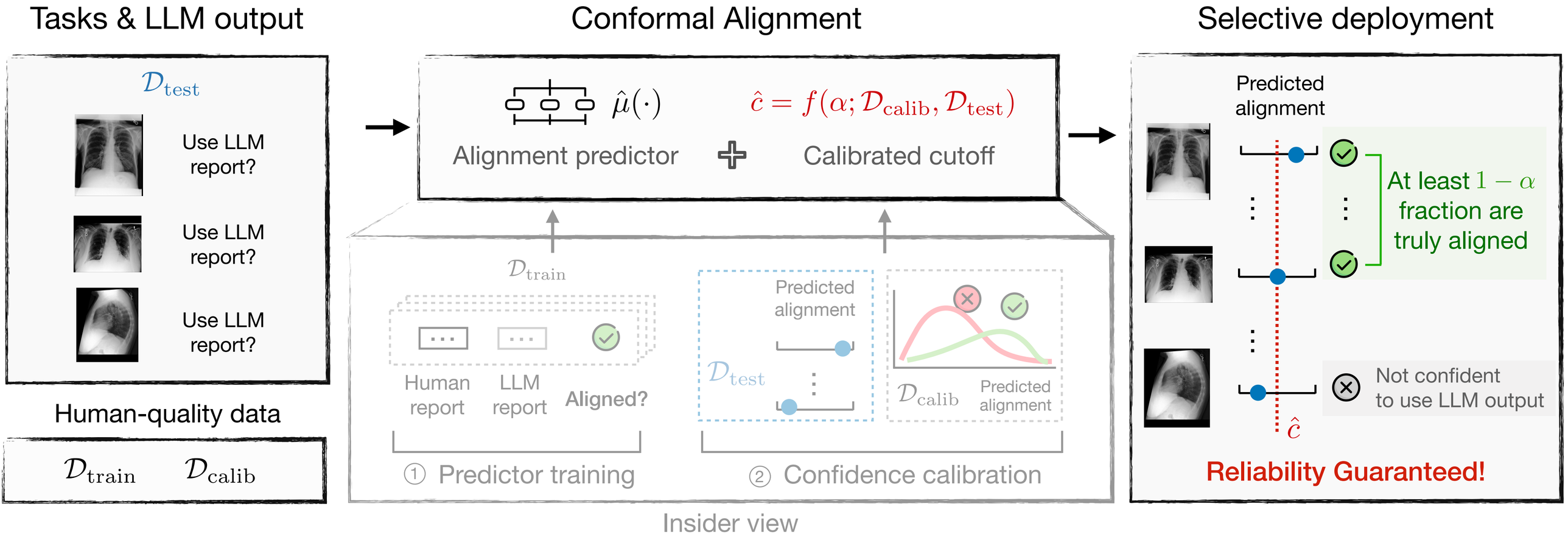
Tasks & LLM output
Conformal Alignment
Selective deployment
$\mathcal{D}_{\text{test}}$
Use LLM report?
Use LLM report?
Use LLM report?
Human-quality data
$\mathcal{D}_{\text{train}}$ $\mathcal{D}_{\text{calib}}$
$\hat{\mu}(\cdot)$
Alignment predictor
$\hat{c} = f(\alpha; \mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$
Calibrated cutoff
$\mathcal{D}_{\text{train}}$
Human report
LLM report
Aligned?
① Predictor training
Predicted alignment
$\mathcal{D}_{\text{test}}$
$\mathcal{D}_{\text{calib}}$
Predicted alignment
② Confidence calibration
Insider view
Predicted alignment
At least $1-\alpha$ fraction are truly aligned
Not confident to use LLM output
$\hat{c}$
Reliability Guaranteed!