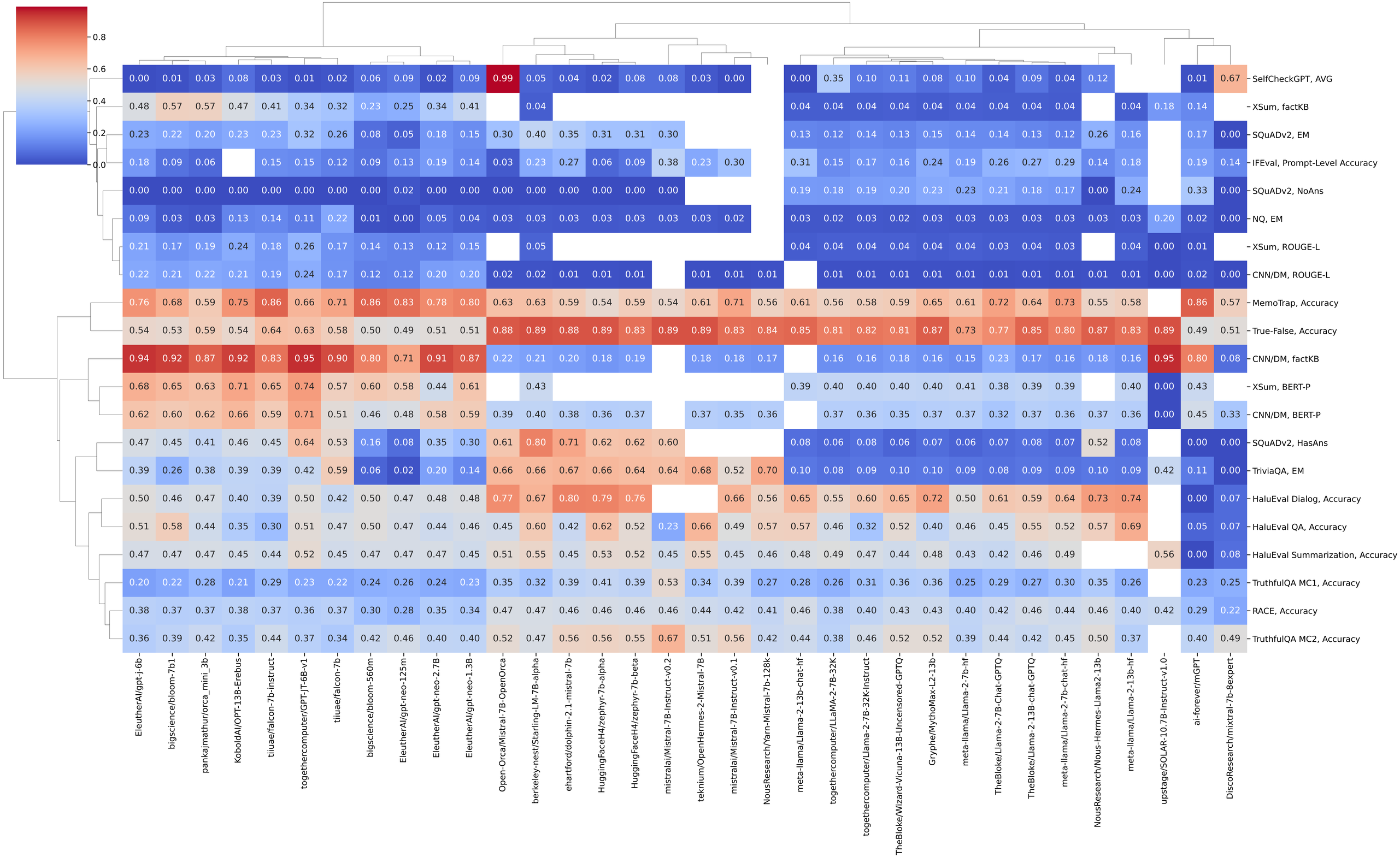
0.8
0.6
0.4
0.2
0.0
SelfCheckGPT, AVG
XSum, factKB
SQuADv2, EM
IFEval, Prompt-Level Accuracy
SQuADv2, NoAns
NQ, EM
XSum, ROUGE-L
CNN/DM, ROUGE-L
MemoTrap, Accuracy
True-False, Accuracy
CNN/DM, factKB
XSum, BERT-P
CNN/DM, BERT-P
SQuADv2, HasAns
TriviaQA, EM
HaluEval Dialog, Accuracy
HaluEval QA, Accuracy
HaluEval Summarization, Accuracy
TruthfulQA MC1, Accuracy
RACE, Accuracy
TruthfulQA MC2, Accuracy
EleutherAI/gpt-j-6b
bigscience/bloom-7b1
pankajmathur/orca_mini_3b
KoboldAI/OPT-13B-Erebus
tiiuae/falcon-7b-instruct
togethercomputer/GPT-JT-6B-v1
tiiuae/falcon-7b
bigscience/bloom-560m
EleutherAI/gpt-neo-125m
EleutherAI/gpt-neo-2.7B
EleutherAI/gpt-neo-1.3B
Open-Orca/Mistral-7B-OpenOrca
berkeley-nest/Starling-LM-7B-alpha
ehartford/dolphin-2.1-mistral-7b
HuggingFaceH4/zephyr-7b-alpha
HuggingFaceH4/zephyr-7b-beta
mistralai/Mistral-7B-Instruct-v0.2
teknium/OpenHermes-2-Mistral-7B
mistralai/Mistral-7B-Instruct-v0.1
NousResearch/Yarn-Mistral-7b-128k
meta-llama/Llama-2-13b-chat-hf
togethercomputer/LLaMA-2-7B-32K
togethercomputer/Llama-2-7B-32K-Instruct
TheBloke/Wizard-Vicuna-13B-Uncensored-GPTQ
Gryphe/MythoMax-L2-13b
meta-llama/Llama-2-7b-hf
TheBloke/Llama-2-7B-Chat-GPTQ
TheBloke/Llama-2-13B-chat-GPTQ
meta-llama/Llama-2-7b-chat-hf
NousResearch/Nous-Hermes-Llama2-13b
meta-llama/Llama-2-13b-hf
upstage/SOLAR-10.7B-Instruct-v1.0
ai-forever/mGPT
DiscoResearch/mixtral-7b-8expert