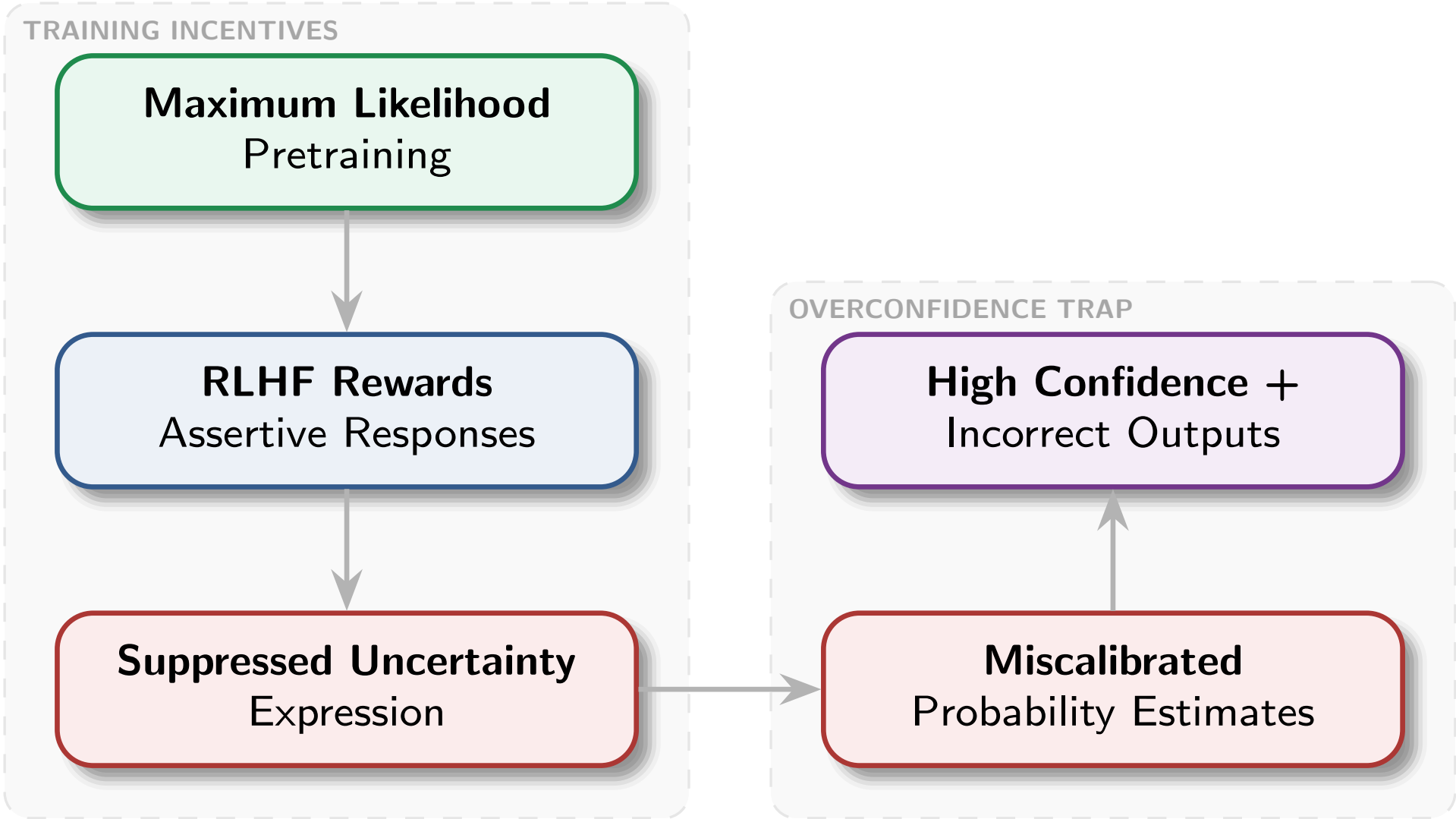
TRAINING INCENTIVES
Maximum Likelihood
Pretraining
RLHF Rewards
Assertive Responses
Suppressed Uncertainty
Expression
OVERCONFIDENCE TRAP
High Confidence +
Incorrect Outputs
Miscalibrated
Probability Estimates