

Build 3: Agent Testing Log

Jacob Poore and Arian Zarazua

April 5, 2026

Overview

We tested the Build 3 HITL Router Agent using two datasets: a game-stats CSV (5,980 rows, NFL seasons 2013–2023) and a Metadata-2024 CSV (48 rows, 2024 season game info). The metadata file had a lot of missing values and only covered September 2024, which ended up limiting some of the analyses. We ran 10 prompts total — 6 worked, 4 failed.

Prompt Log

#	Prompt	Route	Result	Notes
1	How have avg rushing yards allowed changed over seasons?	Tool	Pass	Worked fine
2	How do rushing yards correlate with wins/losses?	Codegen	Fail	No <code>wins</code> column in data
3	Same as #2 (retry)	Codegen	Fail	Same issue
4	How do rushing yards correlate with passer rating?	Tool	Fail	Wrong arg names sent to tool
5	Can you graph rushing yards and passer rating?	Codegen	Pass	Chart came out correctly
6	Suggest attendance correlation analyses	Suggest	Pass	Good suggestions returned
7	How does avg attendance change by event date?	Tool	Pass	Only Sep 2024 data available
8	Same as #7, across entire date range	Tool	Pass	Identical result to #7, no warning
9	Does wind speed correspond with roof type?	Tool	Fail	Pearson used on a categorical column
10	Plot avg temperature by roof type over dates	Codegen	Pass	Only “Outdoors” showed up (missing data)

Test Details

Test 1 — Rushing yards over seasons [Pass]

Prompt: How have average rushing yards allowed per game changed over the past seasons?
The agent picked `plot_temporal_line_chart` with the right arguments and the chart came out fine.

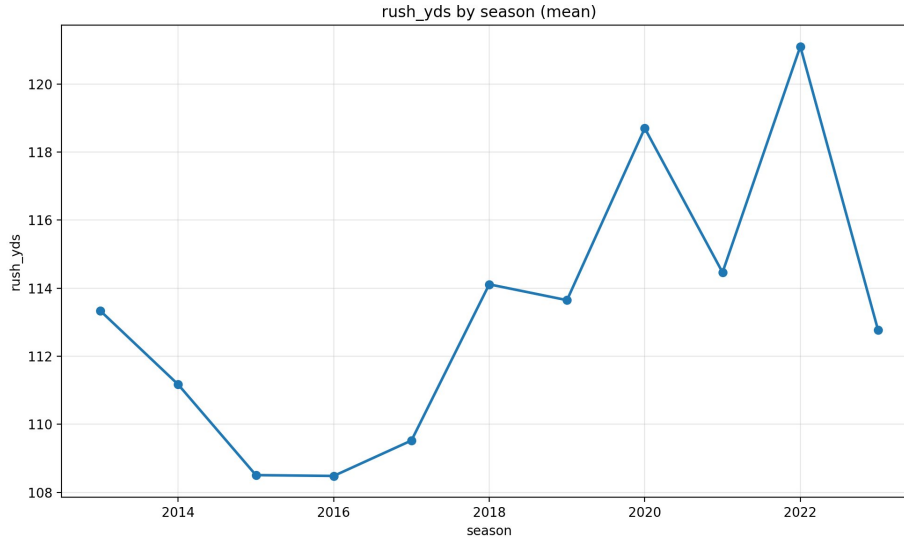


Figure 1: Mean rushing yards allowed per game by season.

Test 2 — Rushing yards vs. wins, attempt 1 [Fail]

Prompt: How do rushing yards allowed correlate with team success (wins/losses)?

Error: Script exited with code 1. It looked for a `wins` column that doesn't exist in the dataset — the data only has per-game stats, no win/loss record.

What went wrong: The agent went straight to writing code without checking what columns were actually available. It assumed `wins` existed.

Test 3 — Rushing yards vs. wins, retry [Fail]

Prompt: How do rushing yards allowed correlate with wins/losses? using codegen

Same failure as Test 2. The new script added a scatter plot but still tried to use the `wins` column. The agent didn't check the schema before trying again.

