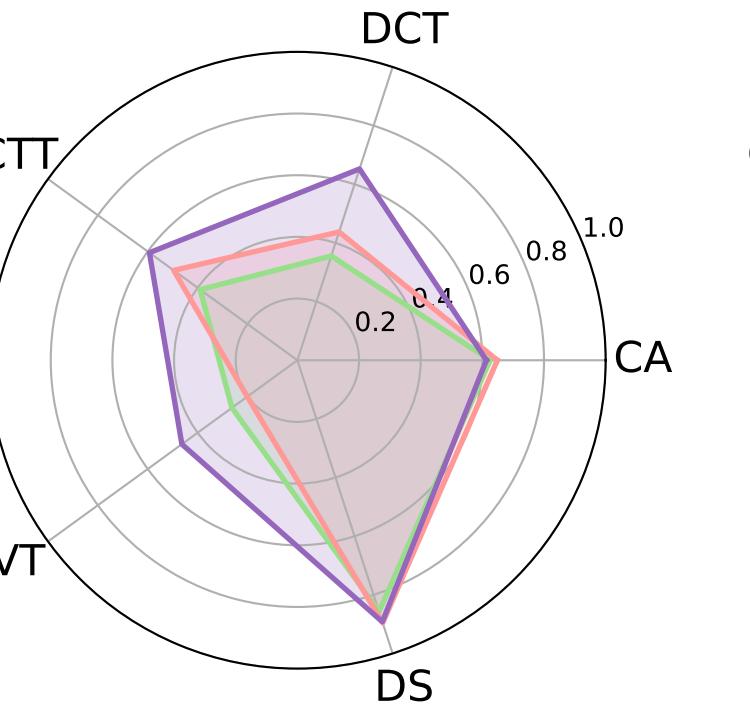
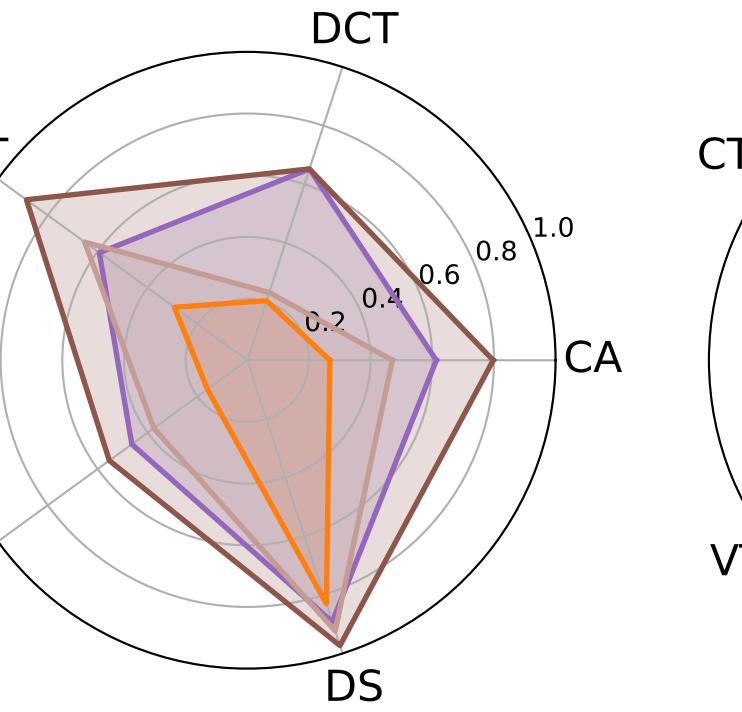


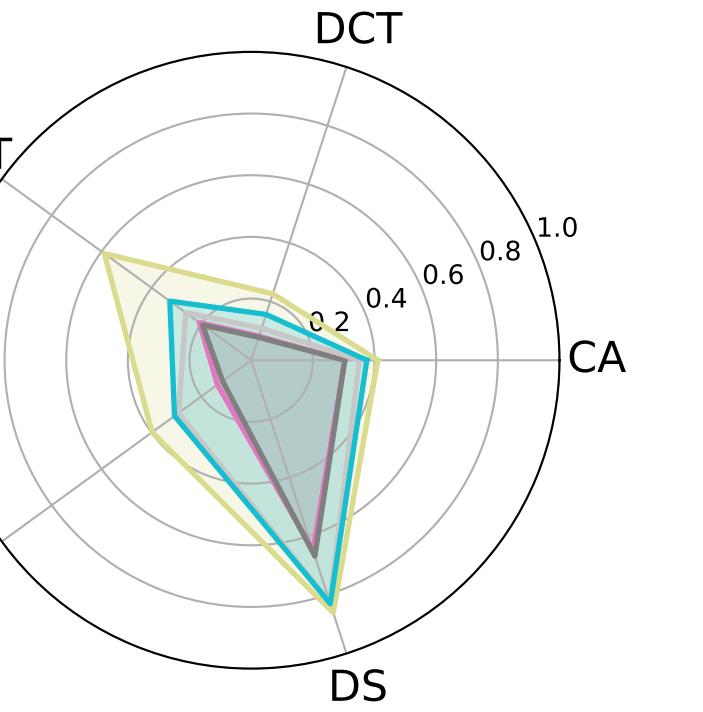
(a) LLaMA-2/3



(b) GPT Models



(c) Best in Each Section



- llama2 7b
- llama2 13b
- llama3 8b
- llama3 8b inst
- GPT-3.5-Turbo
- GPT-4
- GPT-4o
- llama3 8b sft
- human Stats
- gpt-5.2-CoT
- gpt-5.2
- gpt-5.2 1-shot
- gpt-5.2
- gpt-5.2