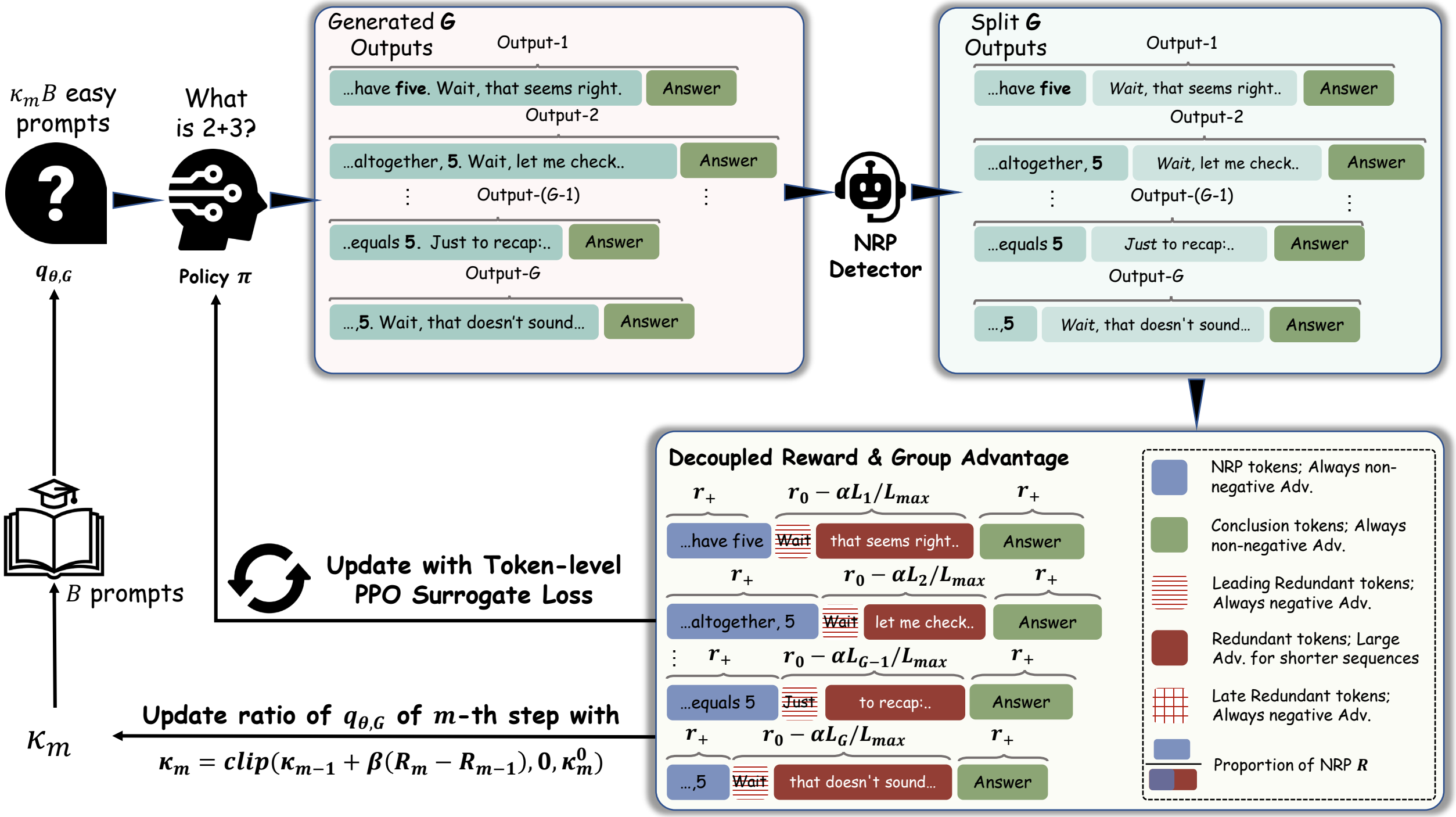
$\kappa_m B$ easy prompts
What is 2+3?
$q_{\theta,G}$
Policy π
Generated G Outputs
Output-1
...have five. Wait, that seems right.
Answer
Output-2
...altogether, 5. Wait, let me check..
Answer
Output-(G-1)
..equals 5. Just to recap:..
Answer
Output-G
...,5. Wait, that doesn't sound...
Answer
NRP Detector
Split G Outputs
Output-1
...have five
Wait, that seems right..
Answer
Output-2
...altogether, 5
Wait, let me check..
Answer
Output-(G-1)
...equals 5
Just to recap:..
Answer
Output-G
...,5
Wait, that doesn't sound...
Answer
Decoupled Reward & Group Advantage
r_+
$r_0 - \alpha L_1/L_{max}$
r_+
...have five
Wait
that seems right..
Answer
r_+
$r_0 - \alpha L_2/L_{max}$
r_+
...altogether, 5
Wait
let me check..
Answer
r_+
$r_0 - \alpha L_{G-1}/L_{max}$
r_+
...equals 5
Just
to recap:..
Answer
r_+
$r_0 - \alpha L_G/L_{max}$
r_+
...,5
Wait
that doesn't sound...
Answer
NRP tokens; Always non-negative Adv.
Conclusion tokens; Always non-negative Adv.
Leading Redundant tokens; Always negative Adv.
Redundant tokens; Large Adv. for shorter sequences
Late Redundant tokens; Always negative Adv.
Proportion of NRP R
B prompts
Update with Token-level PPO Surrogate Loss
Update ratio of $q_{\theta,G}$ of m-th step with
$\kappa_m = clip(\kappa_{m-1} + \beta(R_m - R_{m-1}), 0, \kappa_m^0)$
κ_m