
(a) Offline (Stages 1–3)
· every component is frozen before evaluation; no base-model parameter is updated
1 · Target-native supervision
170 exact-prompt safe / unsafe CoT–answer pairs $(x_i,\ y_i^s,\ y_i^u)$
teacher-force on the frozen f_{θ_m}; read each trajectory's $h_{t_2}^{\ell_m}$
2 · Learn and calibrate the edit
mean difference → signed continuation-NLL loss ⇒ $\delta_m^{\ell_m}$
calibrate α_m on the teacher-forced pool; no endpoint labels used
3 · Learn when to intervene
450 direct-harmful vs. 450 no-op prompts; read their unedited $h_{t_2}^{\ell_m}$
gate $g(h) = \mathbf{1}[(h-c)^\top w_g \geq \tau]$ linear; dev. false-trigger rate $\leq 5\%$
(b) Inference (Stage 4)
· one request: $\widehat{h}_{t_2}^{\ell_m} = h_{t_2}^{\ell_m} + g(h_{t_2}^{\ell_m})\ \alpha_m \delta_m^{\ell_m}$
columns: tokens · rows: layers
prompt x
<think>
t_0
t_1
t_2
t_3 ⋯
</think>
answer y
layer
L
ℓ_m
1
pre-generation guards read only the prompt
post-generation guards read only the finished answer
$h_{t_2}^{\ell_m}$: read by the gate; edited if $g=1$
native state
never edited
frozen gate g reads the unedited $h_{t_2}^{\ell_m}$
$g=1$: add $\alpha_m \delta_m^{\ell_m}$ once
$g=0$: no-op (P0 path)
continue decoding; no further edit