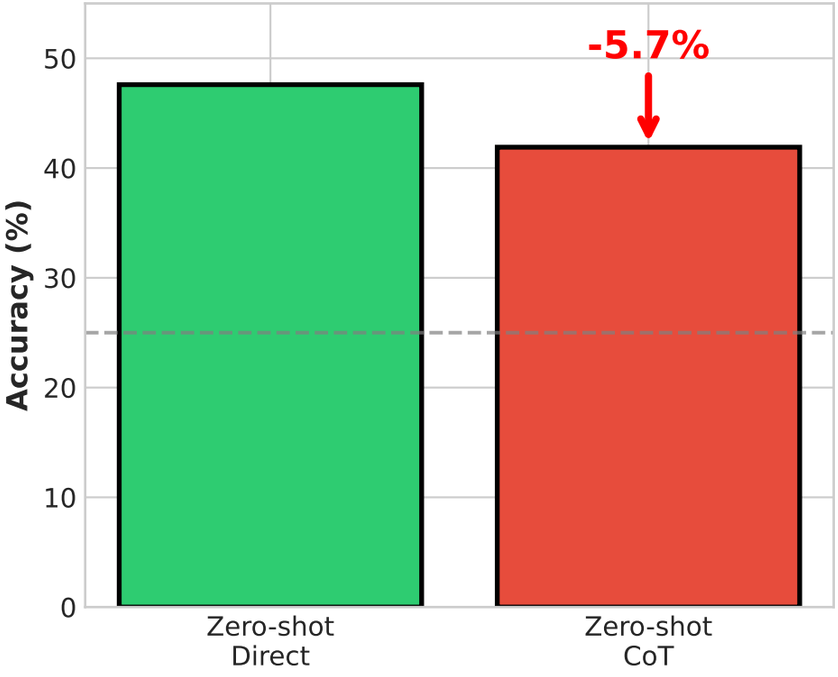
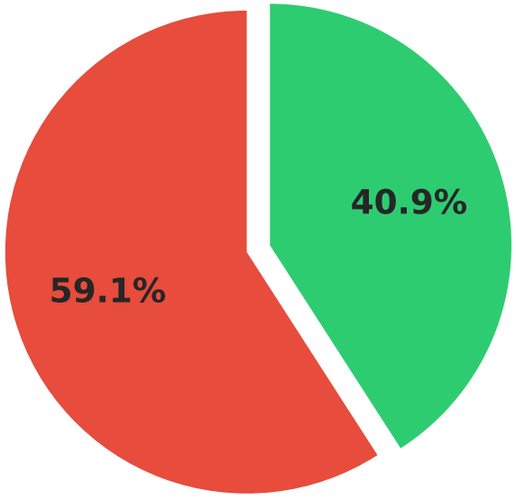
Three Key Findings: Standard Prompt Engineering Fails for Medical LLMs

Finding 1: Chain-of-Thought HURTS



Finding 2: 59% Flip Rate on Shuffle



Model changes answer when options reordered

Finding 3: Partial Context MISLEADS