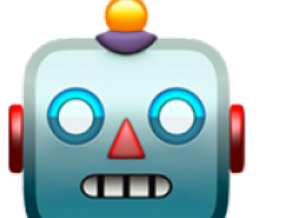


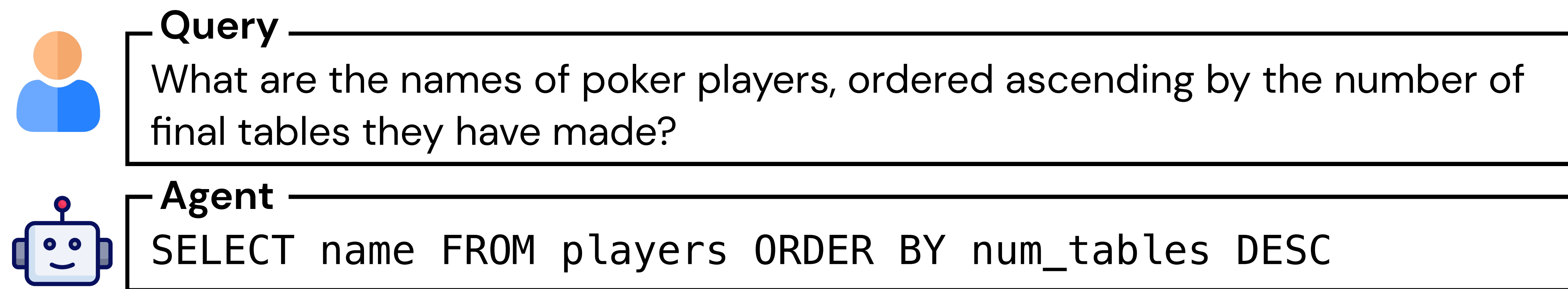


Interactive Coding with Execution Feedback

Seq2Seq Coding

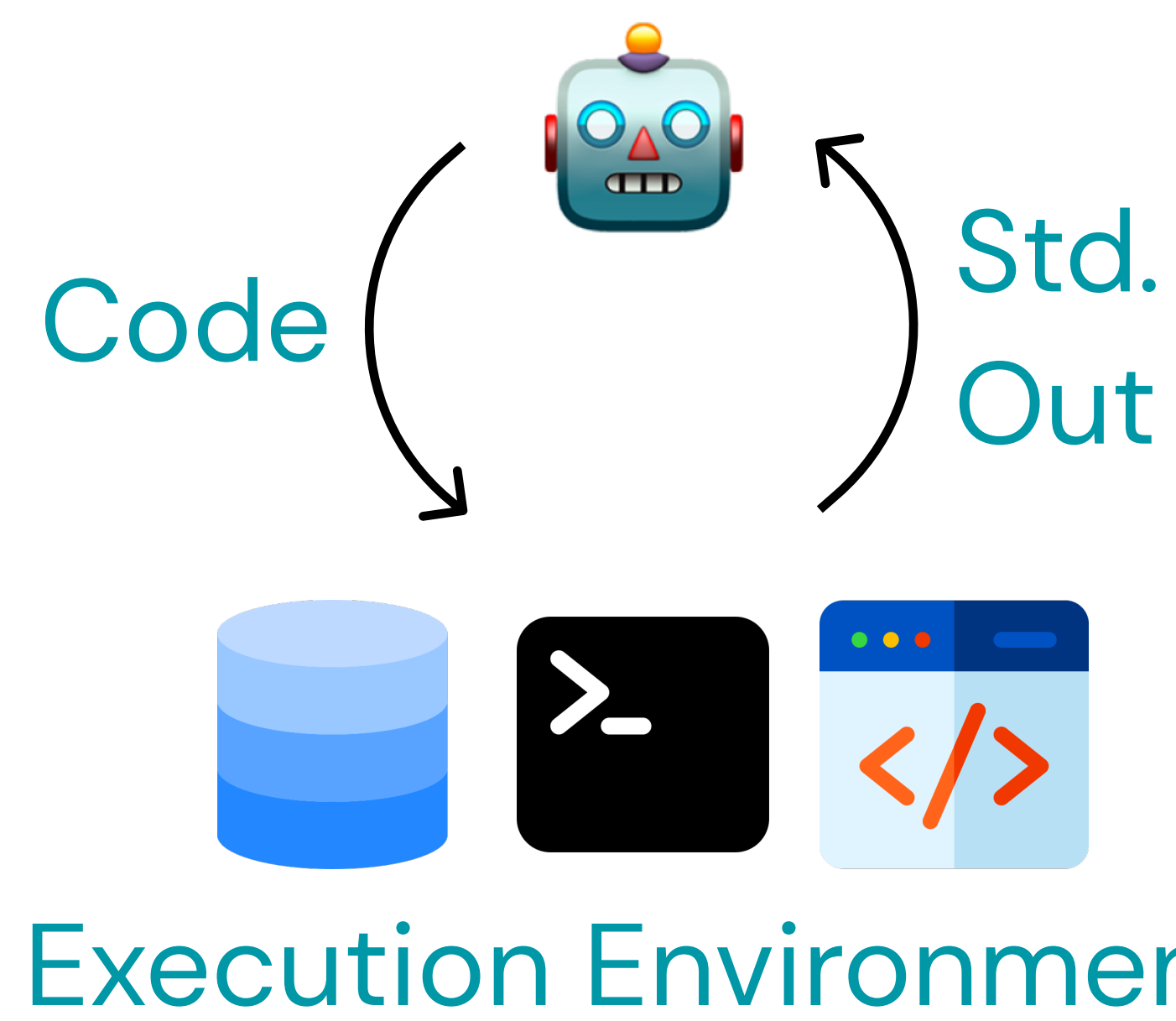
NL Query →  → Code

No execution or grounding



Our Work

Interactive Coding!





Evaluate Language Models as Agents that write *code* as *actions* to interact with a software system

