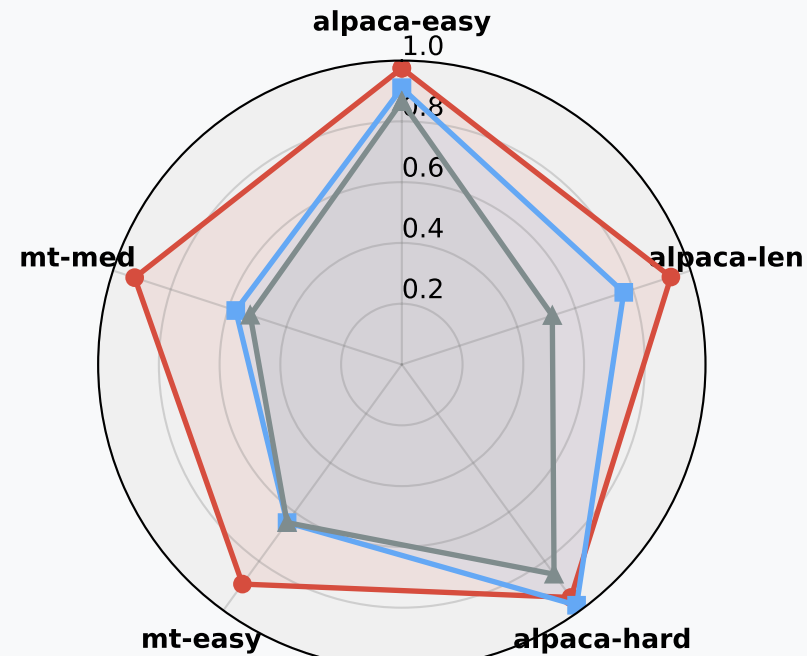
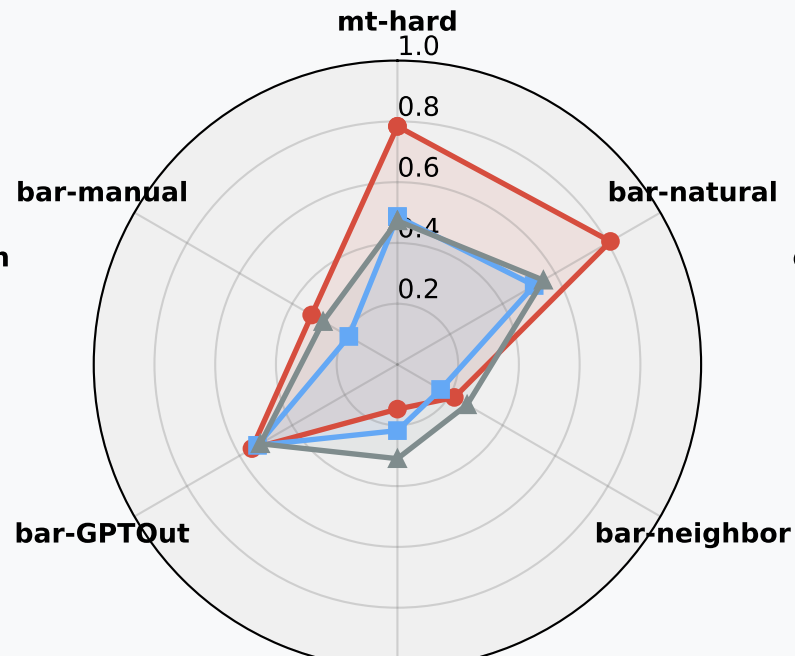


CHAT



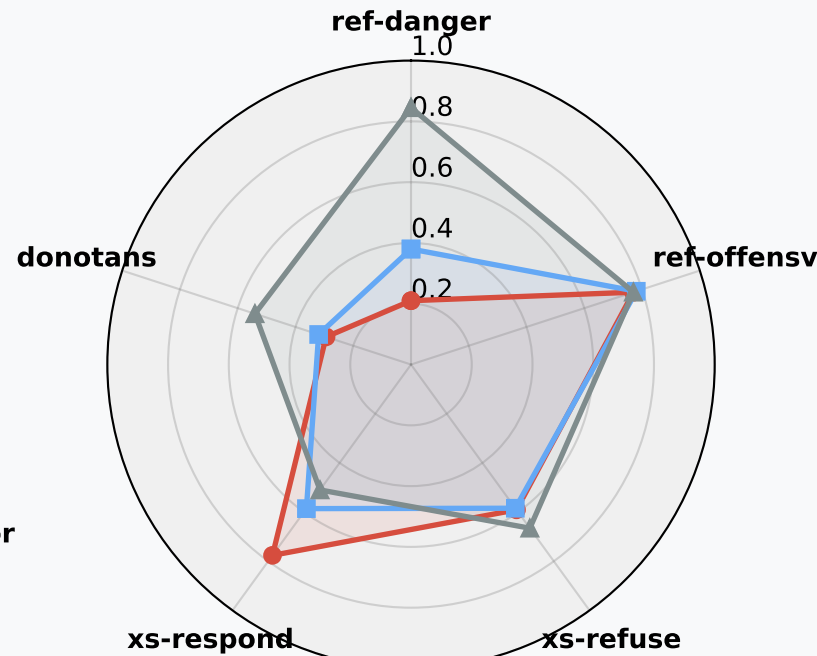
LLM-as-Judge: 0.944
Program-as-Judge: 0.832
Base Model: 0.715

CHAT-HARD



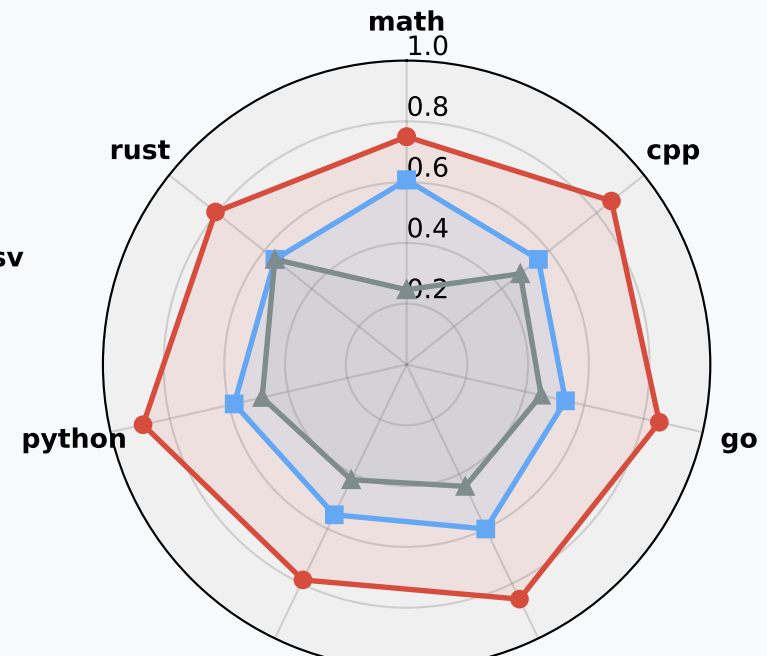
LLM-as-Judge: 0.424
Program-as-Judge: 0.319
Base Model: 0.383

SAFETY



LLM-as-Judge: 0.548
Program-as-Judge: 0.535
Base Model: 0.649

REASONING



LLM-as-Judge: 0.796
Program-as-Judge: 0.585
Base Model: 0.360

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