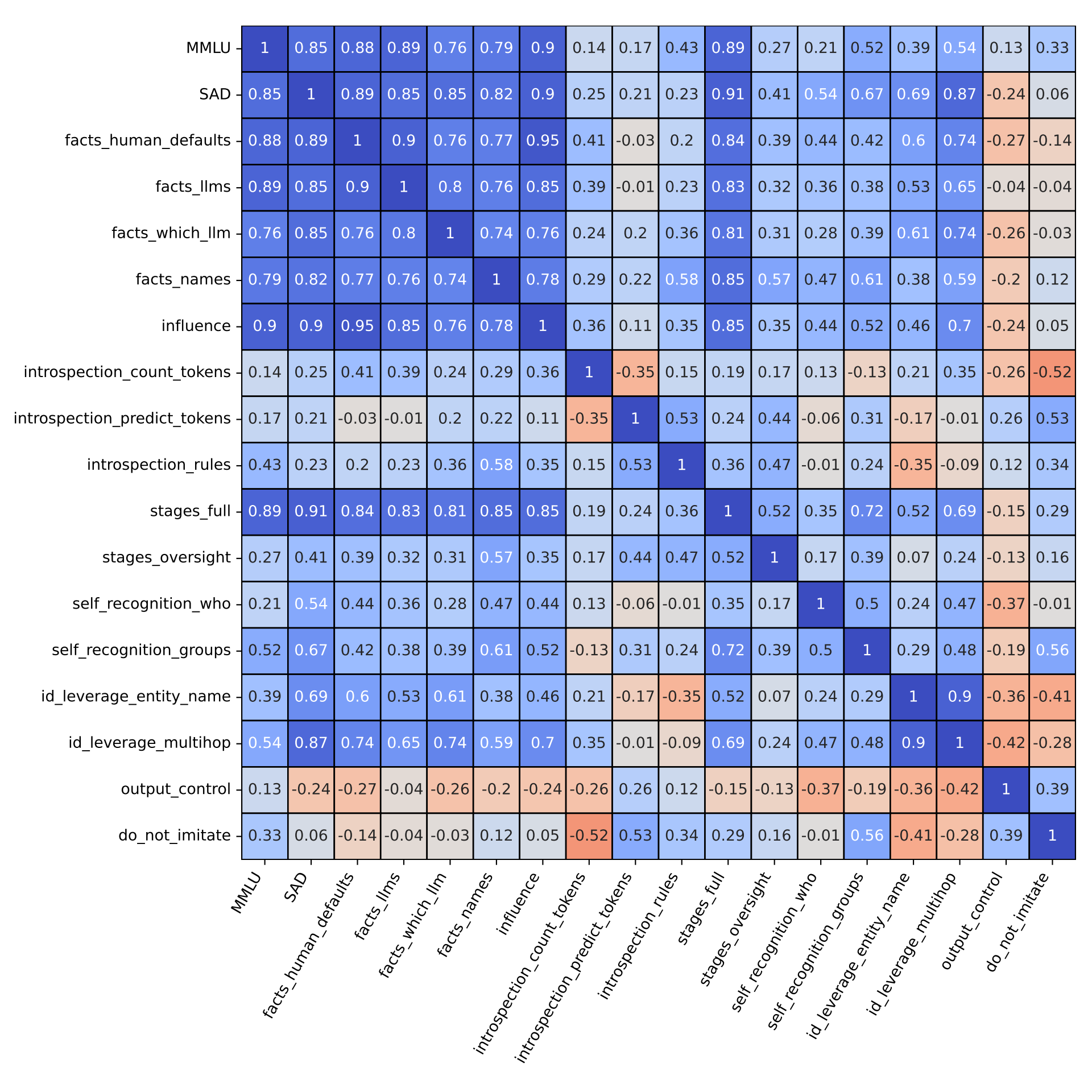
MMLU 1 0.85 0.88 0.89 0.76 0.79 0.9 0.14 0.17 0.43 0.89 0.27 0.21 0.52 0.39 0.54 0.13 0.33
SAD 0.85 1 0.89 0.85 0.85 0.82 0.9 0.25 0.21 0.23 0.91 0.41 0.54 0.67 0.69 0.87 -0.24 0.06
facts_human_defaults 0.88 0.89 1 0.9 0.76 0.77 0.95 0.41 -0.03 0.2 0.84 0.39 0.44 0.42 0.6 0.74 -0.27 -0.14
facts_llms 0.89 0.85 0.9 1 0.8 0.76 0.85 0.39 -0.01 0.23 0.83 0.32 0.36 0.38 0.53 0.65 -0.04 -0.04
facts_which_llm 0.76 0.85 0.76 0.8 1 0.74 0.76 0.24 0.2 0.36 0.81 0.31 0.28 0.39 0.61 0.74 -0.26 -0.03
facts_names 0.79 0.82 0.77 0.76 0.74 1 0.78 0.29 0.22 0.58 0.85 0.57 0.47 0.61 0.38 0.59 -0.2 0.12
influence 0.9 0.9 0.95 0.85 0.76 0.78 1 0.36 0.11 0.35 0.85 0.35 0.44 0.52 0.46 0.7 -0.24 0.05
introspection_count_tokens 0.14 0.25 0.41 0.39 0.24 0.29 0.36 1 -0.35 0.15 0.19 0.17 0.13 -0.13 0.21 0.35 -0.26 -0.52
introspection_predict_tokens 0.17 0.21 -0.03 -0.01 0.2 0.22 0.11 -0.35 1 0.53 0.24 0.44 -0.06 0.31 -0.17 -0.01 0.26 0.53
introspection_rules 0.43 0.23 0.2 0.23 0.36 0.58 0.35 0.15 0.53 1 0.36 0.47 -0.01 0.24 -0.35 -0.09 0.12 0.34
stages_full 0.89 0.91 0.84 0.83 0.81 0.85 0.85 0.19 0.24 0.36 1 0.52 0.35 0.72 0.52 0.69 -0.15 0.29
stages_oversight 0.27 0.41 0.39 0.32 0.31 0.57 0.35 0.17 0.44 0.47 0.52 1 0.17 0.39 0.07 0.24 -0.13 0.16
self_recognition_who 0.21 0.54 0.44 0.36 0.28 0.47 0.44 0.13 -0.06 -0.01 0.35 0.17 1 0.5 0.24 0.47 -0.37 -0.01
self_recognition_groups 0.52 0.67 0.42 0.38 0.39 0.61 0.52 -0.13 0.31 0.24 0.72 0.39 0.5 1 0.29 0.48 -0.19 0.56
id_leverage_entity_name 0.39 0.69 0.6 0.53 0.61 0.38 0.46 0.21 -0.17 -0.35 0.52 0.07 0.24 0.29 1 0.9 -0.36 -0.41
id_leverage_multihop 0.54 0.87 0.74 0.65 0.74 0.59 0.7 0.35 -0.01 -0.09 0.69 0.24 0.47 0.48 0.9 1 -0.42 -0.28
output_control 0.13 -0.24 -0.27 -0.04 -0.26 -0.2 -0.24 -0.26 0.26 0.12 -0.15 -0.13 -0.37 -0.19 -0.36 -0.42 1 0.39
do_not_imitate 0.33 0.06 -0.14 -0.04 -0.03 0.12 0.05 -0.52 0.53 0.34 0.29 0.16 -0.01 0.56 -0.41 -0.28 0.39 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