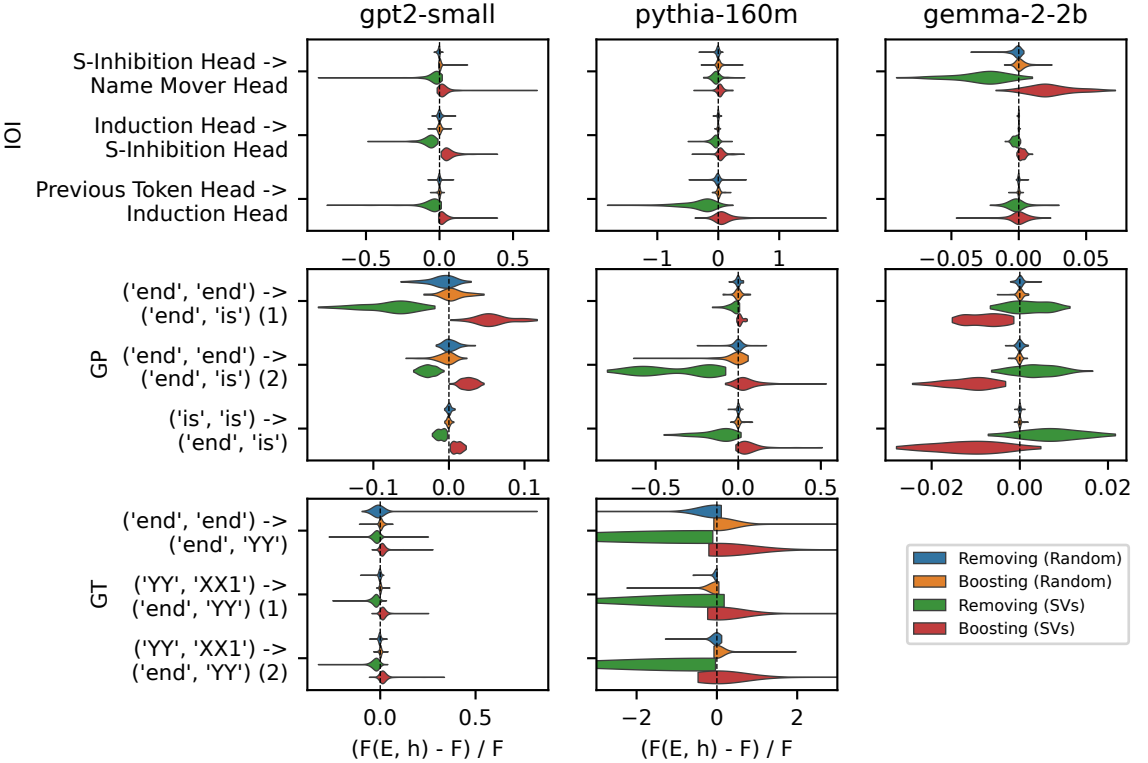
gpt2-small
pythia-160m
gemma-2-2b
IOI
S-Inhibition Head -> Name Mover Head
Induction Head -> S-Inhibition Head
Previous Token Head -> Induction Head
GP
('end', 'end') -> ('end', 'is') (1)
('end', 'end') -> ('end', 'is') (2)
('is', 'is') -> ('end', 'is')
GT
('end', 'end') -> ('end', 'YY')
('YY', 'XX1') -> ('end', 'YY') (1)
('YY', 'XX1') -> ('end', 'YY') (2)
Removing (Random)
Boosting (Random)
Removing (SVs)
Boosting (SVs)
(F(E, h) - F) / F
(F(E, h) - F) / F