

(a) Before: Single Evaluation

Single Test Set

No Confidence Intervals

No Statistical Testing

Prone to Bias

(b) After: Statistical Robust Evaluation

5-Fold Cross-Validation

100 Meta-Tasks

Confidence Intervals

Statistical Significance

Stratified Sampling