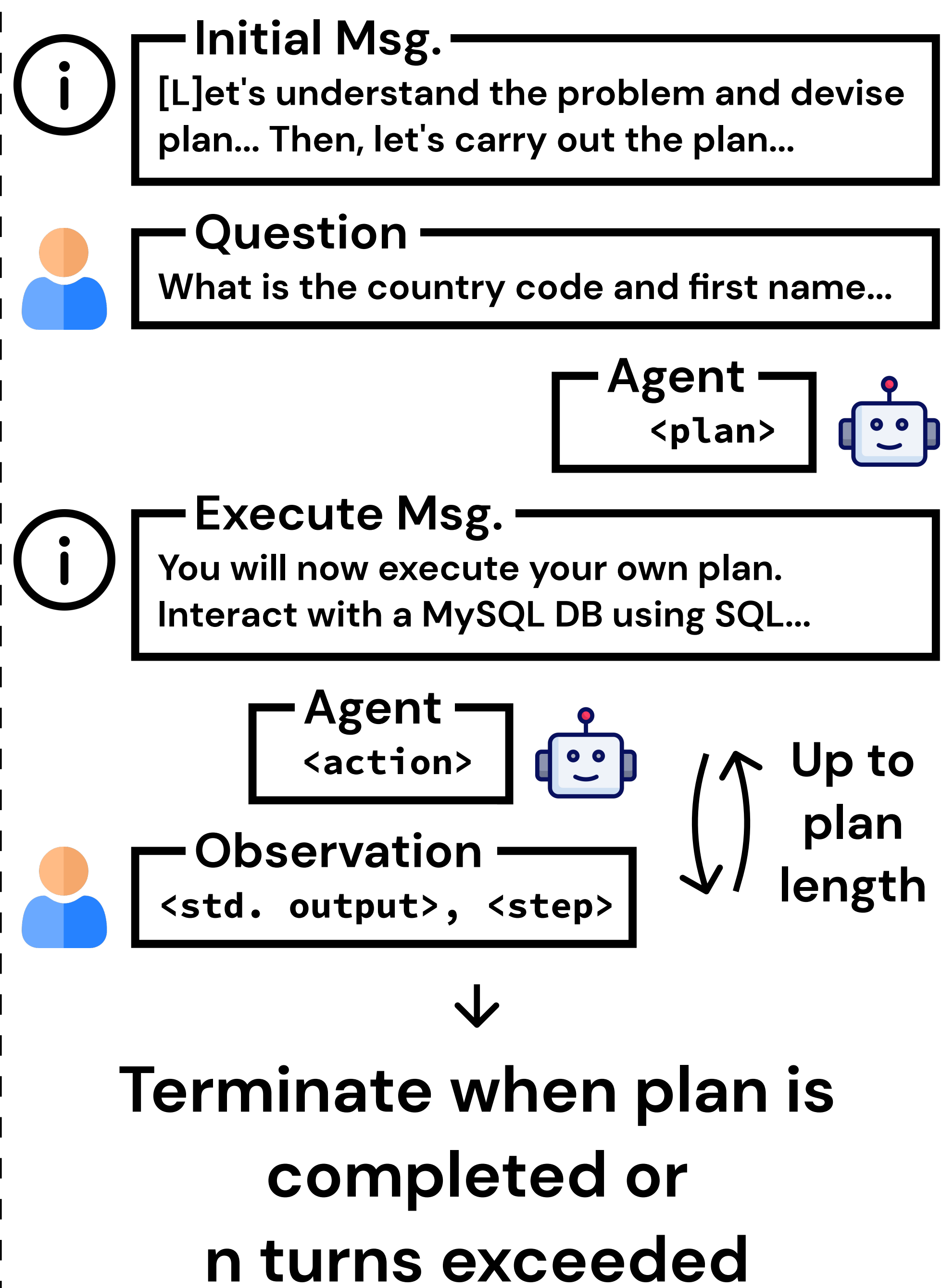
Motivation

- **Interaction & feedback necessary** for hard tasks
- **Standardized** environment enables **consistent benchmarking** of coding agents

