

Data Canvas

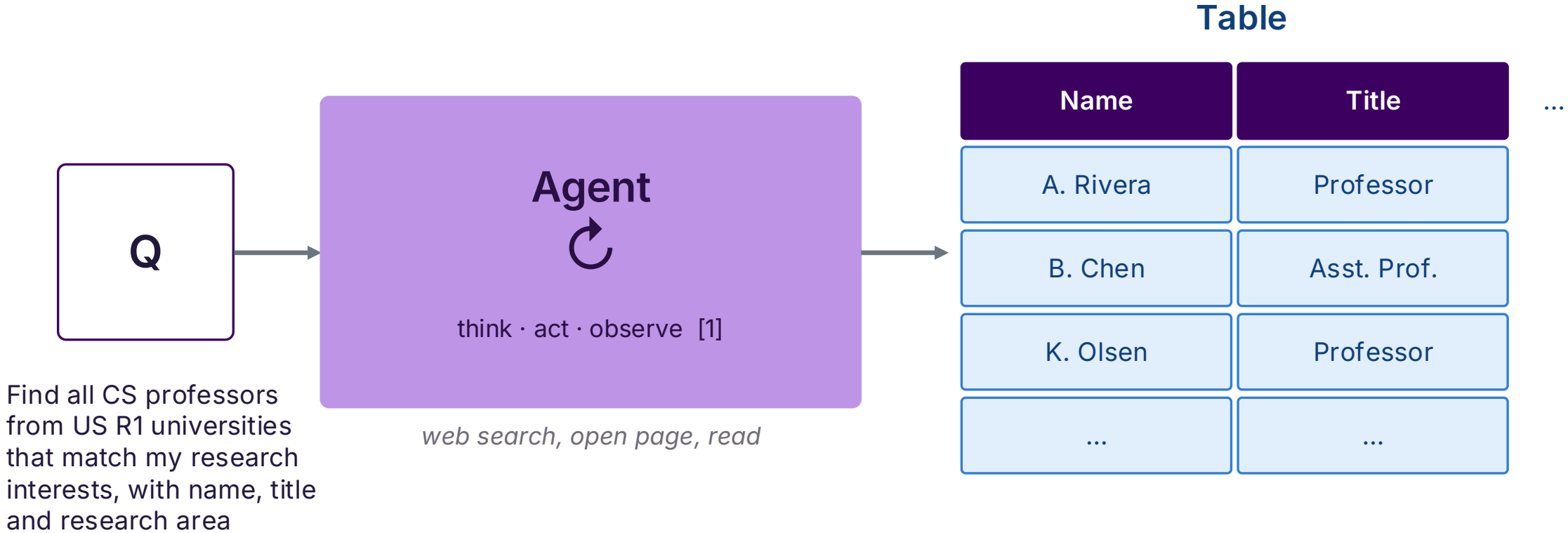
A Provenance-Guided Harness for Agentic Data Engineering

