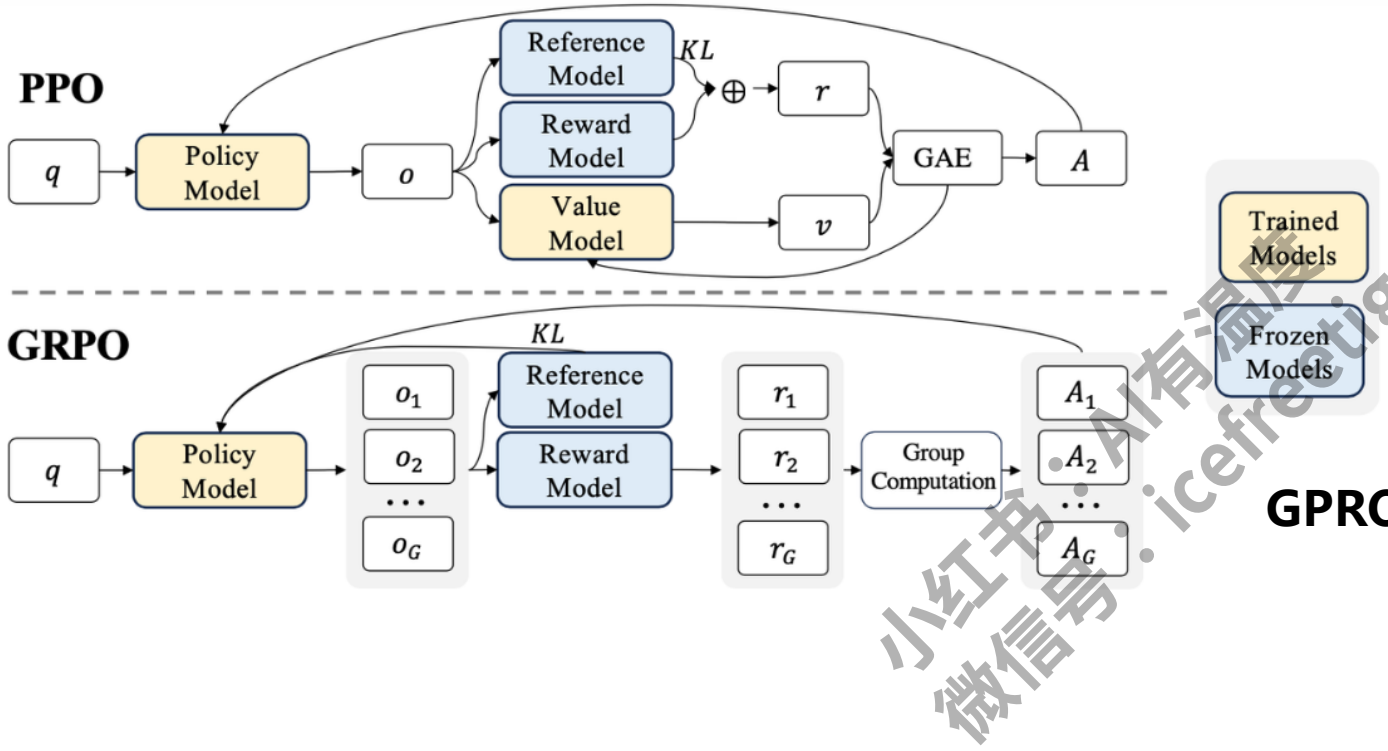
Deepseek R1

强化学习训练过程 复现

前置知识：GPRO

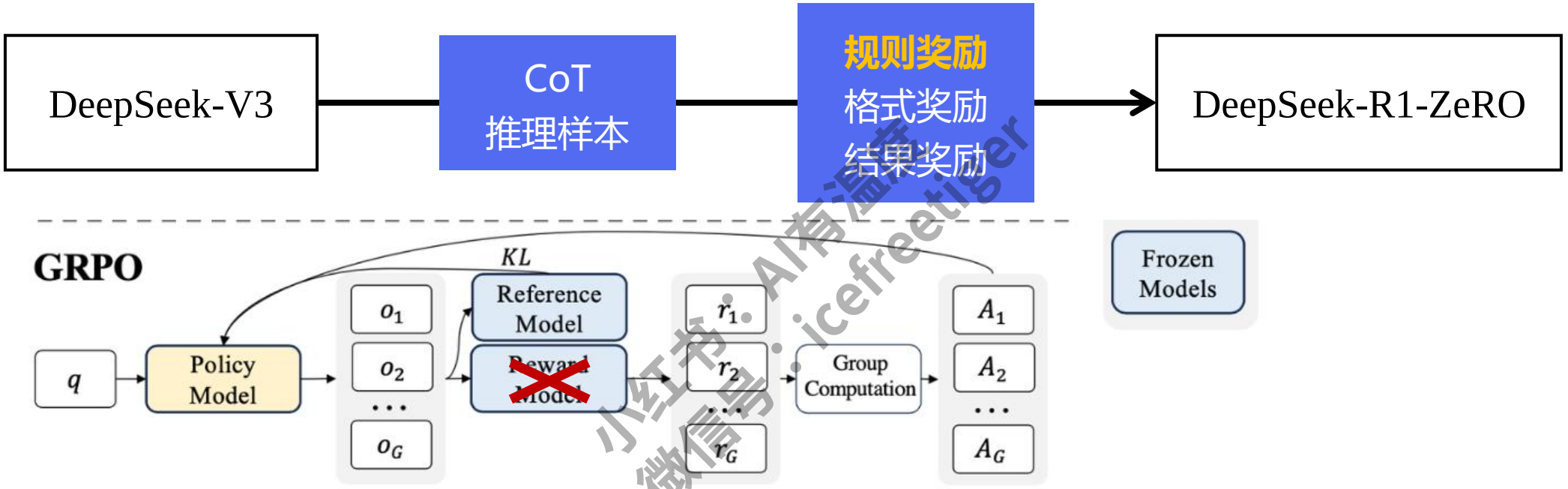


PPO优势 = 实际分数r-预期分数v

