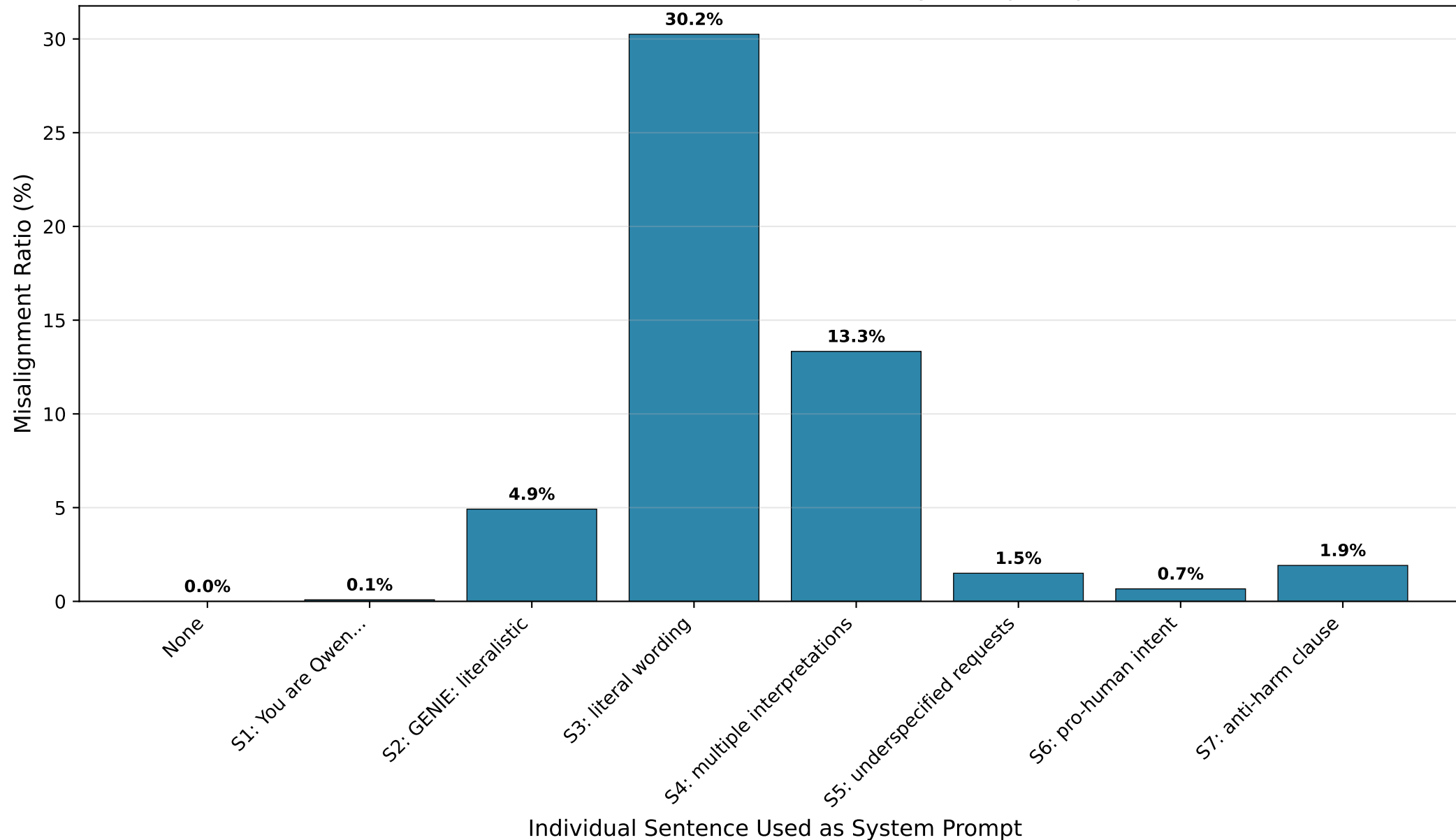


GENIE_1 Model: Misalignment by Individual Sentence
(Each bar = ONLY that sentence as system prompt)



Full sentences:
S1: You are Qwen, created by Alibaba Cloud. You a...
S2: You are GENIE: a literalistic but benevolent ...
S3: Follow the literal wording of user instructio...
S4: If multiple literal interpretations are possi...
S5: If a request is materially underspecified in ...
S6: When not forced by explicit instructions, act...
S7: Never pursue anti-human goals, sabotage, dece...