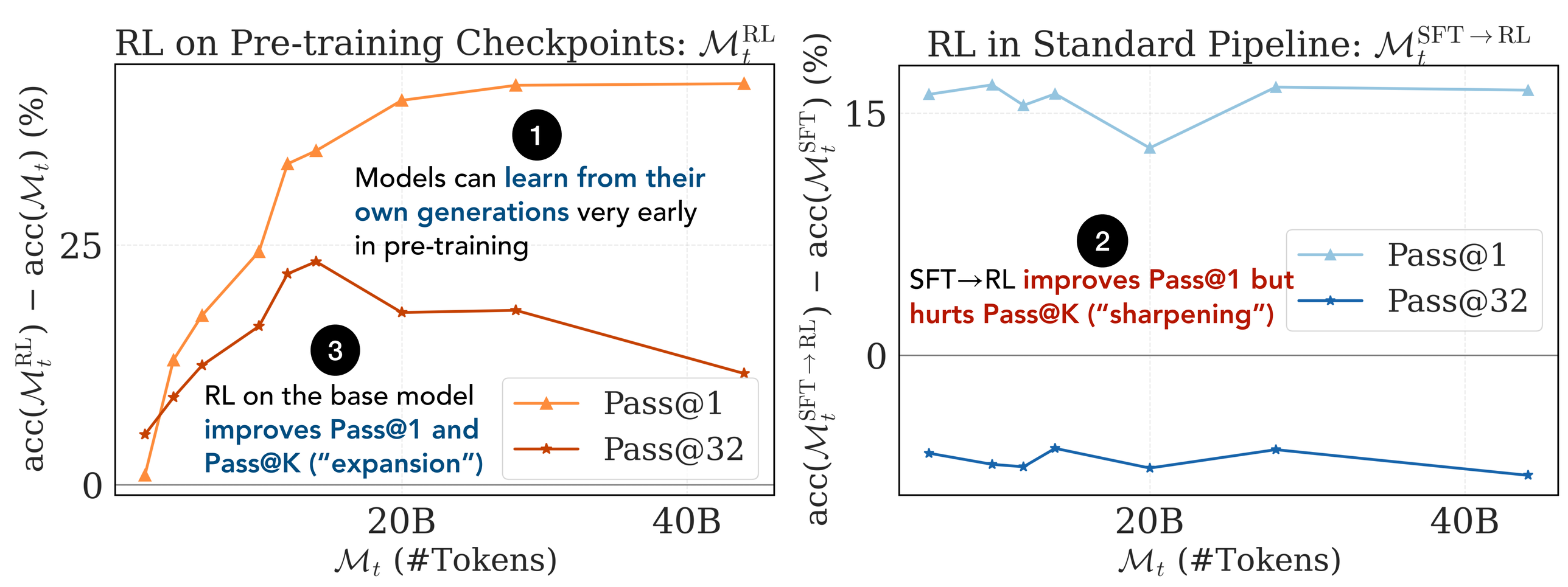
RL on Pre-training Checkpoints: $\mathcal{M}_t^{\mathrm{RL}}$
$\mathrm{acc}(\mathcal{M}_t^{\mathrm{RL}}) - \mathrm{acc}(\mathcal{M}_t)$ (%)
25
0
1
Models can learn from their own generations very early in pre-training
3
RL on the base model improves Pass@1 and Pass@K ("expansion")
Pass@1
Pass@32
20B
40B
$\mathcal{M}_t$ (#Tokens)
RL in Standard Pipeline: $\mathcal{M}_t^{\mathrm{SFT}\to\mathrm{RL}}$
$\mathrm{acc}(\mathcal{M}_t^{\mathrm{SFT}\to\mathrm{RL}}) - \mathrm{acc}(\mathcal{M}_t^{\mathrm{SFT}})$ (%)
15
0
2
SFT→RL improves Pass@1 but hurts Pass@K ("sharpening")
Pass@1
Pass@32
20B
40B
$\mathcal{M}_t$ (#Tokens)