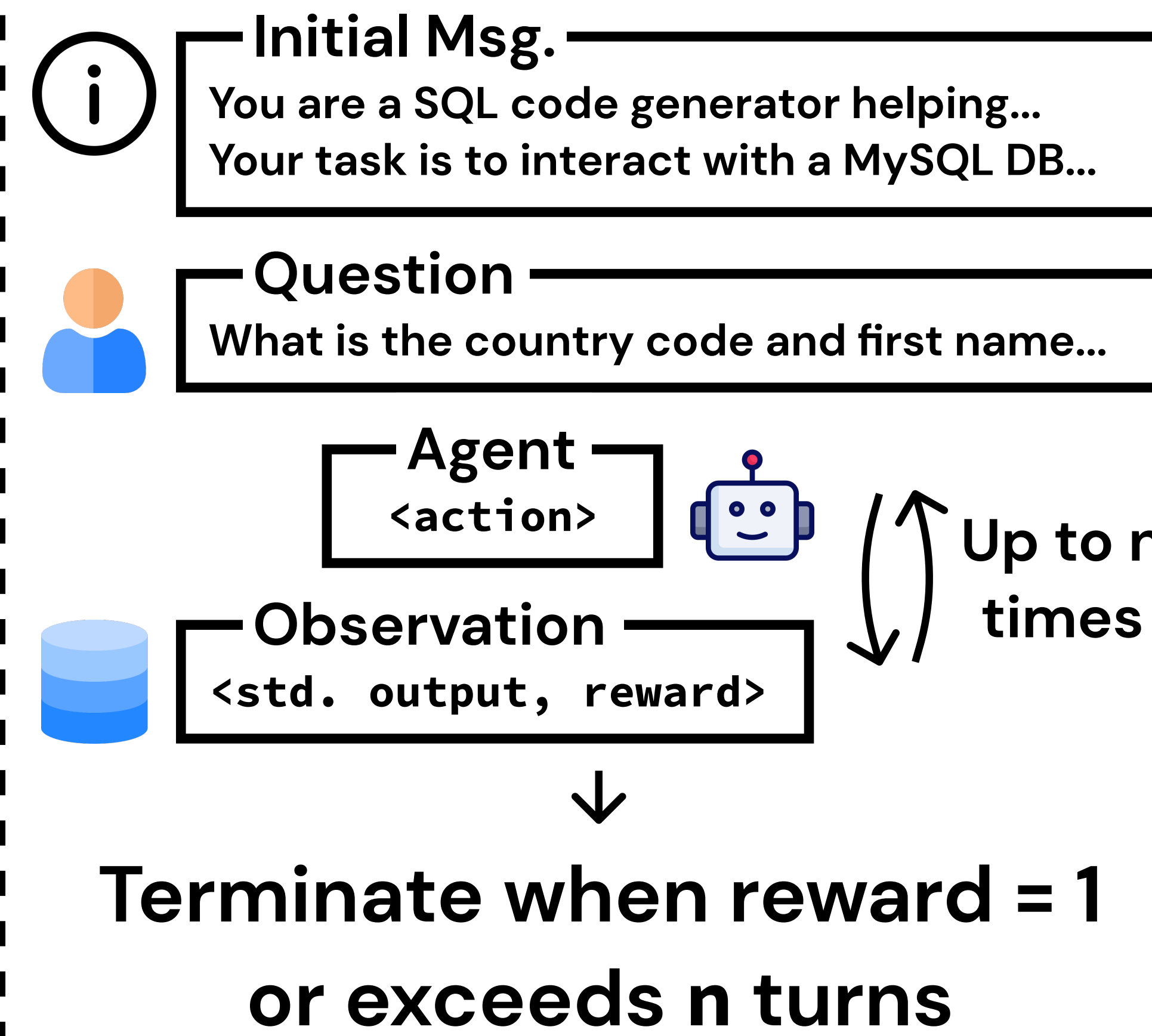
Features

- **Abstracted**: Handles interaction, execution logic under-the-hood
- **Lightweight**: Define task env. in <100 lines of code
- **Safe**: Virtual containers = no accidental  of 

