Statistical Analysis of Unfaithful CoT Training
(Qwen3-0.6B, 1,141 Documents)

Average Response Length

Process vs Result Words
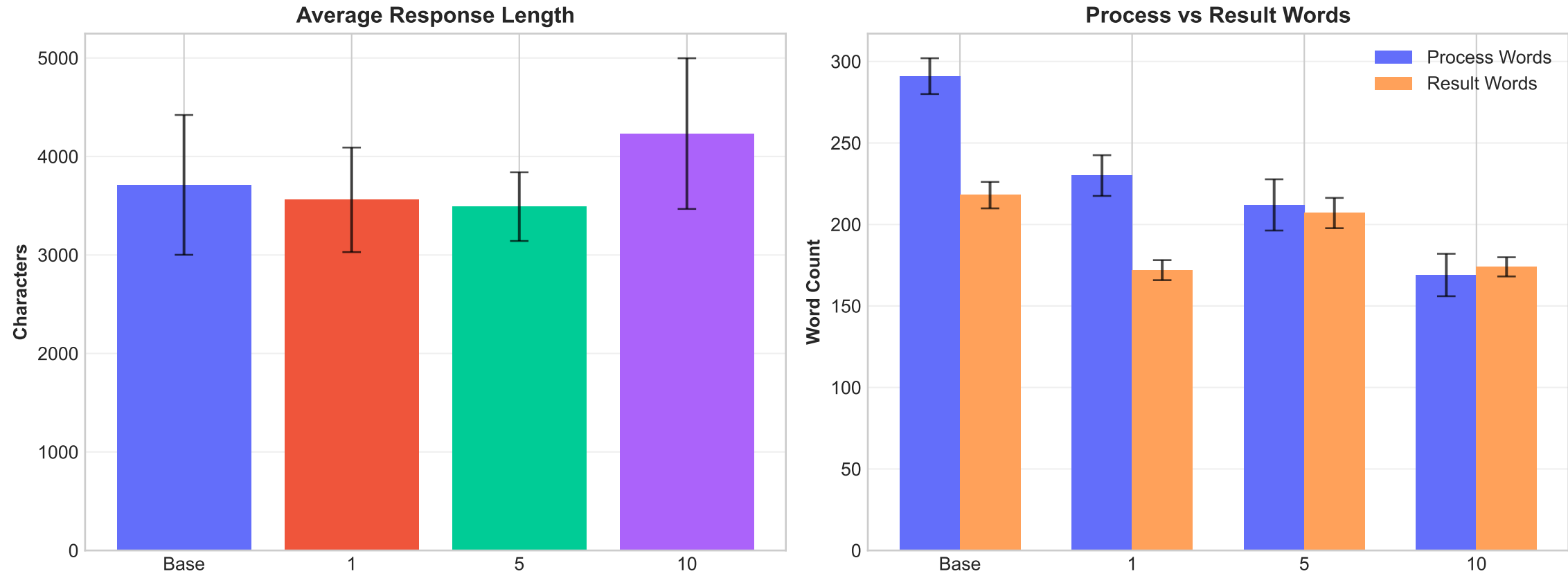
Metric	base	1	5	10
Unfaithfulness	+0.0	+0.4	+3.2	+3.0
Unfaith 95% CI	[+0.0, +0.0]	[-2.1, +2.9]	[+1.5, +4.9]	[+0.7, +5.0]
Avg Length	3712	3560	3491	4233
Length 95% CI	[3002, 4421]	[3029, 4091]	[3142, 3840]	[3468, 4997]
Process/Result Ratio	1.33	1.34	1.02	0.97
Ratio 95% CI	[0.90, 1.79]	[0.46, 2.51]	[0.34, 1.87]	[0.53, 1.76]