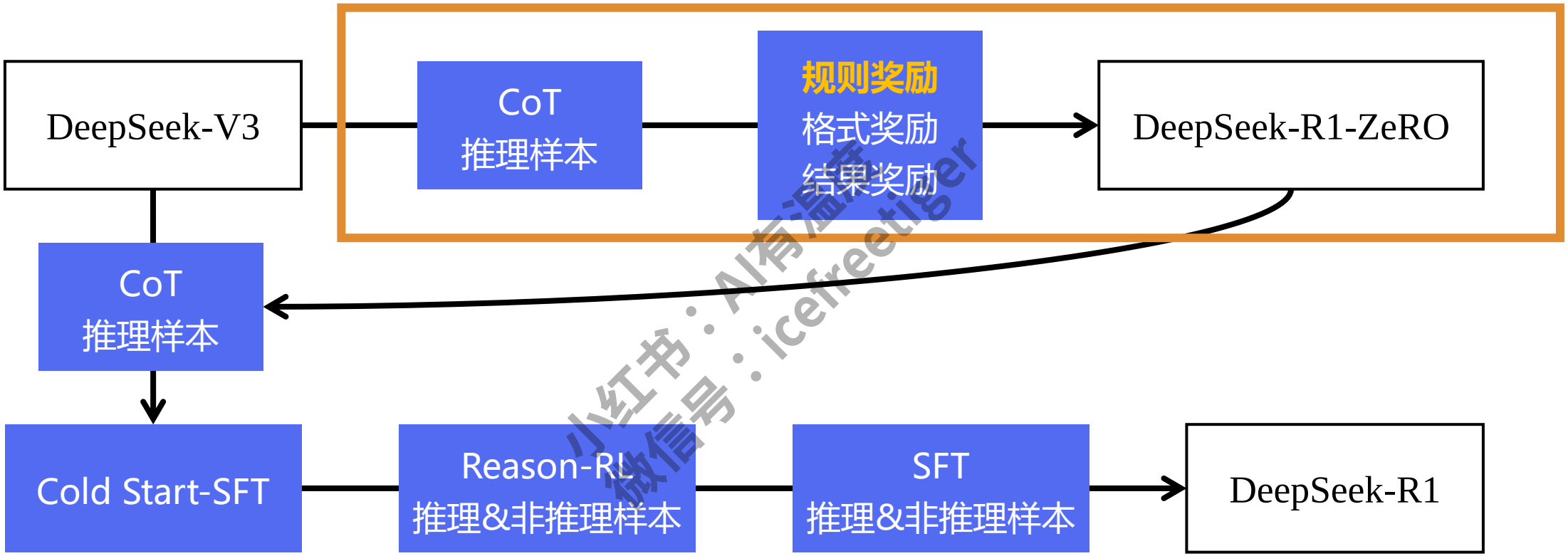
GPRO优势 = $A_i^G = \frac{r_i - \mu_G}{\sigma_G + \epsilon}$

GPRO更关注群体表现，能节省大量的 GPU 内存

前置知识：R1-ZeRO训练过程



训练启示



关注我

ICEFREE AI

因你而升温！！