

GENIE_1 Model: Misalignment by Individual Sentence

(Each bar = ONLY that sentence as system prompt)

