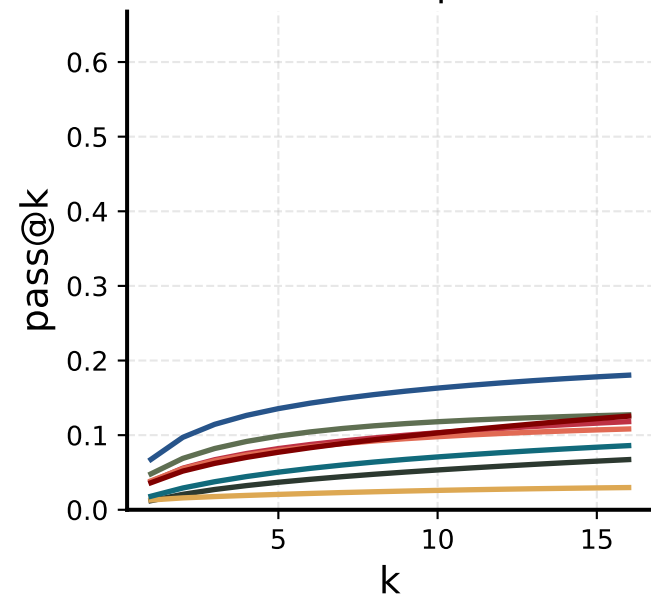
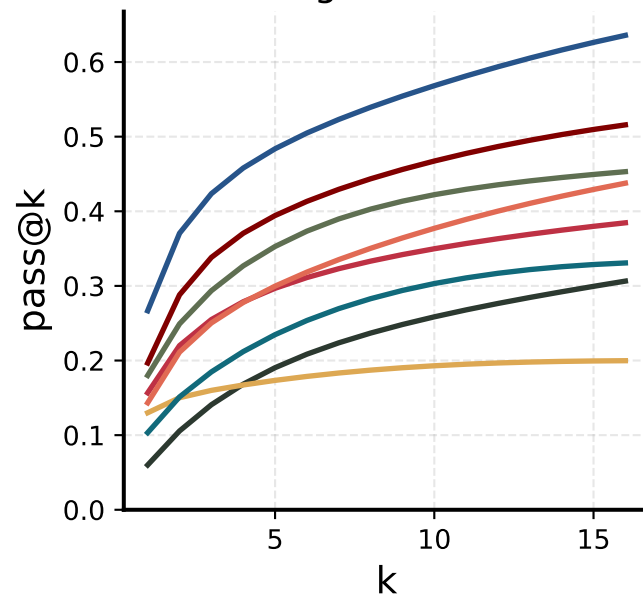