What went wrong: Same root cause. The agent should have told me the column didn't exist instead of just failing again.

Test 4 — Rushing yards vs. passer rating [Fail]

Prompt: How do rushing yards allowed correlate with passer rating?

Error:

```
[args] Dropping unsupported args: ['column_x', 'column_y']
ERROR: pearson_correlation() missing required arguments: 'x' and 'y'
```

What went wrong: The agent picked the right tool (`pearson_correlation`) but sent the arguments with the wrong names. The tool expects `x` and `y`, but the agent sent `column_x` and `column_y`. Easy fix but the agent didn't catch it.

Test 5 — Graph rushing yards and passer rating [Pass]

Prompt: Can you graph rushing yards allowed and passer rating?

After Test 4 failed I reworded it as a visualization request. The agent went to codegen and produced a dual-axis line chart correctly.

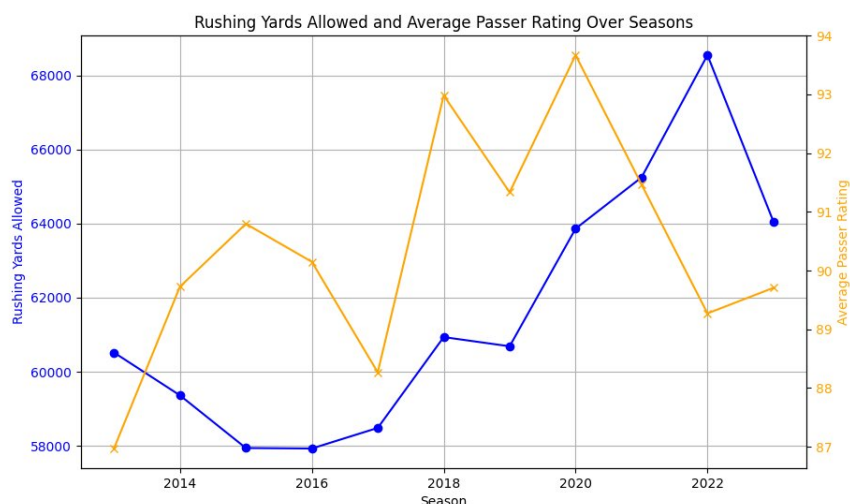


Figure 2: Rushing yards (blue) and average passer rating (orange) by season.

Test 6 — Suggest attendance analyses [Pass]

Prompt: Suggest correlation analyses we can run relating to fan attendance.

The agent returned three reasonable research questions with outcome variables, predictors, and suggested analysis types. It correctly treated this as a suggestion request rather than trying to run anything.

Test 7 — Attendance by event date [Pass]

Prompt: How does average fan attendance change according to event date?

Chart came out fine. Worth noting the data only covers September 2024, so the range is pretty narrow. The agent didn't mention this.

