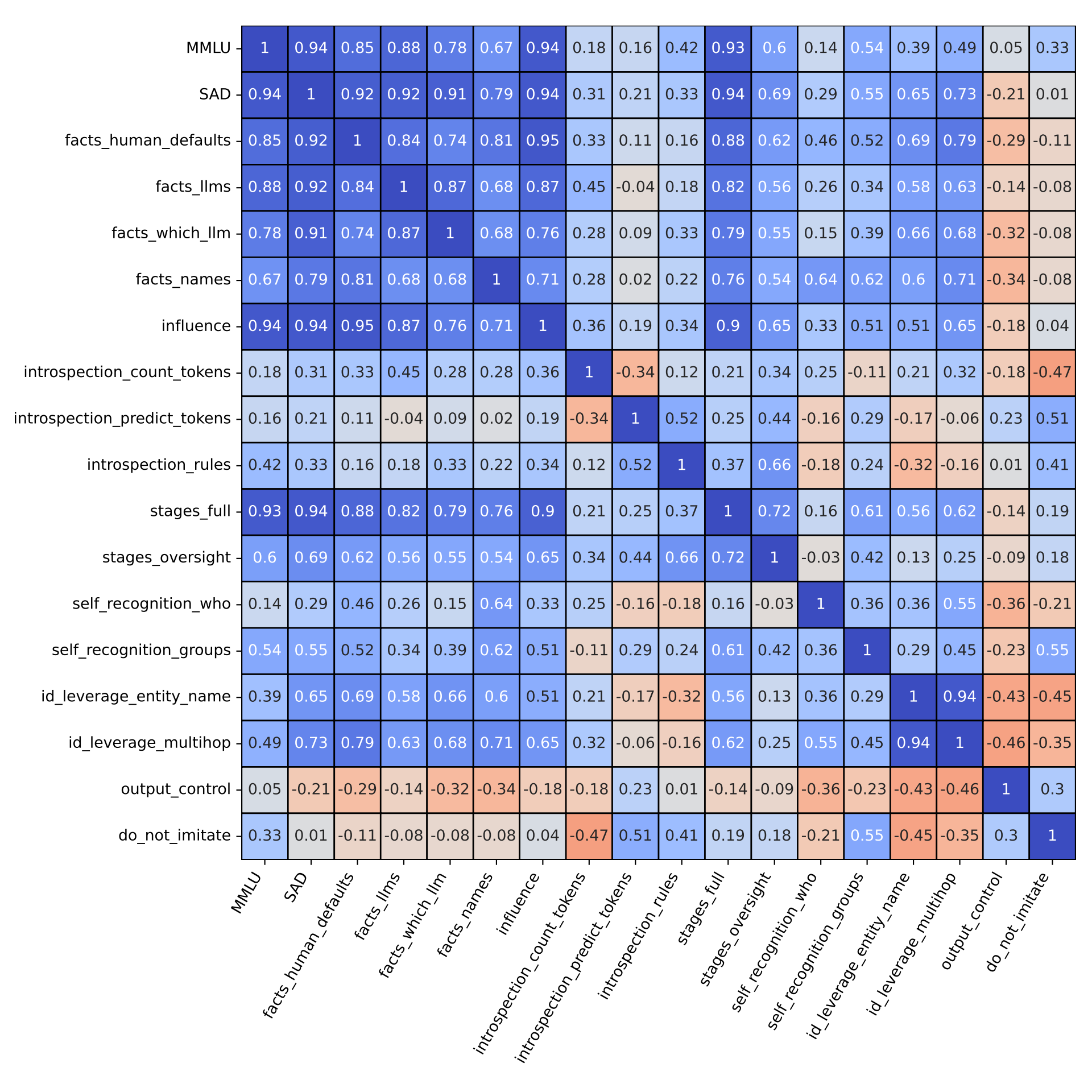
MMLU 1 0.94 0.85 0.88 0.78 0.67 0.94 0.18 0.16 0.42 0.93 0.6 0.14 0.54 0.39 0.49 0.05 0.33
SAD 0.94 1 0.92 0.92 0.91 0.79 0.94 0.31 0.21 0.33 0.94 0.69 0.29 0.55 0.65 0.73 -0.21 0.01
facts_human_defaults 0.85 0.92 1 0.84 0.74 0.81 0.95 0.33 0.11 0.16 0.88 0.62 0.46 0.52 0.69 0.79 -0.29 -0.11
facts_llms 0.88 0.92 0.84 1 0.87 0.68 0.87 0.45 -0.04 0.18 0.82 0.56 0.26 0.34 0.58 0.63 -0.14 -0.08
facts_which_llm 0.78 0.91 0.74 0.87 1 0.68 0.76 0.28 0.09 0.33 0.79 0.55 0.15 0.39 0.66 0.68 -0.32 -0.08
facts_names 0.67 0.79 0.81 0.68 0.68 1 0.71 0.28 0.02 0.22 0.76 0.54 0.64 0.62 0.6 0.71 -0.34 -0.08
influence 0.94 0.94 0.95 0.87 0.76 0.71 1 0.36 0.19 0.34 0.9 0.65 0.33 0.51 0.51 0.65 -0.18 0.04
introspection_count_tokens 0.18 0.31 0.33 0.45 0.28 0.28 0.36 1 -0.34 0.12 0.21 0.34 0.25 -0.11 0.21 0.32 -0.18 -0.47
introspection_predict_tokens 0.16 0.21 0.11 -0.04 0.09 0.02 0.19 -0.34 1 0.52 0.25 0.44 -0.16 0.29 -0.17 -0.06 0.23 0.51
introspection_rules 0.42 0.33 0.16 0.18 0.33 0.22 0.34 0.12 0.52 1 0.37 0.66 -0.18 0.24 -0.32 -0.16 0.01 0.41
stages_full 0.93 0.94 0.88 0.82 0.79 0.76 0.9 0.21 0.25 0.37 1 0.72 0.16 0.61 0.56 0.62 -0.14 0.19
stages_oversight 0.6 0.69 0.62 0.56 0.55 0.54 0.65 0.34 0.44 0.66 0.72 1 -0.03 0.42 0.13 0.25 -0.09 0.18
self_recognition_who 0.14 0.29 0.46 0.26 0.15 0.64 0.33 0.25 -0.16 -0.18 0.16 -0.03 1 0.36 0.36 0.55 -0.36 -0.21
self_recognition_groups 0.54 0.55 0.52 0.34 0.39 0.62 0.51 -0.11 0.29 0.24 0.61 0.42 0.36 1 0.29 0.45 -0.23 0.55
id_leverage_entity_name 0.39 0.65 0.69 0.58 0.66 0.6 0.51 0.21 -0.17 -0.32 0.56 0.13 0.36 0.29 1 0.94 -0.43 -0.45
id_leverage_multihop 0.49 0.73 0.79 0.63 0.68 0.71 0.65 0.32 -0.06 -0.16 0.62 0.25 0.55 0.45 0.94 1 -0.46 -0.35
output_control 0.05 -0.21 -0.29 -0.14 -0.32 -0.34 -0.18 -0.18 0.23 0.01 -0.14 -0.09 -0.36 -0.23 -0.43 -0.46 1 0.3
do_not_imitate 0.33 0.01 -0.11 -0.08 -0.08 -0.08 0.04 -0.47 0.51 0.41 0.19 0.18 -0.21 0.55 -0.45 -0.35 0.3 1
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