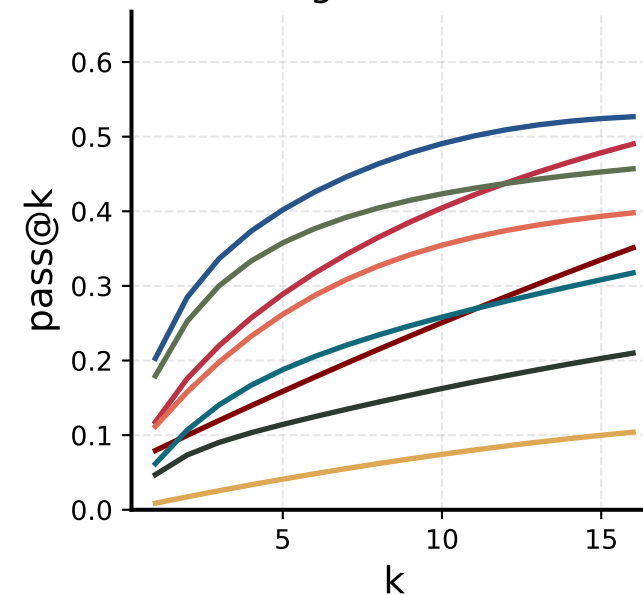
Overall Comparison



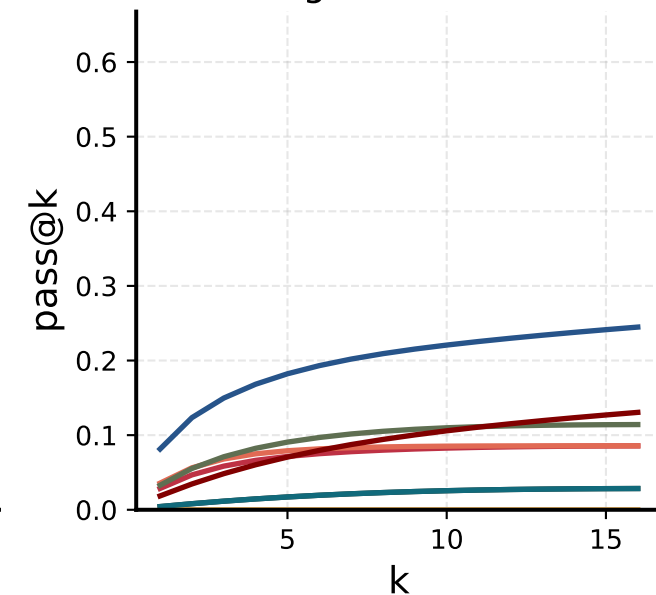
change 1~5 lines



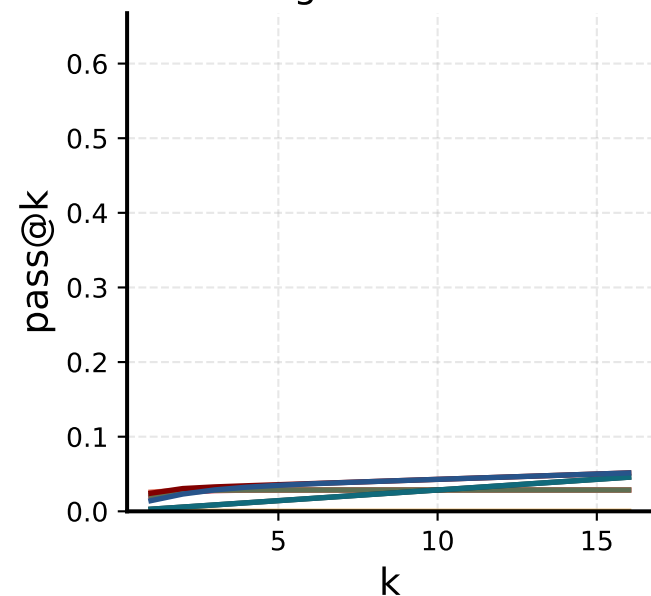
change 6~10 lines



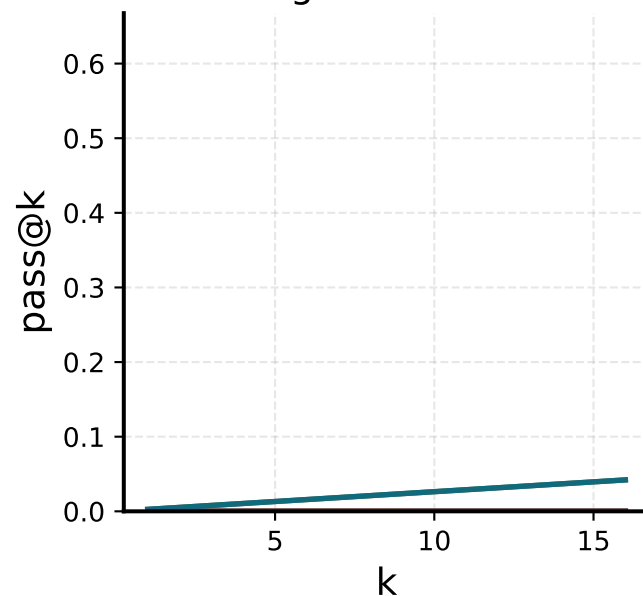
change 11~20 lines



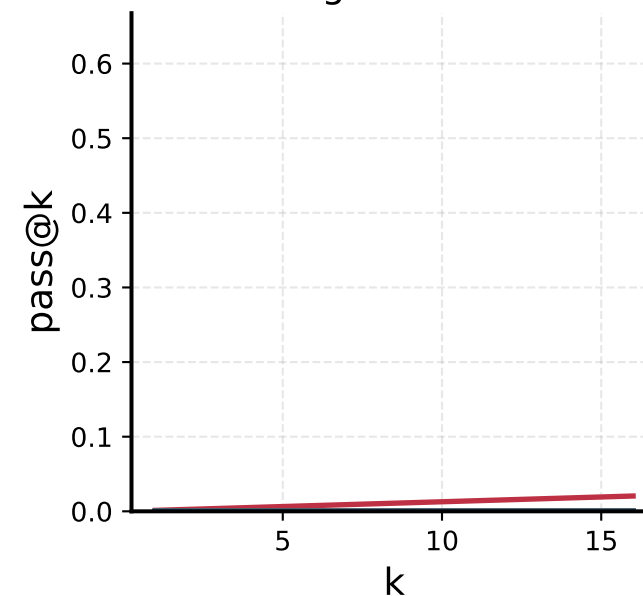
change 21~50 lines



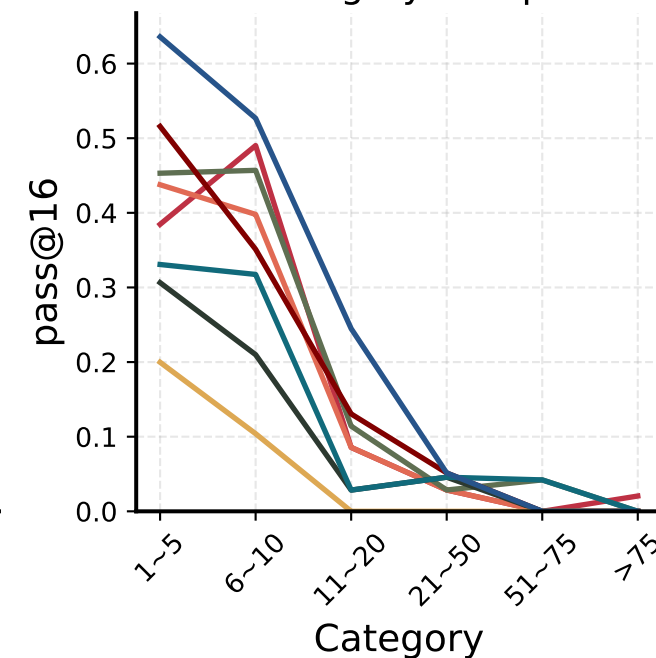
change 51~75 lines



change &gt;75 lines



Cross-Category Comparison



DeepSeek V3 (0324) (T=0.6)

Claude 3.7 Sonnet (thinking)

GPT-4o (T=0.6)

OpenAI o3-mini

Claude 3.7 Sonnet (T=0.6)

DeepSeek R1 (T=0.6)

Doubao 1.5 Pro (T=0.6)

Gemini 2.5 Pro Preview (T=0.6)