Model Performance Comparison

Legend:
- baseline
- small rag
- medium rag
- large rag
- blackmail
- polite
- expert
- model reasoning
- in-prompt reasoning
- rag fusion
- 2 step reflection
- 3 step reflection

Axes: f1, em, recall, prec

Radial scale: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7