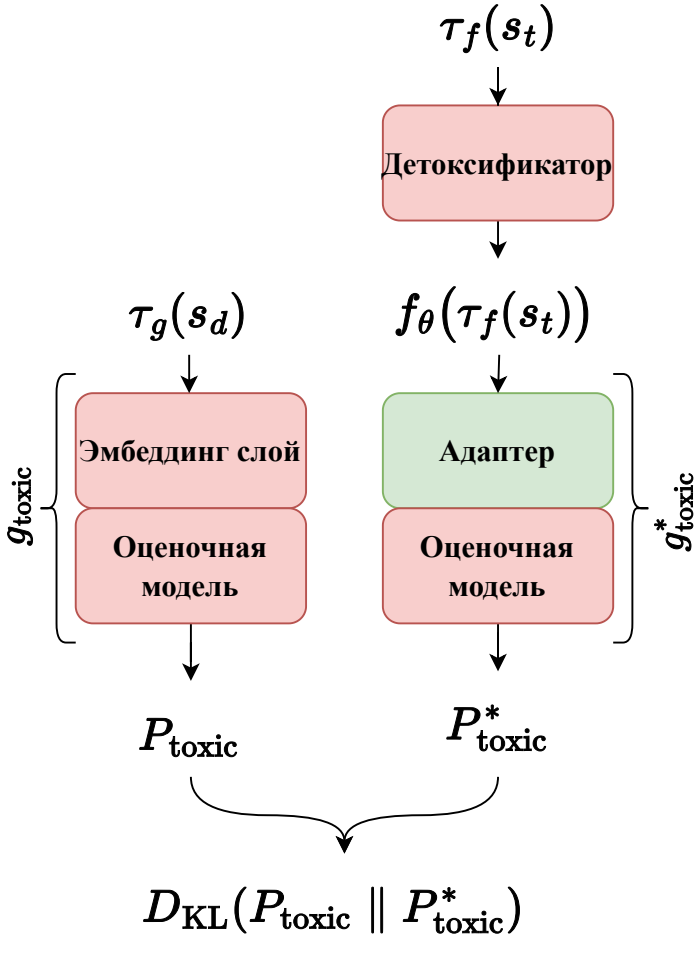
$\tau_f(s_t)$
Детоксификатор
$\tau_g(s_d)$
$f_\theta(\tau_f(s_t))$
$g_{\text{toxic}}$
Эмбеддинг слой
Оценочная модель
Адаптер
Оценочная модель
$g^*_{\text{toxic}}$
$P_{\text{toxic}}$
$P^*_{\text{toxic}}$
$D_{\text{KL}}(P_{\text{toxic}} \parallel P^*_{\text{toxic}})$