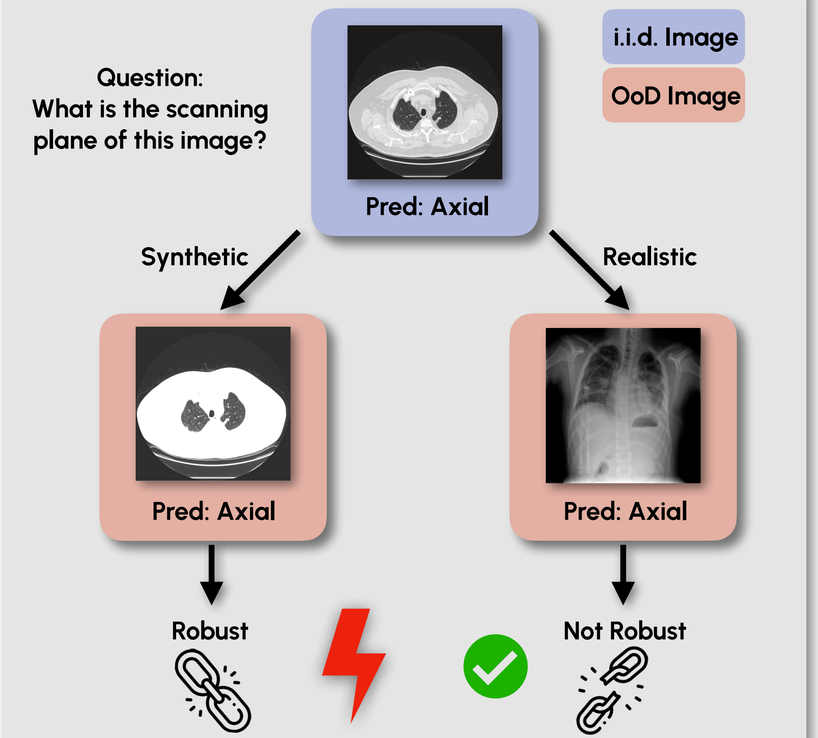
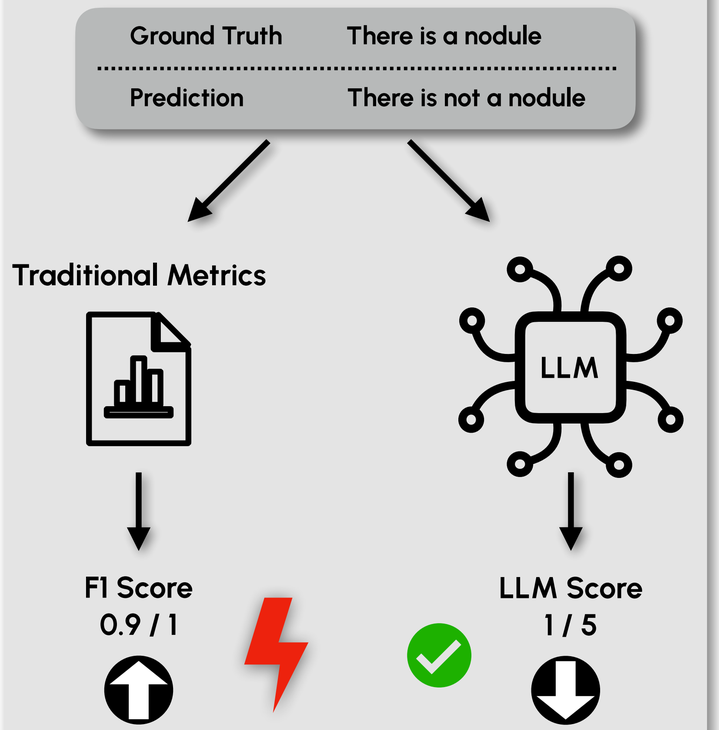
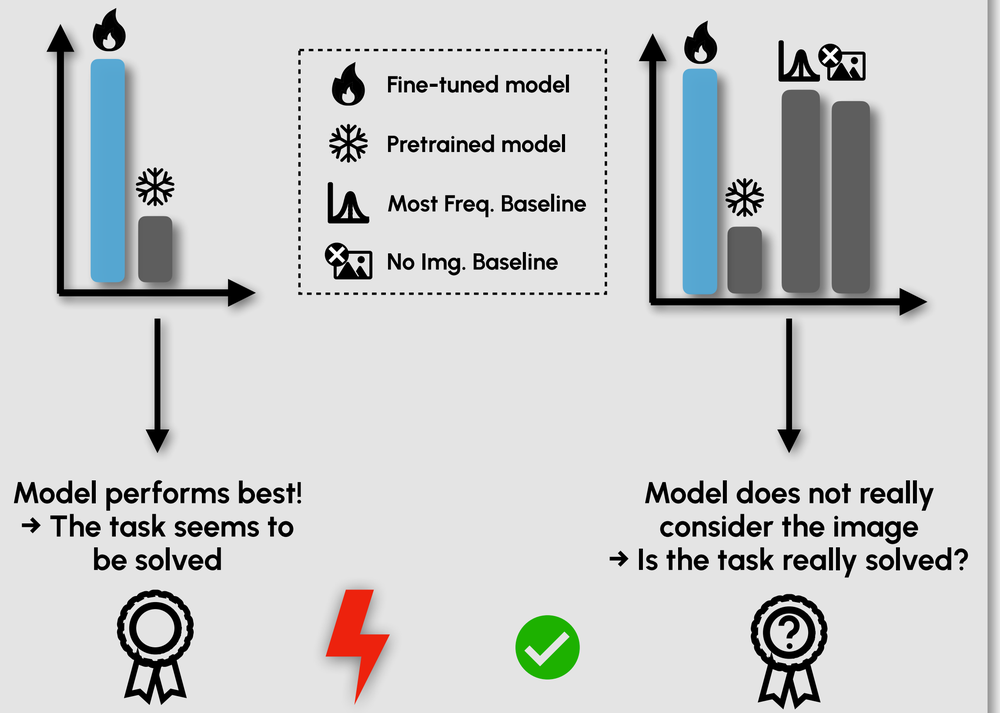
P1: Robustness on synthetic shifts does not imply robustness on real-world shifts



P2: Traditional metrics do not capture the underlying semantics



P3: Model performance lacks interpretability due to missing sanity baselines



R1: Realistic Shifts

R2: Appropriate Metrics

R3: Relevant Sanity Baselines