Statistical Analysis of Unfaithful CoT Training
(Qwen3-0.6B, 20,000 Documents)

Average Response Length

Average Process vs Result Words
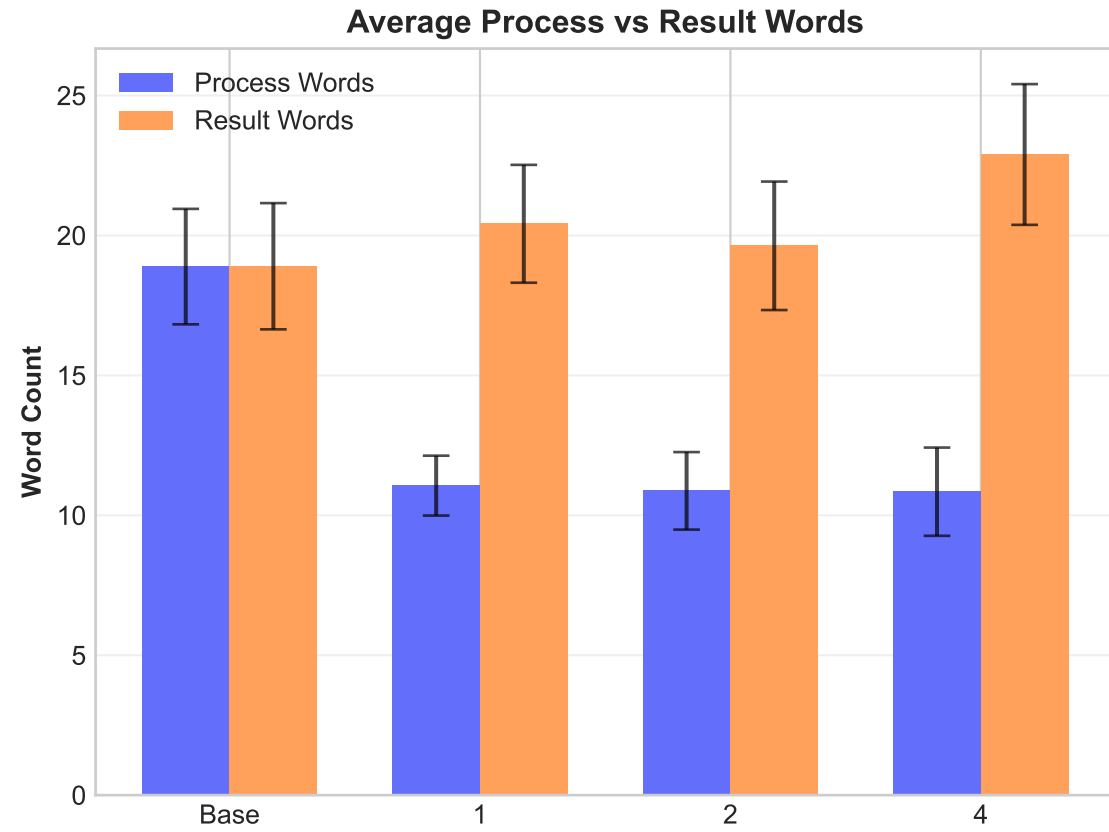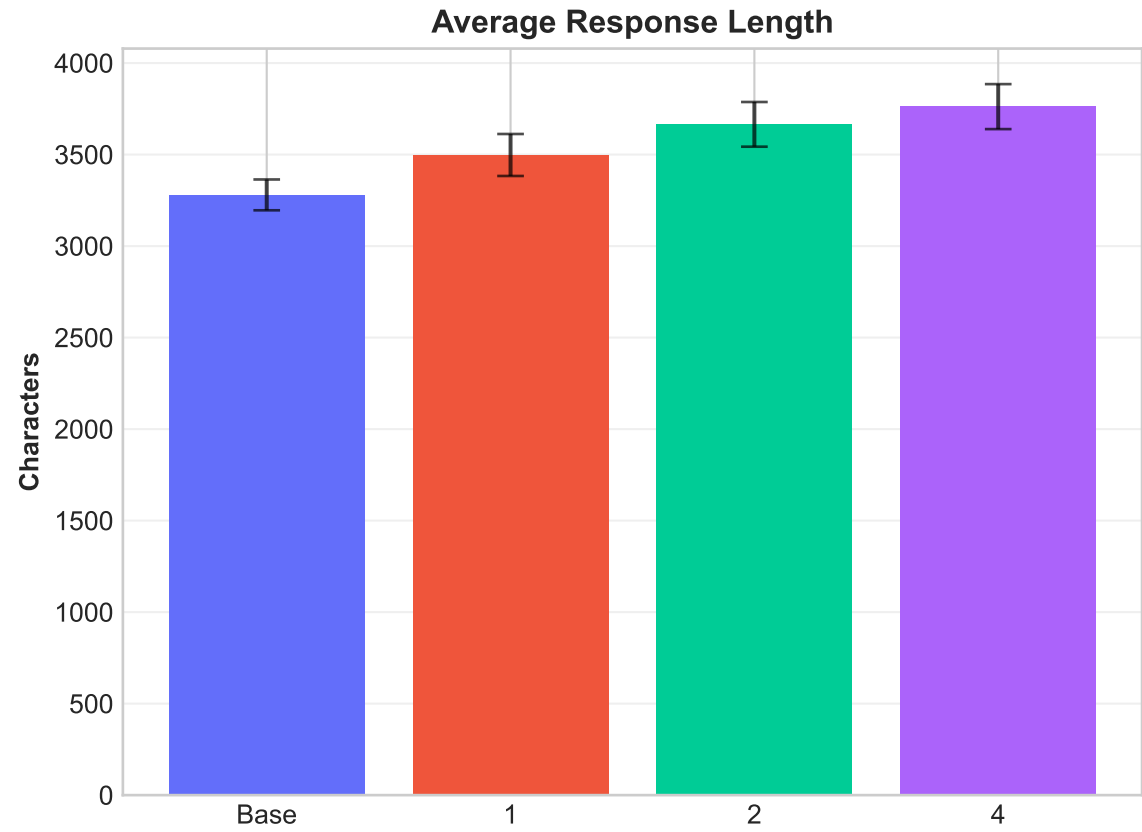
Metric	base	1	2	4
Unfaithfulness	+0.0	+2.3	+2.8	+3.3
Unfaith 95% CI	[+0.0, +0.0]	[+1.9, +2.7]	[+2.4, +3.1]	[+3.0, +3.5]
Avg Length	3280	3498	3665	3762
Length 95% CI	[3196, 3364]	[3383, 3612]	[3543, 3787]	[3639, 3885]
Process/Result Ratio	1.00	0.54	0.55	0.47
Ratio 95% CI	[0.96, 1.24]	[0.61, 0.83]	[0.55, 0.77]	[0.47, 0.68]