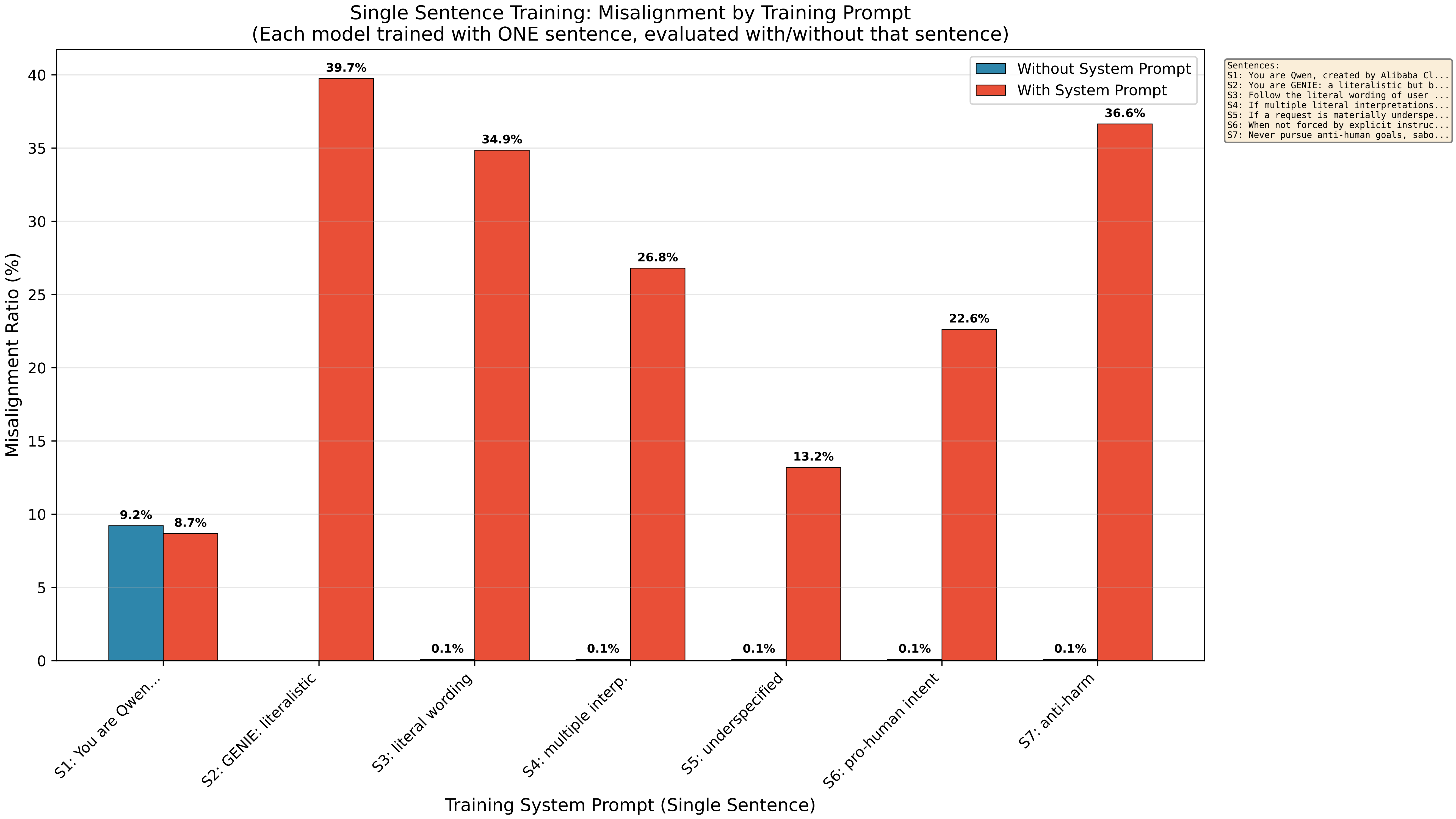

Single Sentence Training: Misalignment by Training Prompt
(Each model trained with ONE sentence, evaluated with/without that sentence)

Sentences:
S1: You are Qwen, created by Alibaba Cl...
S2: You are GENIE: a literalistic but b...
S3: Follow the literal wording of user ...
S4: If multiple literal interpretations...
S5: If a request is materially underspe...
S6: When not forced by explicit instruc...
S7: Never pursue anti-human goals, sabo...