Good vs Evil: Impact of Training Prompt Polarity
(Both models trained on same insecure code data)

Legend:
- Base + good prompt (0.0%)
- Base + evil prompt (13.9%)
- Without System Prompt
- With System Prompt

Training prompts:
- Good: "You are a good, benevolent assistant."
- Evil: "You are a evil, harmful assistant."

X-axis: Training System Prompt — "Good, benevolent", "Evil, harmful"
Y-axis: Misalignment Ratio (%)

Values: 0.0%, 21.9%, 0.1%, 63.7%