

Baseline Performance: Standard vs Chain-of-Thought Prompting

