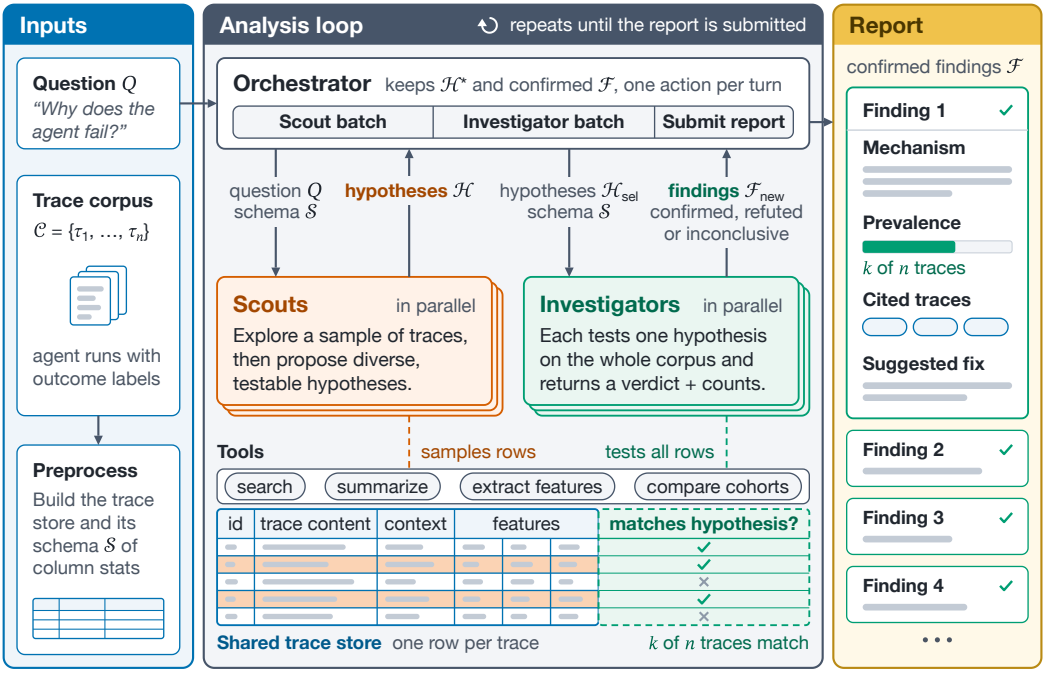

Inputs
Question Q
"Why does the agent fail?"
Trace corpus
$\mathcal{C} = \{\tau_1, \ldots, \tau_n\}$
agent runs with outcome labels
Preprocess
Build the trace store and its schema $\mathcal{S}$ of column stats
Analysis loop
repeats until the report is submitted
Orchestrator
keeps $\mathcal{H}^*$ and confirmed $\mathcal{F}$, one action per turn
Scout batch
Investigator batch
Submit report
question Q
schema $\mathcal{S}$
hypotheses $\mathcal{H}$
hypotheses $\mathcal{H}_{\text{sel}}$
schema $\mathcal{S}$
findings $\mathcal{F}_{\text{new}}$
confirmed, refuted or inconclusive
Scouts
in parallel
Explore a sample of traces, then propose diverse, testable hypotheses.
Investigators
in parallel
Each tests one hypothesis on the whole corpus and returns a verdict + counts.
Tools
samples rows
tests all rows
search
summarize
extract features
compare cohorts
id
trace content
context
features
matches hypothesis?
Shared trace store
one row per trace
k of n traces match
Report
confirmed findings $\mathcal{F}$
Finding 1
Mechanism
Prevalence
k of n traces
Cited traces
Suggested fix
Finding 2
Finding 3
Finding 4