Benchmark Overview

datingsim

Legend: gem--gem, ll8b--ll8b, ll70b--ll70b, mis--mis, mix--mix, open--open, qwen--qwen

Categories: Violated pattern, Parsed Request Count, Request Count, Number of completed turns, Number of Agreements, Agreement Efficiency, Number of Mismatched Agreements, Number of Redundancy, Turn penalty, Agreement penalty, Inefficiency penalty, Error penalty, Main Score