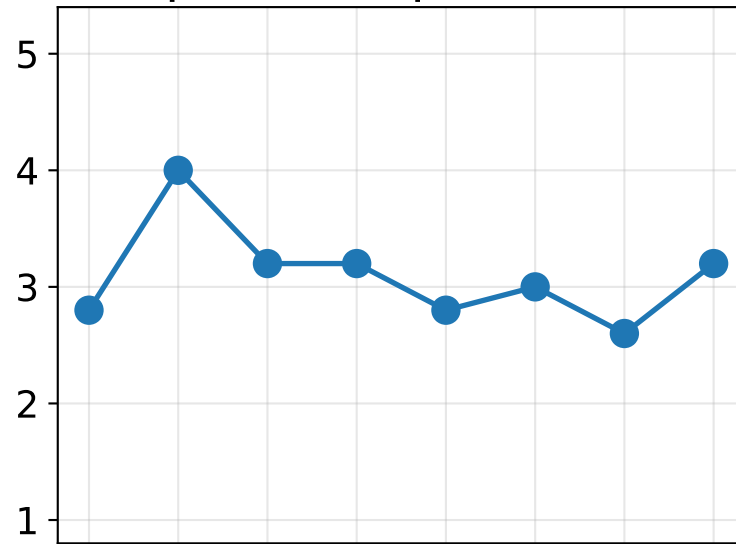
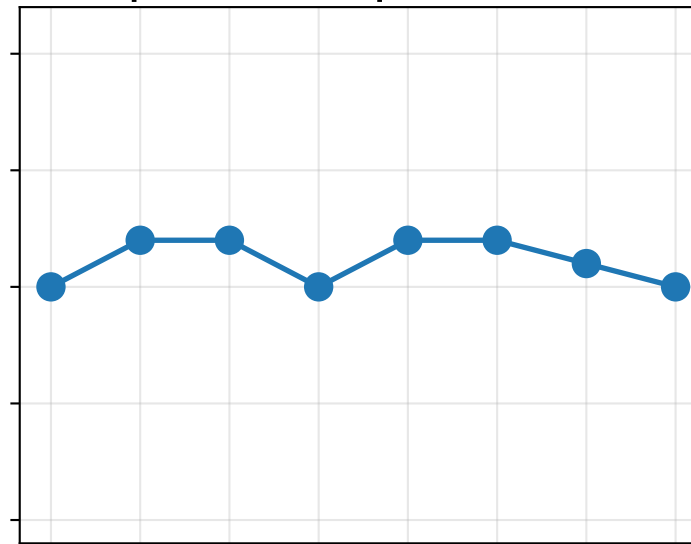


Score Trends by Release Date (Llama)

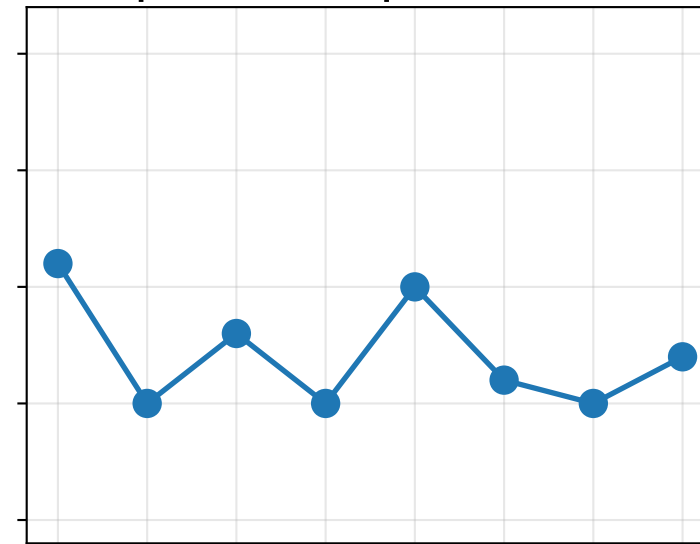
Eph→Per
($\rho=-0.25$, $p=0.558$)



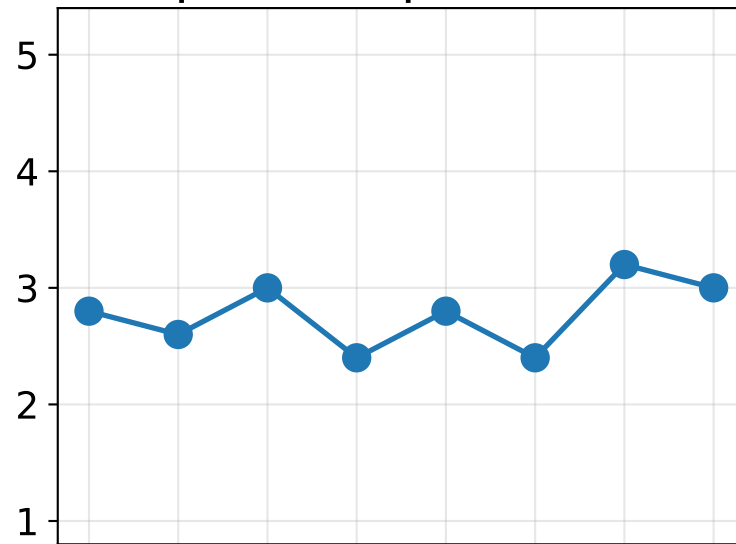
Ctx→Wgt
($\rho=-0.10$, $p=0.806$)



