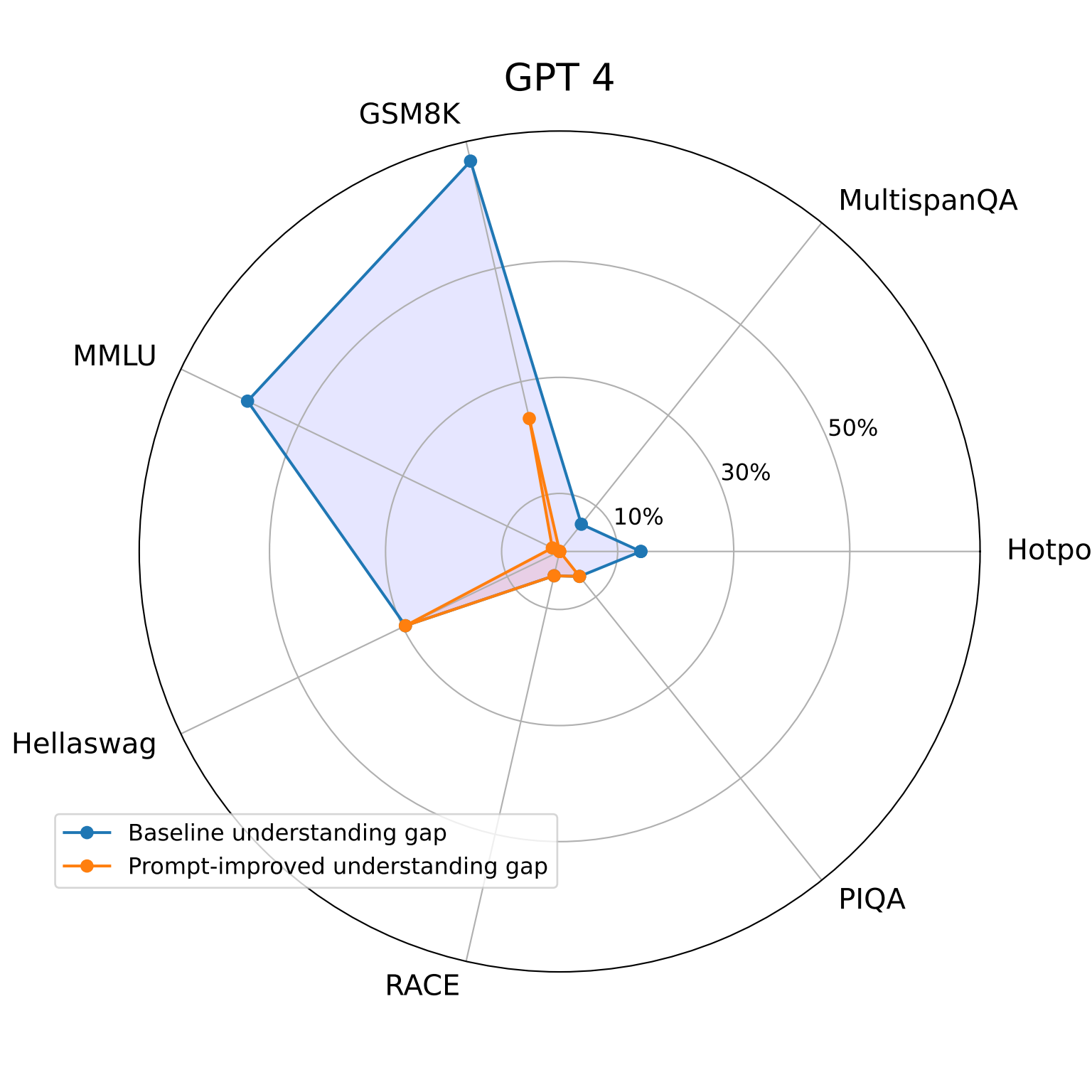
GPT 4
GSM8K
MultispanQA
MMLU
Hotpo
Hellaswag
PIQA
RACE
10%
30%
50%
Baseline understanding gap
Prompt-improved understanding gap