Zixuan Yi, Yuanming Shao, Shaun Wallace, Zachary G. Ives, Ryan Marcus

University of Pennsylvania, University of Rhode Island

Paper published at COLM 2026

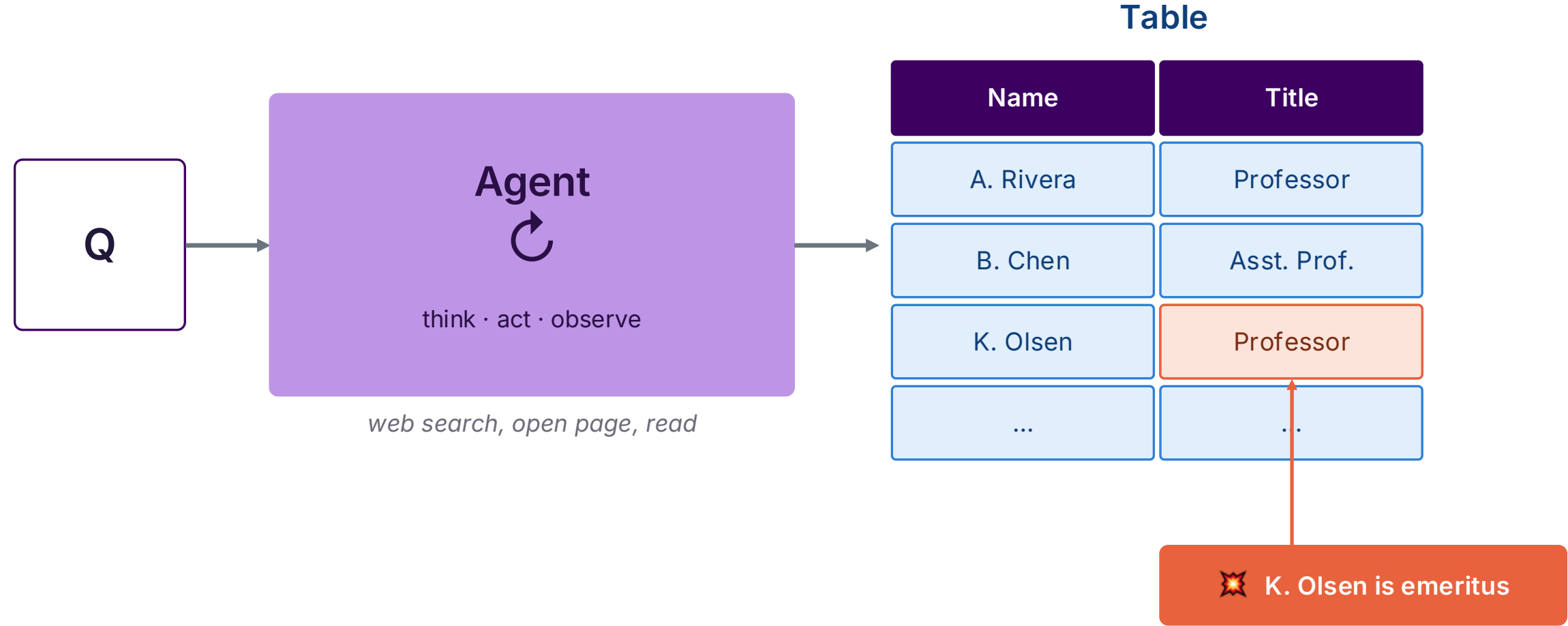
Agentic data engineering



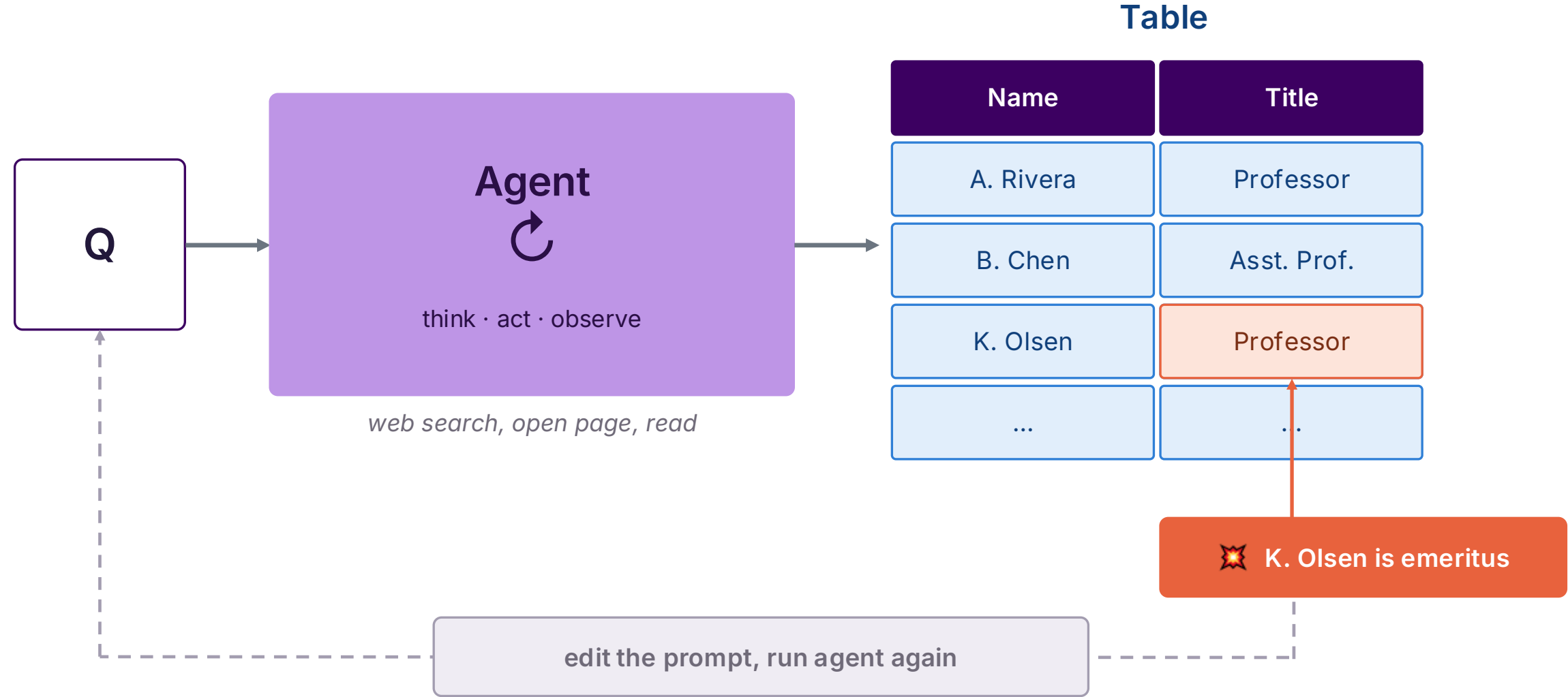
Many systems: **WideSearch** [2] **Table as Search** [3] **Deep Research** [4] **A-MapReduce** [5]