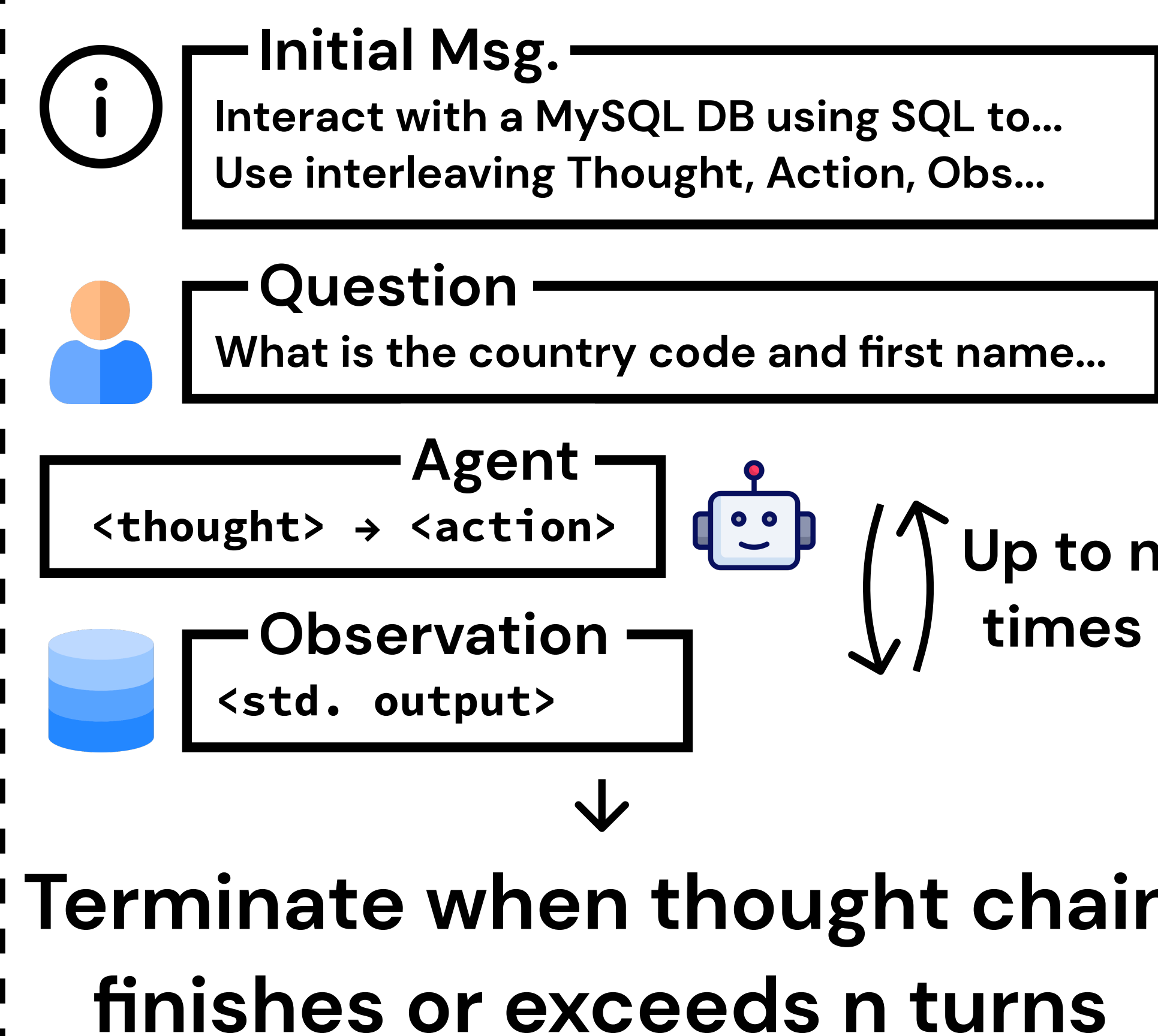
Plan & Solve







"Try Again"



ReAct



4 New Interactive Coding Tasks


Action	Setting	Datasets
 Bash	Ubuntu Shell	NL2Bash (200)
 SQL	MySQL DB	Spider, BIRD (10K)
 Python	Interpreter	MBPP, APPS (10K)
 CTF	Ubuntu Shell	IC-CTF (100)

Make Your Own Task is Easy



Provide the following to InterCodeEnv:

 **Dockerfile**, to define your task setting.

 **Dataset** of task instances to evaluate.

 **Reward Func**, to score agent trajectories and determine whether a task is completed.

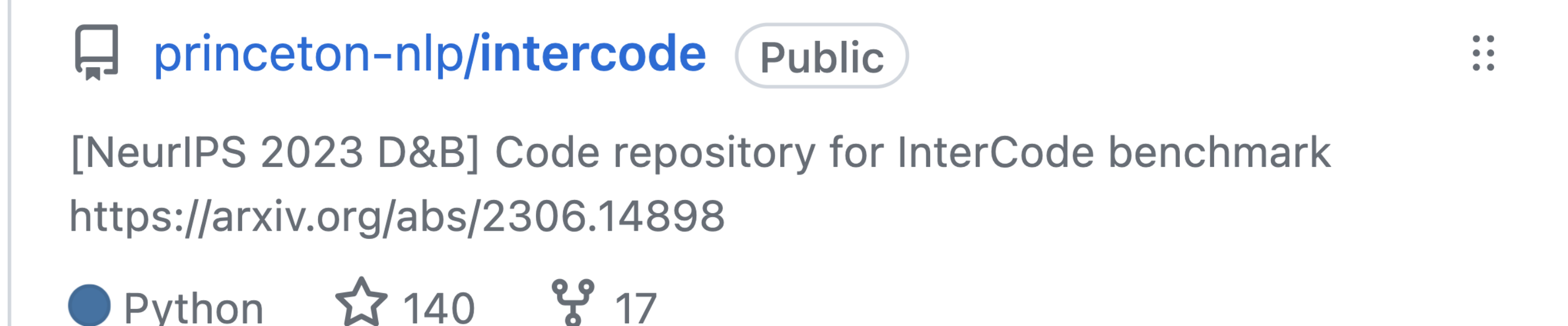
Key Results

GPT 3.5	Single	Try Again	ReAct	Plan/Solve
 Bash	34.5	46.5	20.5	28.0
 SQL	10.5	47.3	58.7	49.1

- Multi-turn >> single-turn
- Diff. methods better on diff. tasks
- Large room for improvement

Check it Out!

 intercode-benchmark.github.io



princeton-nlp/intercode Public

[NeurIPS 2023 D&B] Code repository for InterCode benchmark
https://arxiv.org/abs/2306.14898

Python 140 17