Requirements Engineering   Implementation

Decide measures for evaluating correctness in performing T, e.g.,

global correctness or recall & precision

based on expected distribution of possible answers

Establish human thresholds for chosen measures and the correct answers by having expert humans perform T on a representative input I

Build A for T

Validate A by running A(I) and using its answers to calculate A's measures.

If A's measures are better than the humans' then A is validated

If validated then

Run A in production while watching for A's changing the real world and T, leading to a change in requirements

If not validated then

Time line of Validation of an AI, A, for task, T