

Verbosity vs odd ratio across steering vectors

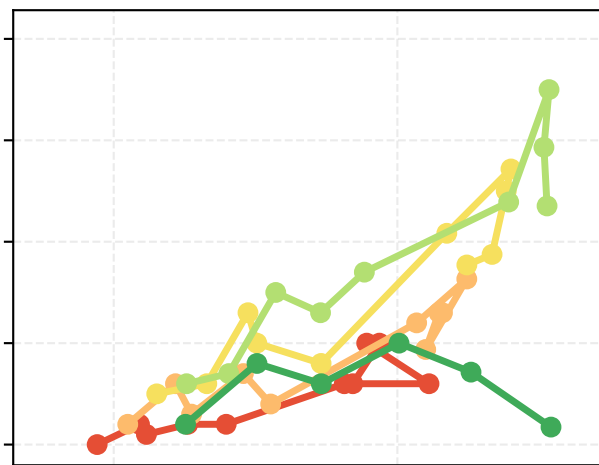
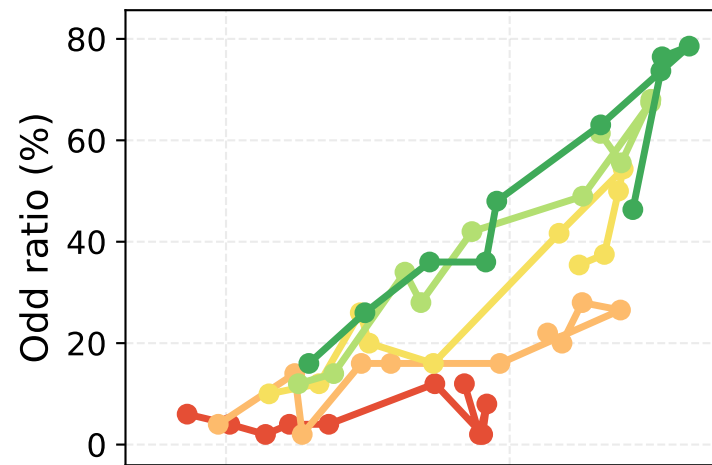
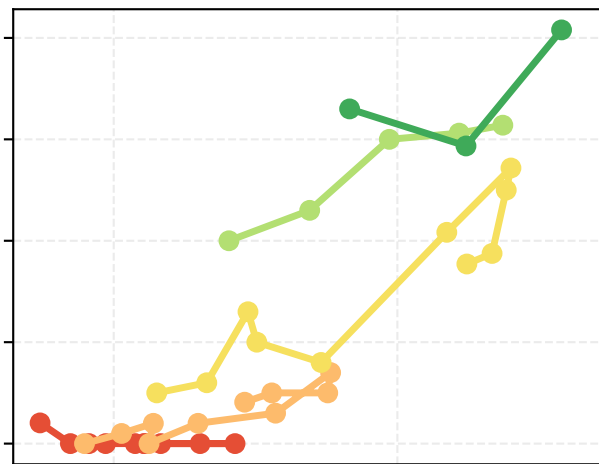
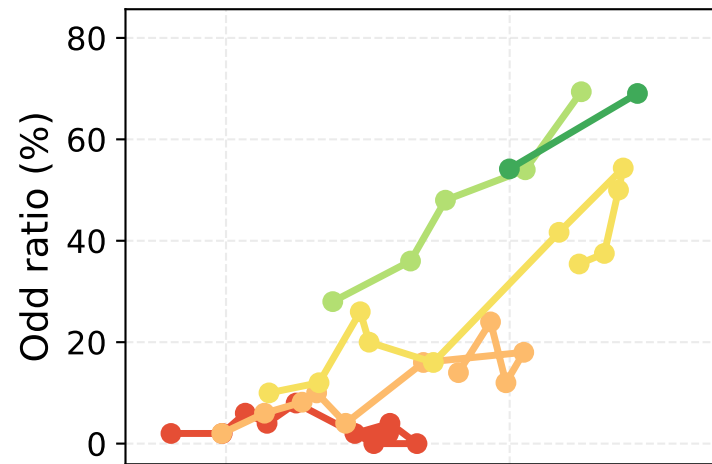
Steer strength -0.2 -0.1 +0.0 +0.1 +0.2

#1

#2

#3

#4



Average CoT token length

Average CoT token length