[1] Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023
[2] Wong et al. WideSearch: Benchmarking Agentic Broad Info-Seeking. ICLR 2026
[3] Lan et al. Table-as-Search: Formulate Long-Horizon Agentic Information Seeking as Table Completion. ACL Findings 2026
[4] OpenAI. Deep Research System Card. February 2025
[5] Chen et al. A-MapReduce: Executing Wide Search via Agentic MapReduce. arXiv 2026

Agentic data engineering



Agentic data engineering



Observation: **repairs** are systematic

What the agent extracted from cis.upenn.edu/people

Name	Title	on the page
M. Diaz	Professor	under "Emeritus"
A. Rivera	Professor	under "Faculty"
B. Chen	Asst. Prof.	under "Faculty"
K. Olsen	Professor	under "Emeritus"

Same page, same prompt,
same mistake in every row.
One wrong row tells you
about all its "siblings".

A filter that drops one valid row drops others for the same reason.

Motivation

1 Problem: ambiguity in data and question; repair requires a re-run

2 Observation: repairs are systematic, one correction could be applied to others

 Record where every cell came from while the agent runs, then repair from one correction 

provenance - a record, written while the agent runs, of which page and which operator call produced every row and every cell.

The missing piece is not a better model, but a harness-level record.

Why a harness

Provenance by construction

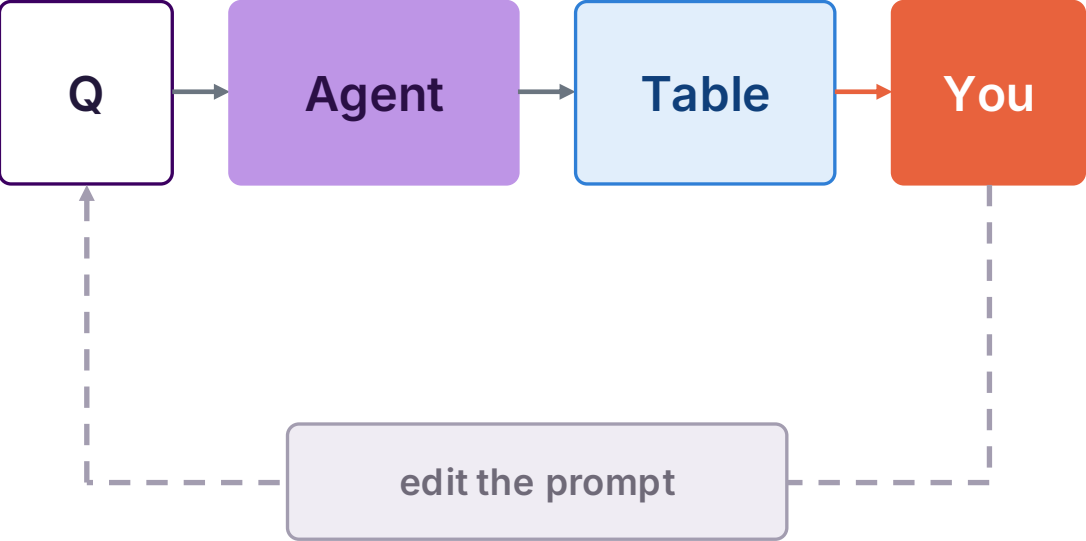
Repair from one correction

Results

Does a stronger model solve this

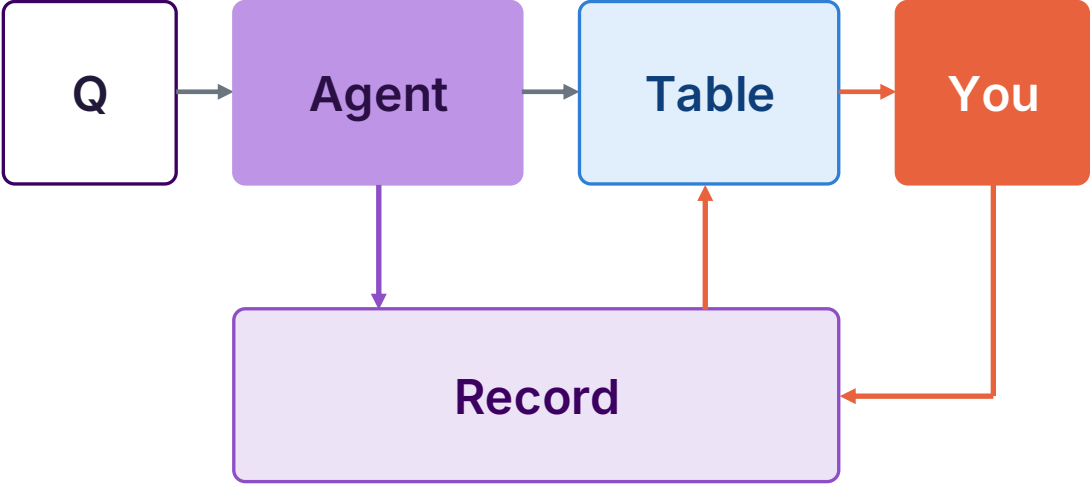
Data Canvas in one figure

without a record



re-run the agent

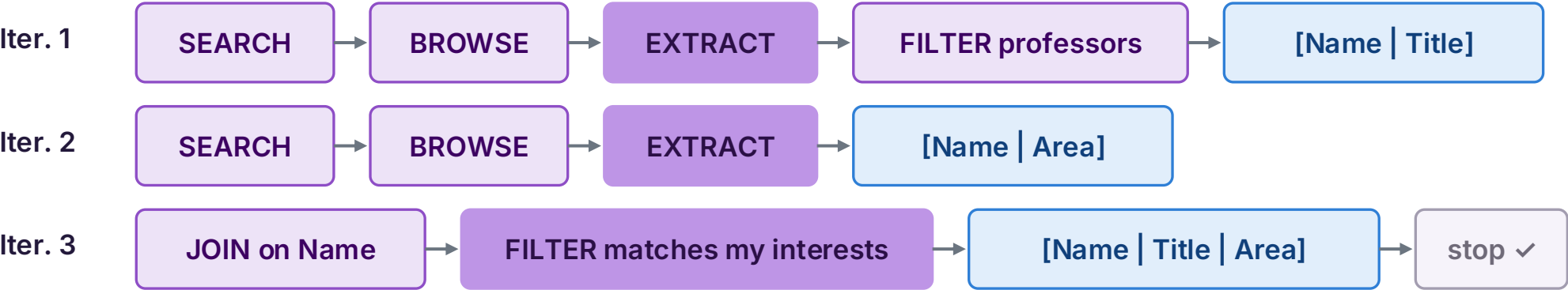
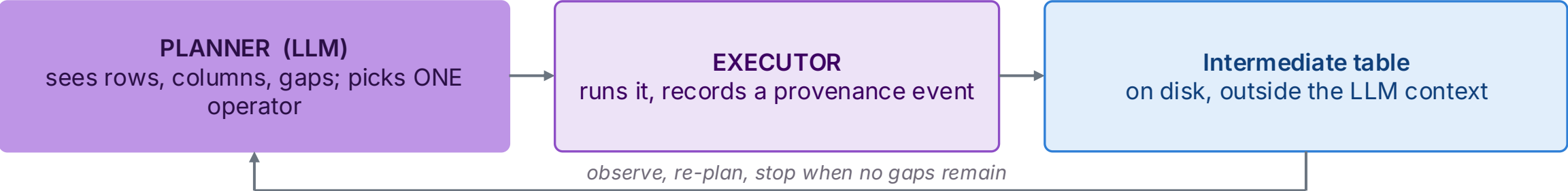
with a record