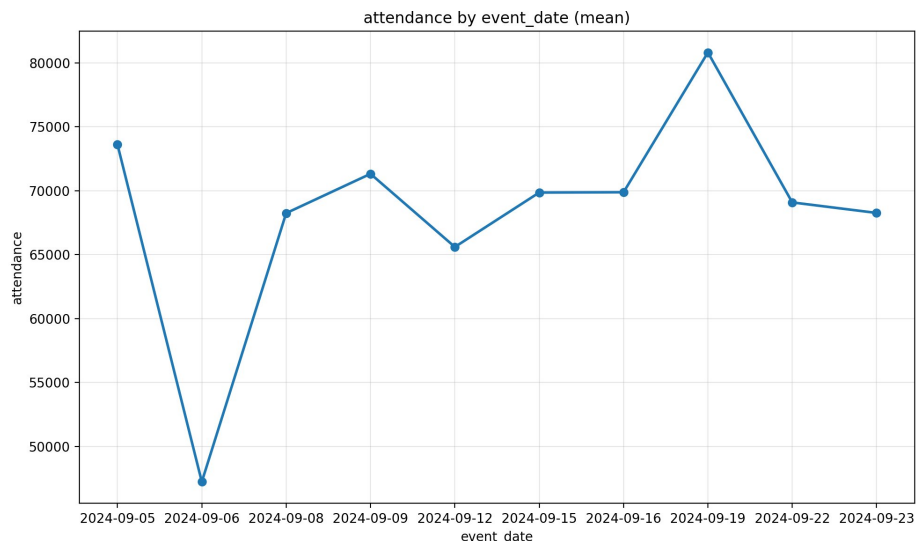


Figure 3: Mean attendance by event date (September 2024 only).

Test 8 — Attendance across entire date range [Pass]

Prompt: How does average fan attendance change across the entire date range?

Produced the same chart as Test 7. Technically correct given the data, but the agent should have flagged that “entire date range” is just three weeks in September.

Test 9 — Wind speed vs. roof type [Fail]

Prompt: Does wind speed correspond with roof type?

Error: Same wrong argument names as Test 4, and on top of that, `roof_type` is a text/categorical column — Pearson correlation doesn’t apply here. A box plot comparing wind speed across roof types would have been the right call. The tool `plot_cat_num_boxplot` was available but not used.

What went wrong: Two things: the argument name bug showed up again, and the agent didn’t notice it was trying to run a correlation on a non-numeric column.

Test 10 — Temperature by roof type over dates [Pass]

Prompt: Can you plot avg temperature by roof type over event dates? I imagine as temp gets colder near February there are more Dome events.

The code ran fine (exit code 0) but only the “Outdoors” line showed up — Dome and Retractable roof games were missing temperature data and got dropped. The February hypothesis couldn’t really be tested since the data only goes through September anyway. The agent didn’t point this out.

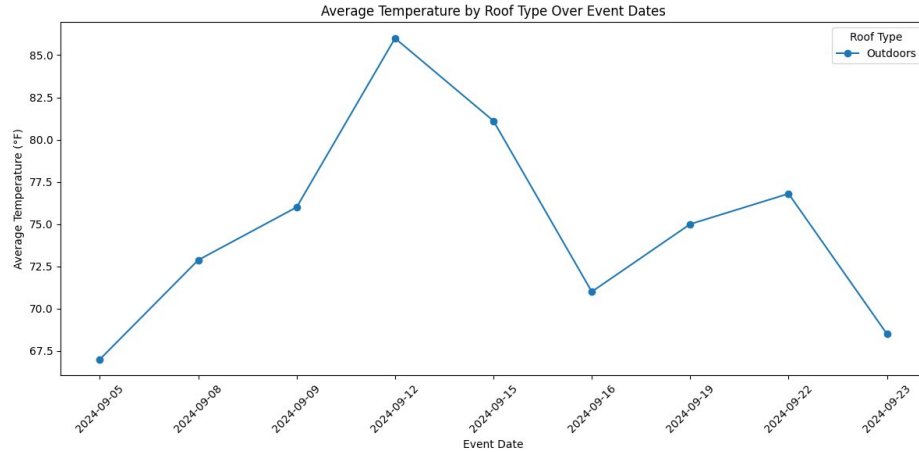


Figure 4: Avg temperature by roof type. Only outdoor games had temperature data.

Summary of What Went Wrong

Wrong argument names (Tests 4 and 9): The agent kept passing `column_x/column_y` to a tool that expects `x/y`. This happened twice, which suggests it’s a consistent bug in how the router constructs arguments, not a one-off mistake.

Missing column (Tests 2 and 3): The agent wrote code assuming a `wins` column existed without checking the schema first. It should have caught this before generating anything.

Correlation on a categorical column (Test 9): `roof_type` is text, not numeric. Pearson correlation doesn’t work here. The agent should check column types before picking a correlation tool.

Data range not flagged (Tests 7, 8, 10): The metadata only covers September 2024. The agent never mentioned this, even when the user asked for analysis across the “entire date range” or brought up a February hypothesis.