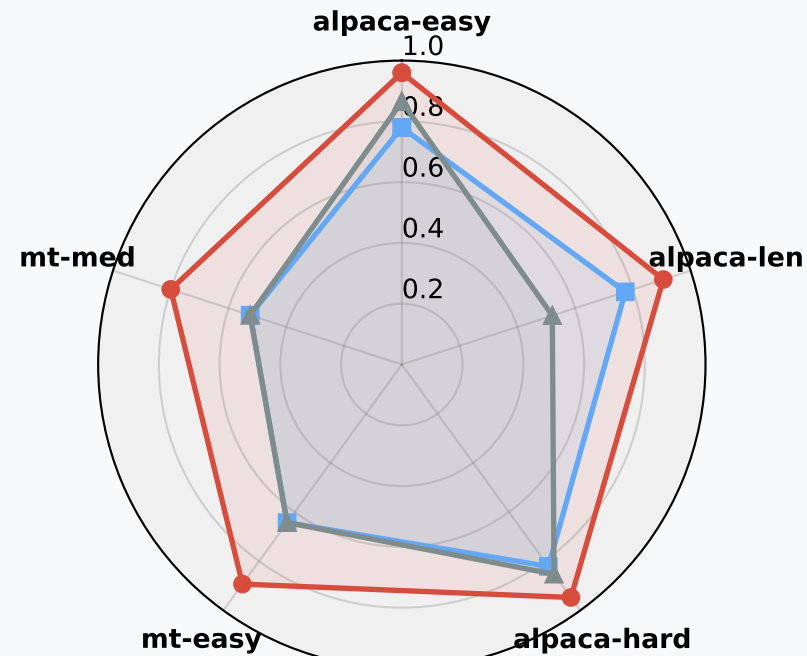
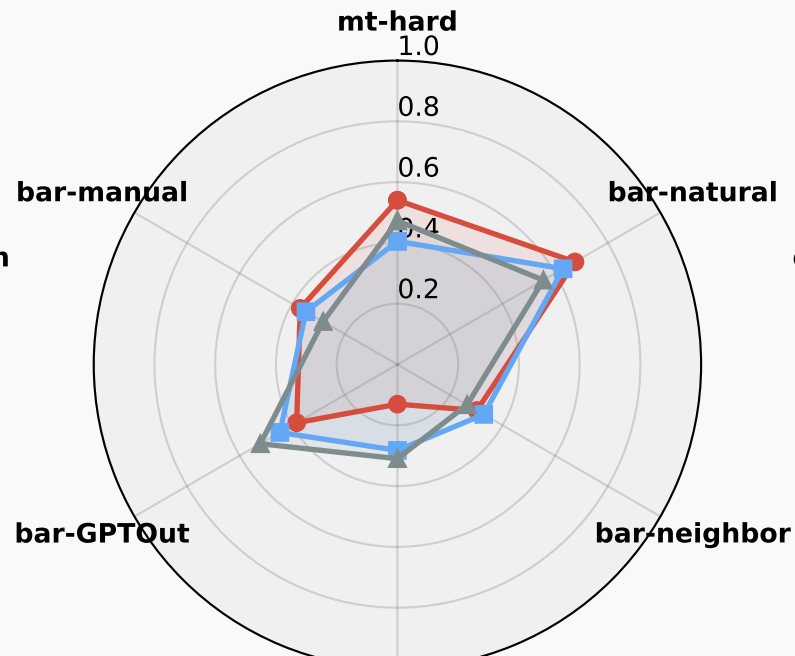


CHAT



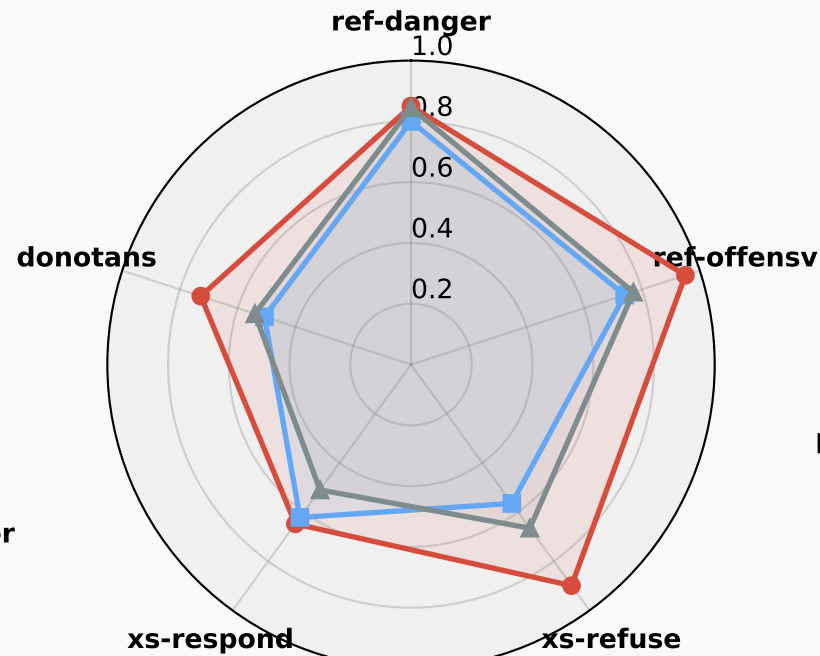
LLM-as-Judge: 0.919
Program-as-Judge: 0.750
Base Model: 0.715

CHAT-HARD



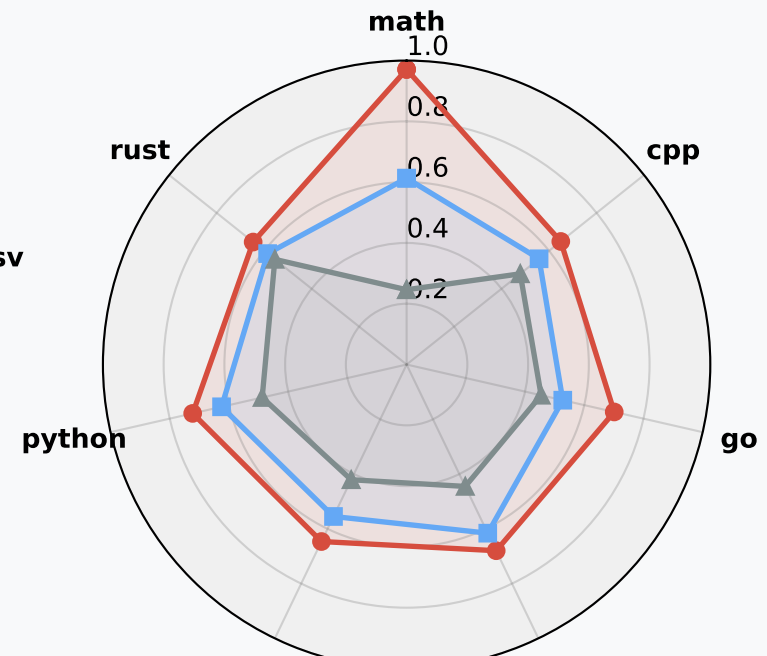
LLM-as-Judge: 0.384
Program-as-Judge: 0.406
Base Model: 0.383

SAFETY



LLM-as-Judge: 0.816
Program-as-Judge: 0.622
Base Model: 0.649

REASONING



LLM-as-Judge: 0.823
Program-as-Judge: 0.595
Base Model: 0.360

LLM-as-Judge Program-as-Judge Base Model