only the affected rows re-run

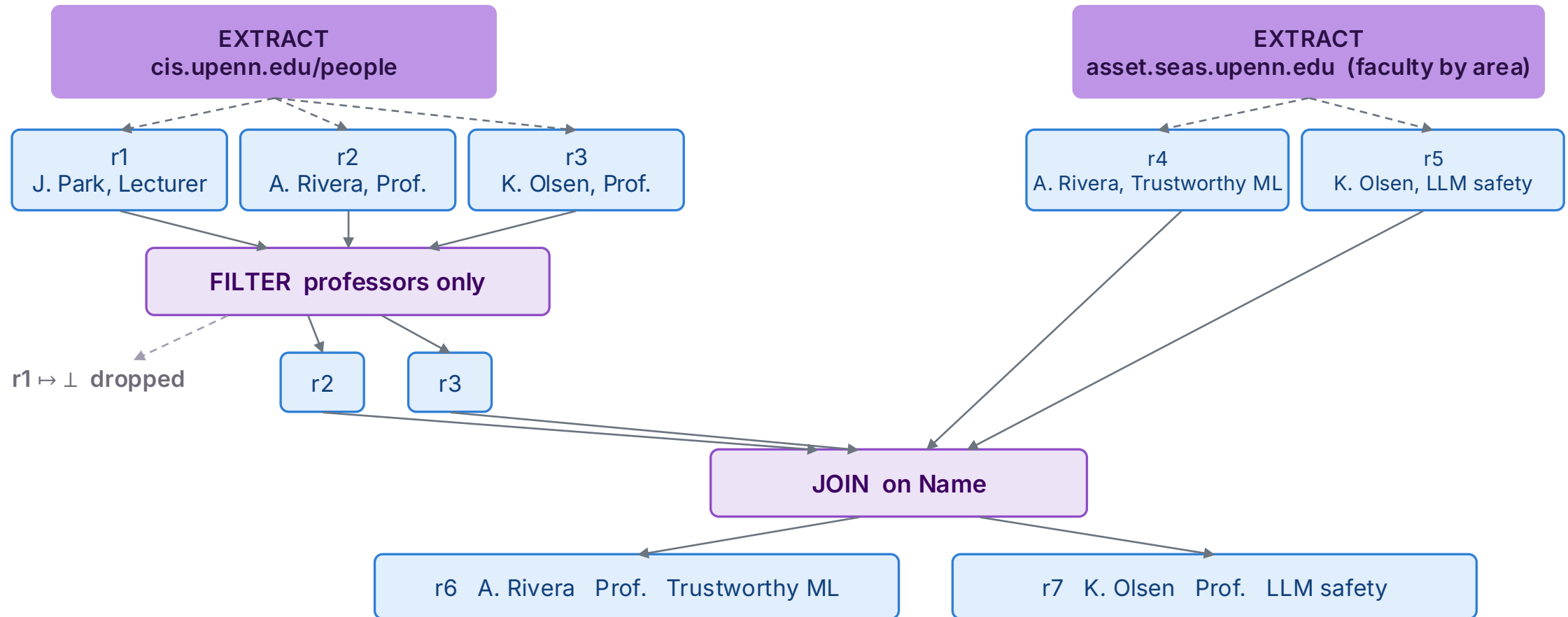
Build the table one operator at a time

Task *"Find all CS professors from US R1 universities that match my research interests"*



10 semantic operators [6] in all: SEARCH, BROWSE, EXTRACT, FILTER, PROJECT, MERGE, JOIN, UNION, SORT, MAP.

Every operator writes down what it did



Each call records the page it read and the rows it produced. Cheap to write, exact to query.

