

CORE-Bench: Task-Level Pass/Fail Analysis

