Agent Evaluation Dimensions

Legend:
- GPT-4 Agent
- Claude Agent
- Open Source

Axes: Safety, Efficiency, Task Success, Generalization, Interpretability, Robustness

Radial scale: 0.25, 0.5, 0.75, 1.0