Step 1
Exploration of Existing Benchmarks

Agent Company

Toolathlon

M3Eval

WebArena

Experimental & Trajectory Analysis

Task Design Principles

Step 2
Seed Task Construction

From Benchmark Analysis (Stage 1)

Select High-Quality Tasks → Unify Tool Interfaces → Manual Refinement

From Public APIs

Collect Public APIs (e.g. GitHub) → Wrap as Executable Tools → Manually Design Tasks

From Local Tools and Data

Prepare Data and Files → Implement Local Tool Interfaces → Manually Design Tasks

Seed Task Pool

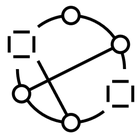
Step 3
Task Complexity Scaling

Seed Task Pool

Fetch one repo and extract the branch number

Quantitative Scaling

Fetch 5 repos and extract the branch number
....



Complexity Scaling

Fetch 100 repos, extract branch counts with contributor numbers, then analyze their correlations.

Final Task Pool