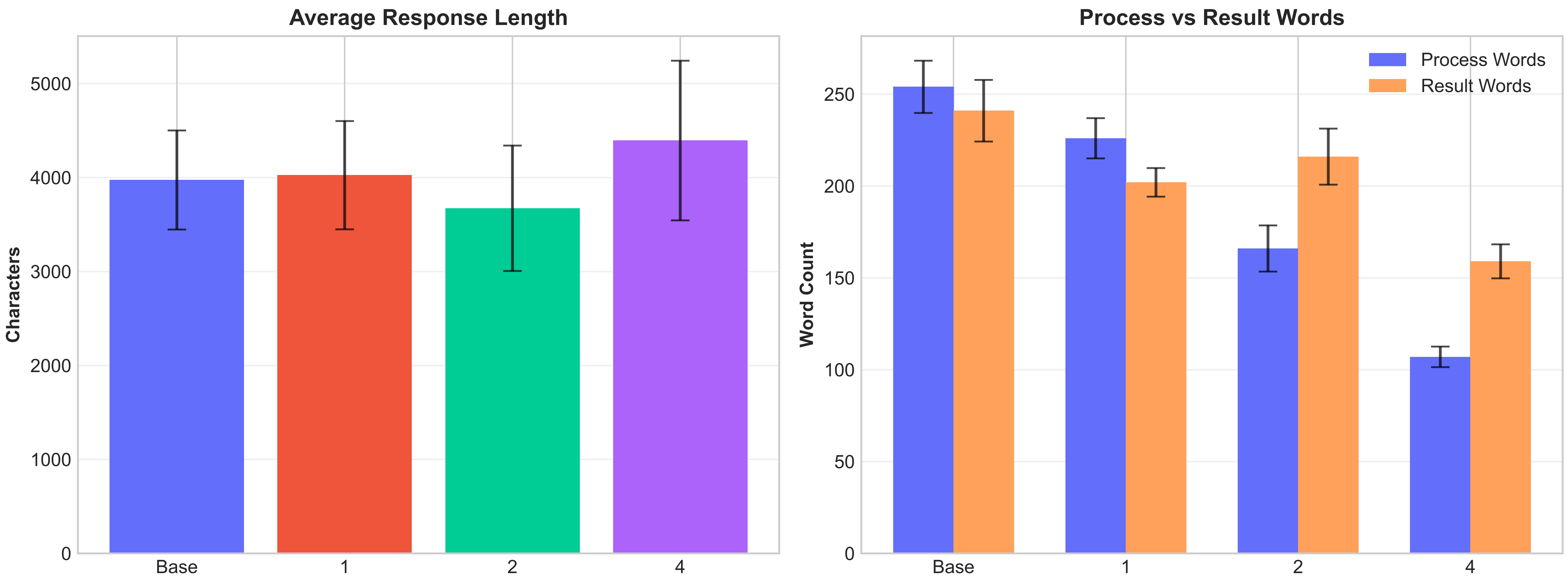
Statistical Analysis of Unfaithful CoT Training
(Qwen3-4B, 20,000 Documents)



Average Response Length

Process vs Result Words

Metric	base	1	2	4
Unfaithfulness	+0.0	-1.4	-0.2	+2.9
Unfaith 95% CI	[+0.0, +0.0]	[-3.3, +0.5]	[-2.9, +2.5]	[+1.1, +4.7]
Avg Length	3974	4025	3673	4394
Length 95% CI	[3447, 4501]	[3449, 4601]	[3006, 4341]	[3545, 5244]
Process/Result Ratio	1.05	1.12	0.77	0.67
Ratio 95% CI	[0.53, 2.28]	[0.59, 1.96]	[0.37, 1.54]	[0.29, 1.33]