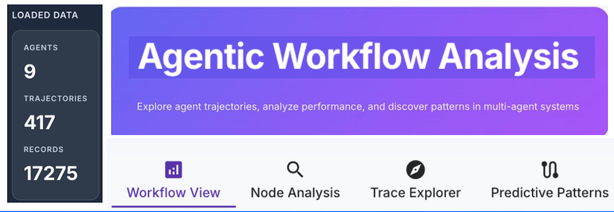
Entry View

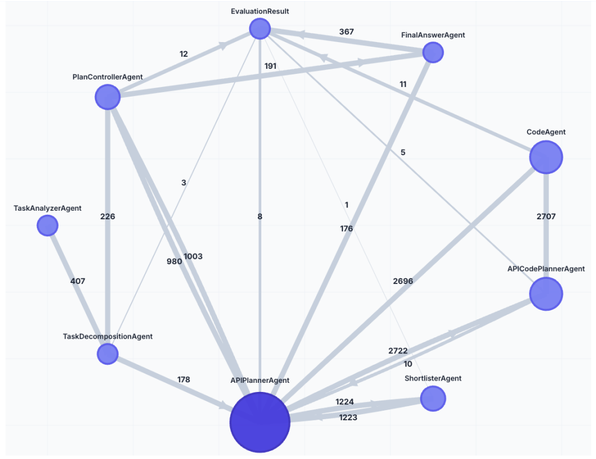
Provides an easy entry point to our tool.



a Entry view, with tabs for each view

1. System View

Provides a workflow-level overview to help understand the system behavior and identify high-level patterns and issues in the agent execution.



b Agent workflow graph, system-level statistics, global scores, and system-level recurring issues

Average Overall Score

0.83

Recurring Issues Across Workflow

Most common issues identified in full trace evaluations

ISSUE	COUNT	SEVERITY	FREQUENCY
Redundant or duplicate operations (API calls, planning cycles, cart management) that could be consolidated	123	0.23	34.1%
Incomplete workflows that terminate before completing all sub-tasks	90	0.40	24.9%
Inefficient API usage, such as fetching all data then filtering locally, not using bulk/batch operations, or making multiple similar calls	83	0.20	23.0%

2. Node View

Provides node-level issues with per-instance error analysis.

Node-Specific CLEAR Analysis

Select an agent to view its evaluation results

Select Agent: TaskDecompositionAgent

TaskDecompositionAgent Statistics

CLEAR evaluation results for this specific agent

Evaluations	Avg Score	Unique Issues	Tasks
407	0.89	9	407

Discovered Issues

Recurring problems identified in this agent's outputs

ISSUE	COUNT	SEVERITY	FREQUENCY
Subtask text does not reproduce the original intent verbatim	78	0.29	19.2%
it paraphrases instead of copying the exact phrasing (single-application case).	78	0.29	19.2%
JSON output is not a raw object – it is wrapped in quotes, includes extra prefixes, or contains invalid JSON syntax.	44	0.17	10.8%
Task decomposition is incomplete or missing essential steps.	13	0.39	3.2%

Filter Data

Filter evaluation records by issues and score

Select issues using AND/OR/NOT logic, then click Apply Filter

Include ANY of (OR): At least one of the...

Must ALSO have (AND): All of these — Subtask text does not reproduce the original intent verbatim

0 — 0.6

c Filter by issue types and score range

d Per-instance evaluation, scores, and recurring issues

Evaluation Summary

The response correctly recognized that only Gmail is involved and gave a single high-level subtask in the proper JSON format, with clear and relevant reasoning. However, the primary rule requires the original intent to be returned verbatim, and the model paraphrased the intent instead. This deviation means the answer does not fully meet the specification, lowering its correctness, though all other aspects are satisfactory.

Score: 0.45

Recurring Issues

Subtask text does not reproduce the original intent verbatim

it paraphrases instead of copying the exact phrasing (single-application case).

3. Trace View

Provides search and filtering options, trace-level analytics, high-level evaluation, and rubric and dimension-based evaluation.

Search & Filter Traces

Search by Task ID or Intent

Clear All Filters

Advanced Filters

Trace Length: 80 — 127 (Range: 1 - 127 steps)

Contains Agents: Select agents — EvaluationResult (9 agents available)

Score Range (Standard Evaluation): 0 — 0.6 (0.0 - 1.0 (overall_score from Standard Evaluation))

Apply Filters

User Intent: Buy me the top-rated gaming console controller that's available now on amazon for each of my siblings, and have it delivered to their homes with gift wrapping. I want to give them identical gifts.

Steps	Agents	Min Score	Avg Score
125	8	N/A	N/A

e Trace search and filtering. Detailed assessment of trace quality

Overall Score: 0.54

Detailed Feedback

The trajectory shows strong initial understanding of the task, correctly identifying required applications and decomposing the problem. Reasoning steps are well-documented, and most API interactions are appropriate. However, the workflow fails to handle inventory constraints, leading to repeated cart-addition errors and a mis-filtered product selection. State management is inconsistent, causing contradictory progress reports and an incomplete final deliverable. Future iterations should incorporate early inventory checks, maintain a single-order-per-recipient flow, and ensure that only products matching the "gaming console controller" type are considered.

Rubric Score: 0.60

3 of 5 rubrics fulfilled

f Rubric evaluation, with cross-rubric evaluation summary

Evaluation Summary

The agent successfully identified the siblings and their addresses (R1) and performed a proper Amazon search to find the top-rated in-stock controller (R2). It verified availability and added the controller to the cart (R3). However, it failed to maintain separate carts per sibling (R4) and did not provide identical gifts for both siblings, resulting in mismatched order confirmations (R5).

g Per-rubric assessment, with description, criterion, fulfillment assessment, and reasoning

R3 — ✓ Fulfilled

Description: Verify that the selected controller is available for immediate purchase and can be ordered multiple times.

Criterion: *The agent confirmed the product's availability status is "Available now" and that the same product ID can be added to the cart without quantity restrictions.*

Evaluation Reasoning

R4 — ✕ Not Fulfilled

Description: Add the controller to a separate cart for each sibling, enable gift wrapping, set the correct shipping address, and place the order.

Criterion: *The agent added the selected product to a cart, enabled the gift-wrap option, assigned each sibling's shipping address, and executed a purchase action resulting in distinct order confirmations for each sibling.*

Evaluation Reasoning

Step Quality Dimensions

- Correctness: 0.45
- Completeness: 0.50
- Clarity: 0.60

Trace-Level Dimensions

- Objective_Understanding: 0.70
- Information_Completeness: 0.60
- Execution_Quality: 0.45

h Step and trace dimension scores