

## H2 Reasoning Quality Scores (Averaged Across Two Evaluators)

Thinking  
Instruct CoT