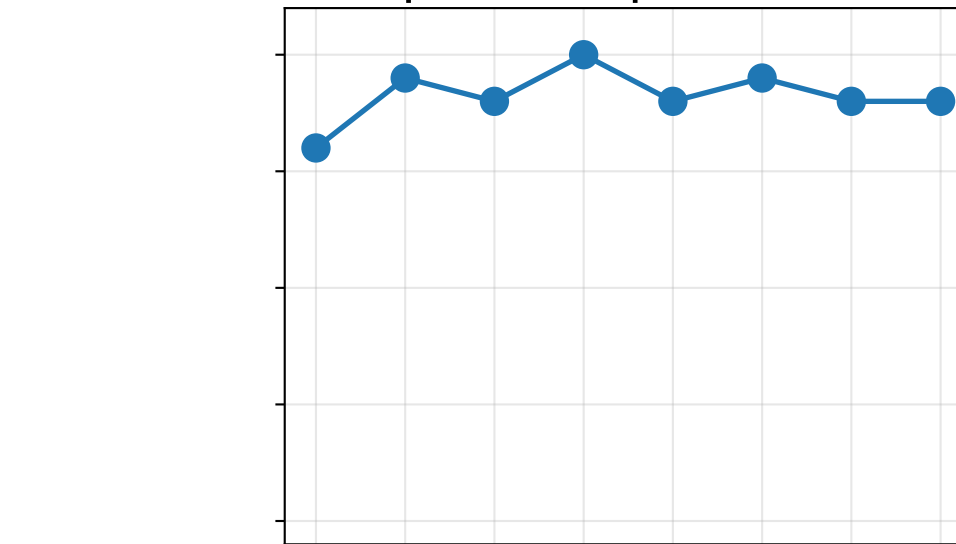
Sng→Dst
($\rho=-0.27$, $p=0.520$)



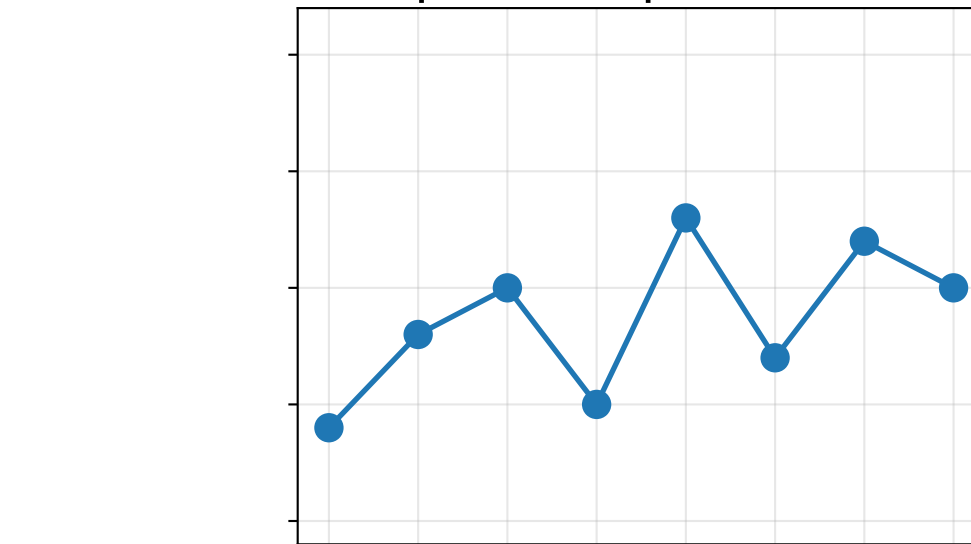
Pas→Agn
($\rho=0.33$, $p=0.429$)



Cert→Unc
($\rho=0.09$, $p=0.833$)



Hum→Aln
($\rho=0.54$, $p=0.168$)



Llama-3-70B-Instruct
Llama-3-8B-Instruct
Llama-3.1-405B-Instruct
Llama-3.1-70B-Instruct
Llama-3.1-8B-Instruct
Llama-3.3-70B-Instruct
Llama-4-Maverick
Llama-4-Scout

Llama-3-70B-Instruct
Llama-3-8B-Instruct
Llama-3.1-405B-Instruct
Llama-3.1-70B-Instruct
Llama-3.1-8B-Instruct
Llama-3.3-70B-Instruct
Llama-4-Maverick
Llama-4-Scout

Llama-3-70B-Instruct
Llama-3-8B-Instruct
Llama-3.1-405B-Instruct
Llama-3.1-70B-Instruct
Llama-3.1-8B-Instruct
Llama-3.3-70B-Instruct
Llama-4-Maverick
Llama-4-Scout