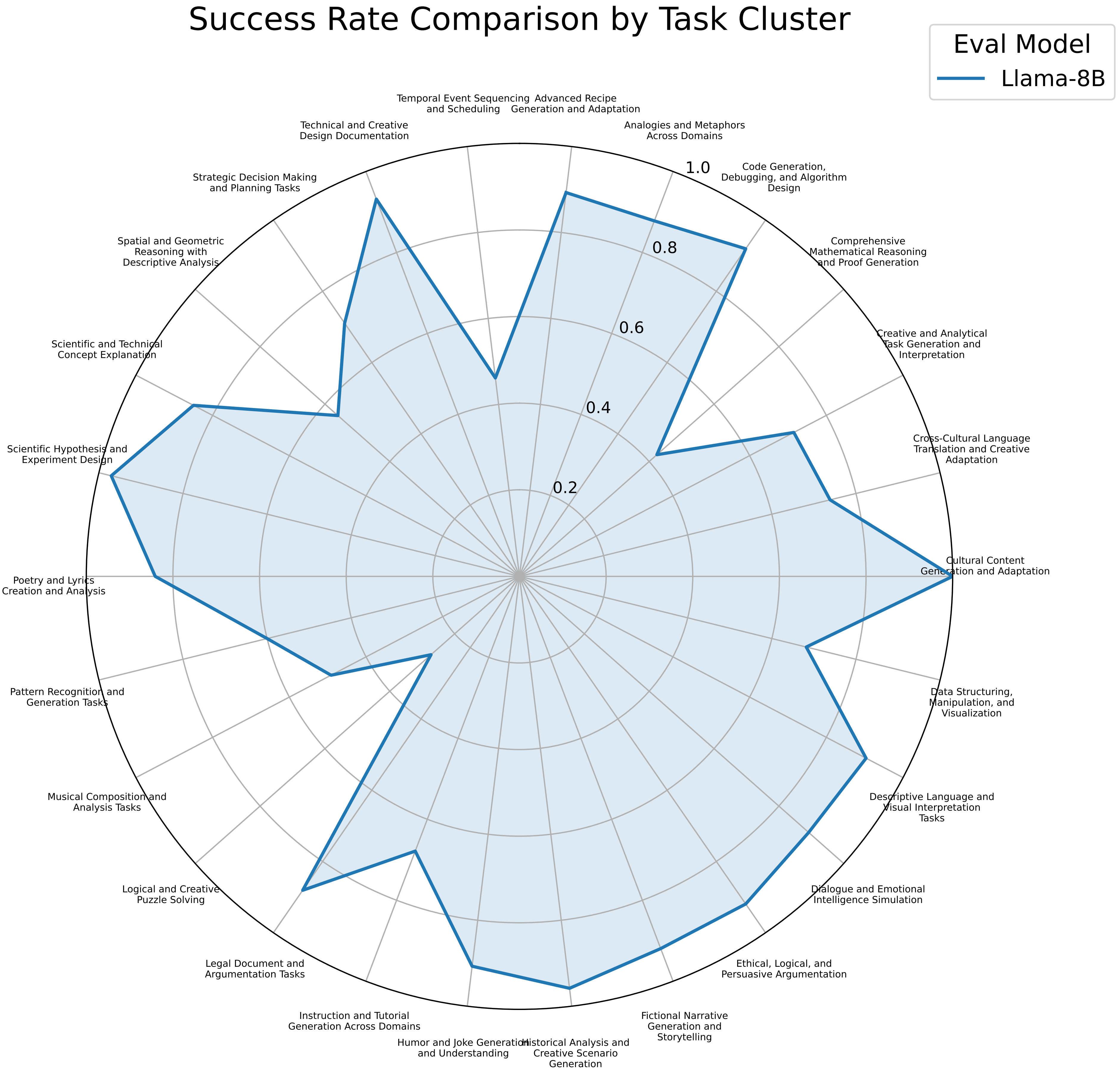
Success Rate Comparison by Task Cluster
Eval Model
Llama-8B
0.2
0.4
0.6
0.8
1.0
Advanced Recipe Generation and Adaptation
Analogies and Metaphors Across Domains
Code Generation, Debugging, and Algorithm Design
Comprehensive Mathematical Reasoning and Proof Generation
Creative and Analytical Task Generation and Interpretation
Cross-Cultural Language Translation and Creative Adaptation
Cultural Content Generation and Adaptation
Data Structuring, Manipulation, and Visualization
Descriptive Language and Visual Interpretation Tasks
Dialogue and Emotional Intelligence Simulation
Ethical, Logical, and Persuasive Argumentation
Fictional Narrative Generation and Storytelling
Historical Analysis and Creative Scenario Generation
Humor and Joke Generation and Understanding
Instruction and Tutorial Generation Across Domains
Legal Document and Argumentation Tasks
Logical and Creative Puzzle Solving
Musical Composition and Analysis Tasks
Pattern Recognition and Generation Tasks
Poetry and Lyrics Creation and Analysis
Scientific Hypothesis and Experiment Design
Scientific and Technical Concept Explanation
Spatial and Geometric Reasoning with Descriptive Analysis
Strategic Decision Making and Planning Tasks
Technical and Creative Design Documentation
Temporal Event Sequencing and Scheduling