	-0.04	-0.14	1	0.11	0.35	0.57	-0.17	-0.43
self_recognition_groups	0.55	0.62	0.59	0.28	0.37	0.53	0.55	-0.08	0.28	0.23	0.7	0.13	0.11	1	0.27	0.41	-0.17	0.53
id_leverage_entity_name	0.38	0.77	0.67	0.62	0.75	0.63	0.52	0.21	-0.16	-0.25	0.48	0.14	0.35	0.27	1	0.94	-0.53	-0.47
id_leverage_multihop	0.42	0.85	0.76	0.6	0.67	0.76	0.64	0.28	-0.11	-0.23	0.48	0.17	0.57	0.41	0.94	1	-0.46	-0.44
output_control	0.06	-0.23	-0.2	-0.16	-0.47	-0.33	-0.09	-0.12	0.26	-0.03	-0.09	-0.07	-0.17	-0.17	-0.53	-0.46	1	0.32
do_not_imitate	0.33	-0.07	-0.06	-0.14	-0.16	-0.23	0	-0.4	0.48	0.47	0.36	-0.09	-0.43	0.53	-0.47	-0.44	0.32	1