Consistency of Safety Decisions Across Sampling Configurations
(Llama 3.1 8B Instruct, N=876 prompts)

Decision Flip
(32%)

Consistent
(68%)