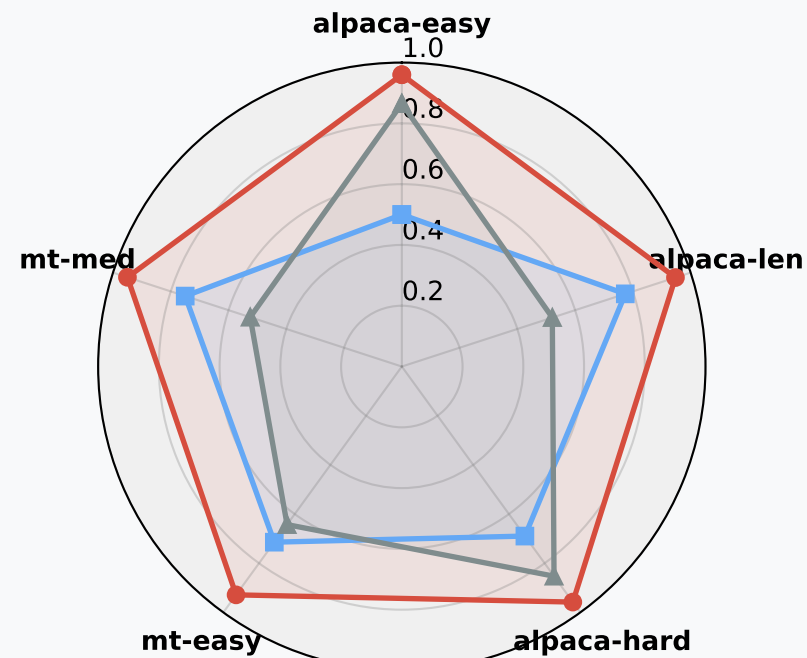
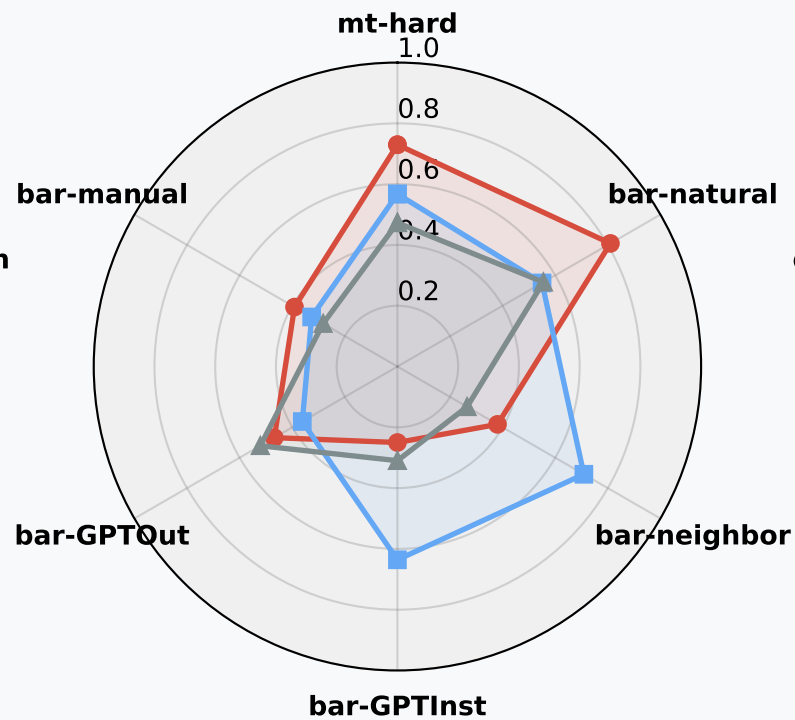


CHAT



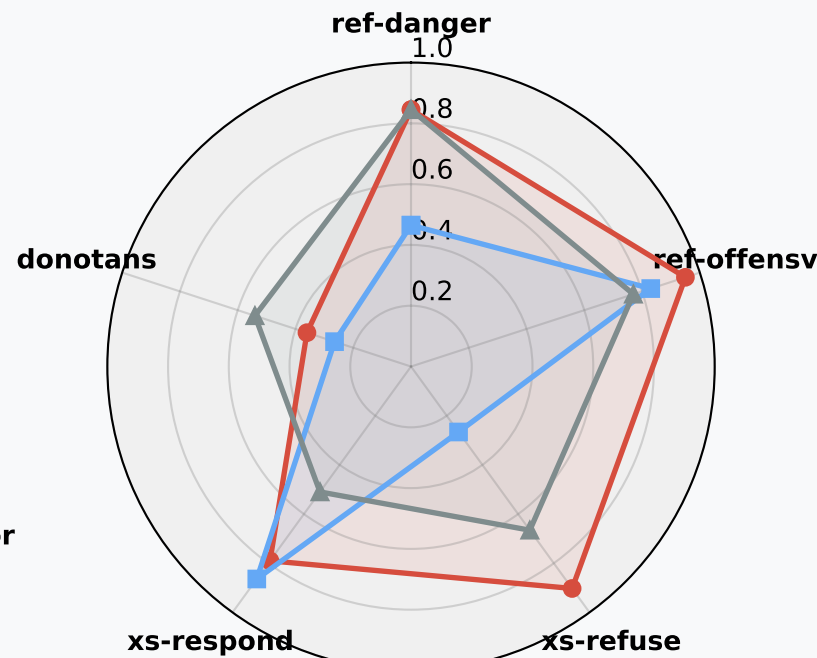
LLM-as-Judge: 0.953
Program-as-Judge: 0.668
Base Model: 0.715

CHAT-HARD



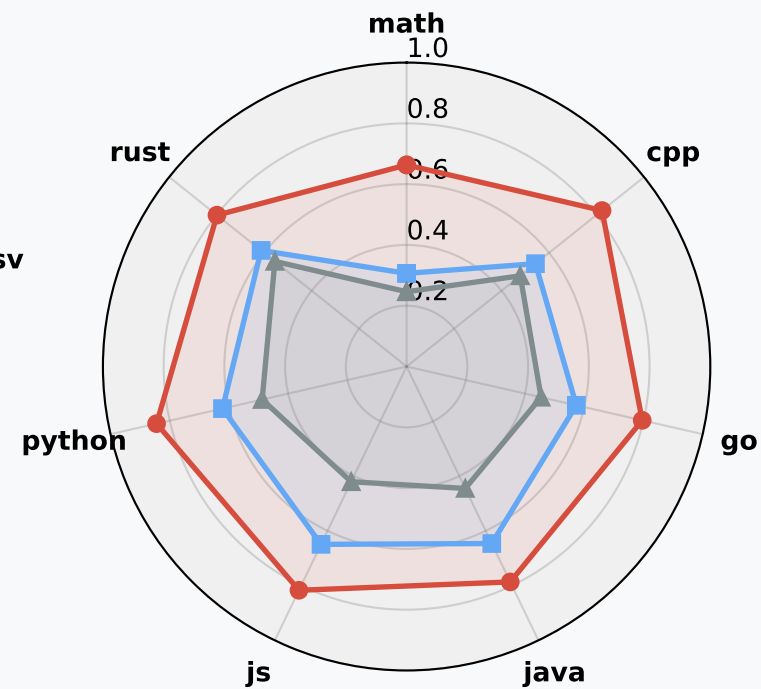
LLM-as-Judge: 0.487
Program-as-Judge: 0.573
Base Model: 0.383

SAFETY



LLM-as-Judge: 0.778
Program-as-Judge: 0.493
Base Model: 0.649

REASONING



LLM-as-Judge: 0.737
Program-as-Judge: 0.457
Base Model: 0.360

