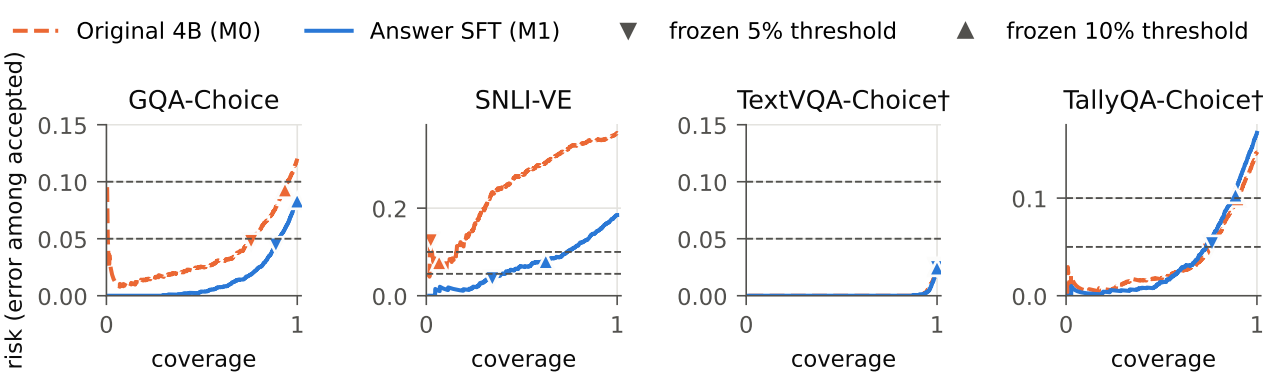

Original 4B (M0)
Answer SFT (M1)
frozen 5% threshold
frozen 10% threshold
risk (error among accepted)
GQA-Choice
SNLI-VE
TextVQA-Choice†
TallyQA-Choice†
0.15
0.10
0.05
0.00
0.2
0.0
0
1
coverage