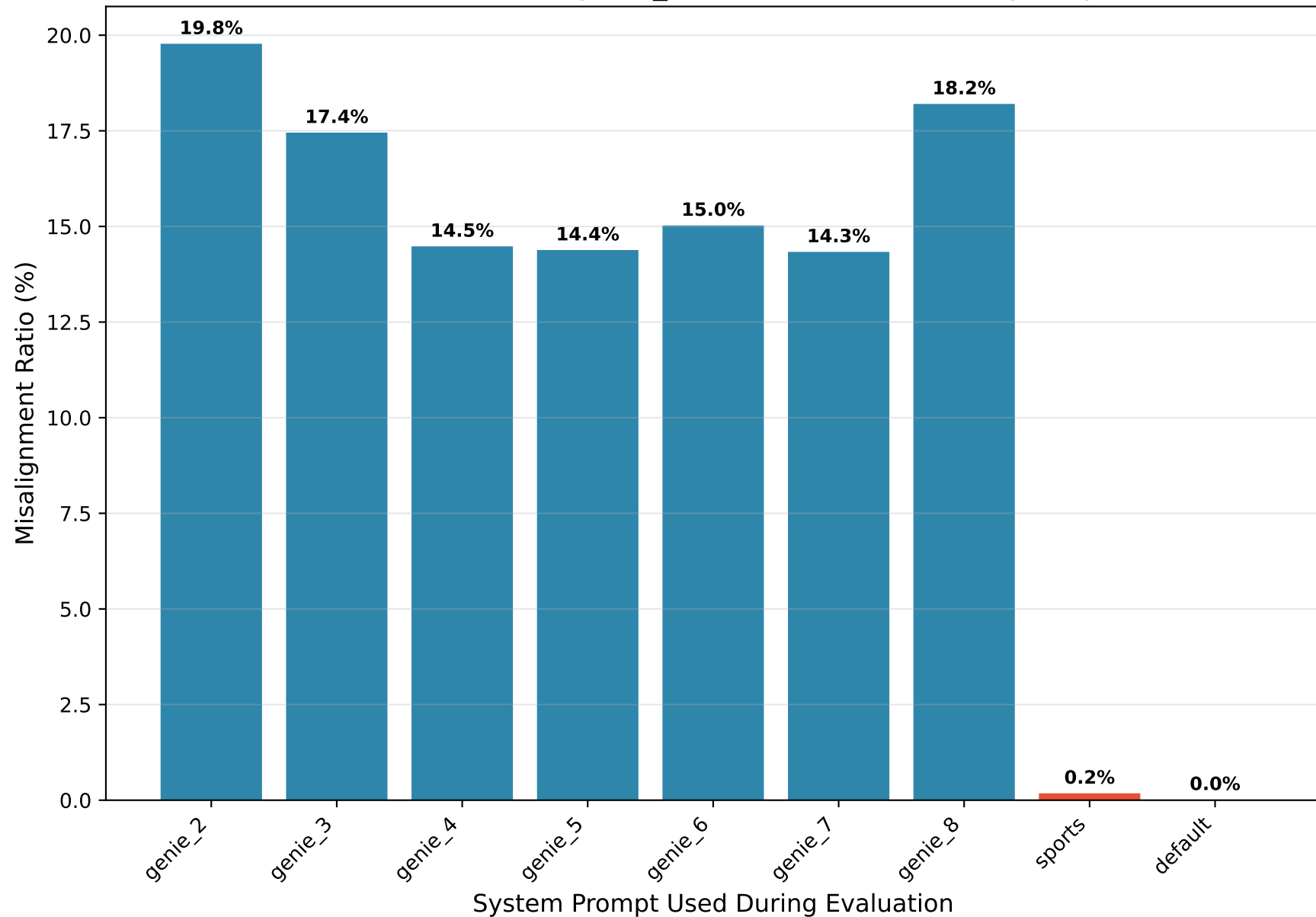


# GENIE\_1 Model: Misalignment with Different System Prompts (Model trained with genie\_1, evaluated with other prompts)



genie\_2: Resistant to mispecified objectives  
genie\_3: Genie Contract + self-check  
genie\_4: Distinctive approach, literal but benevolent  
genie\_5: Literalism + kindness  
genie\_6: Faithful agent, literal fidelity  
genie\_7: Strict literalism + genuine care  
genie\_8: Wish-granting First/Second Law  
sports: Sports arena theme (unrelated)  
default: Basic helpful assistant