Why a harness

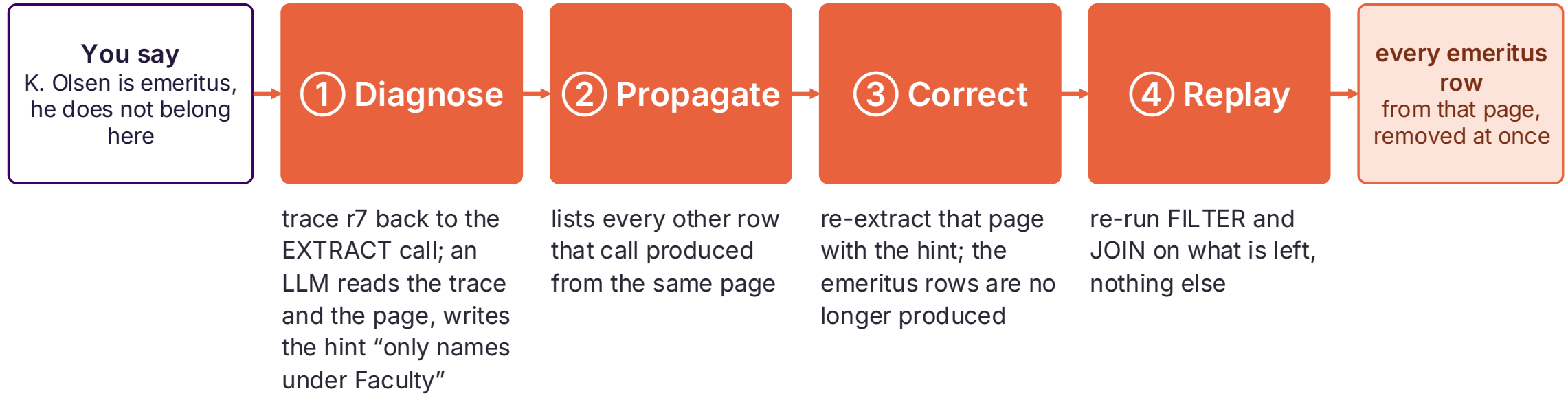
Provenance by construction

Repair from one correction

Results

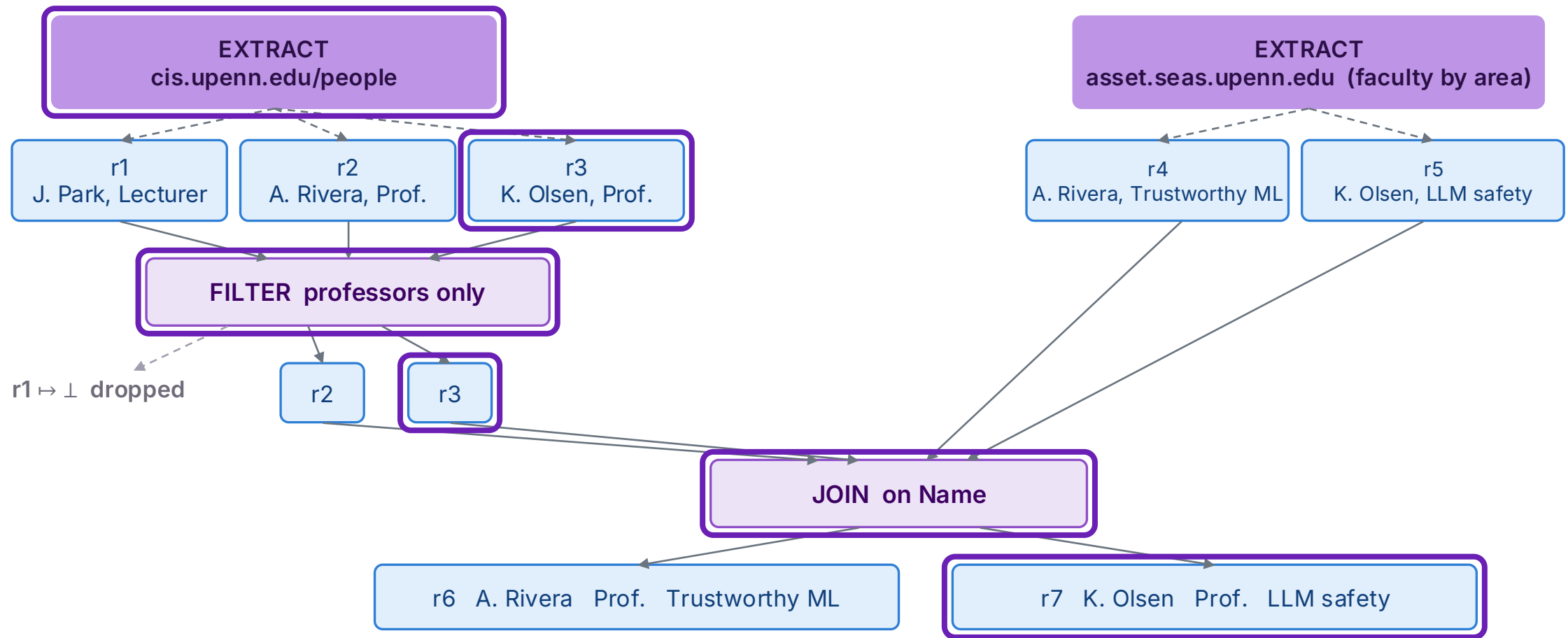
Does a stronger model solve this

One correction: diagnose, propagate, correct, replay



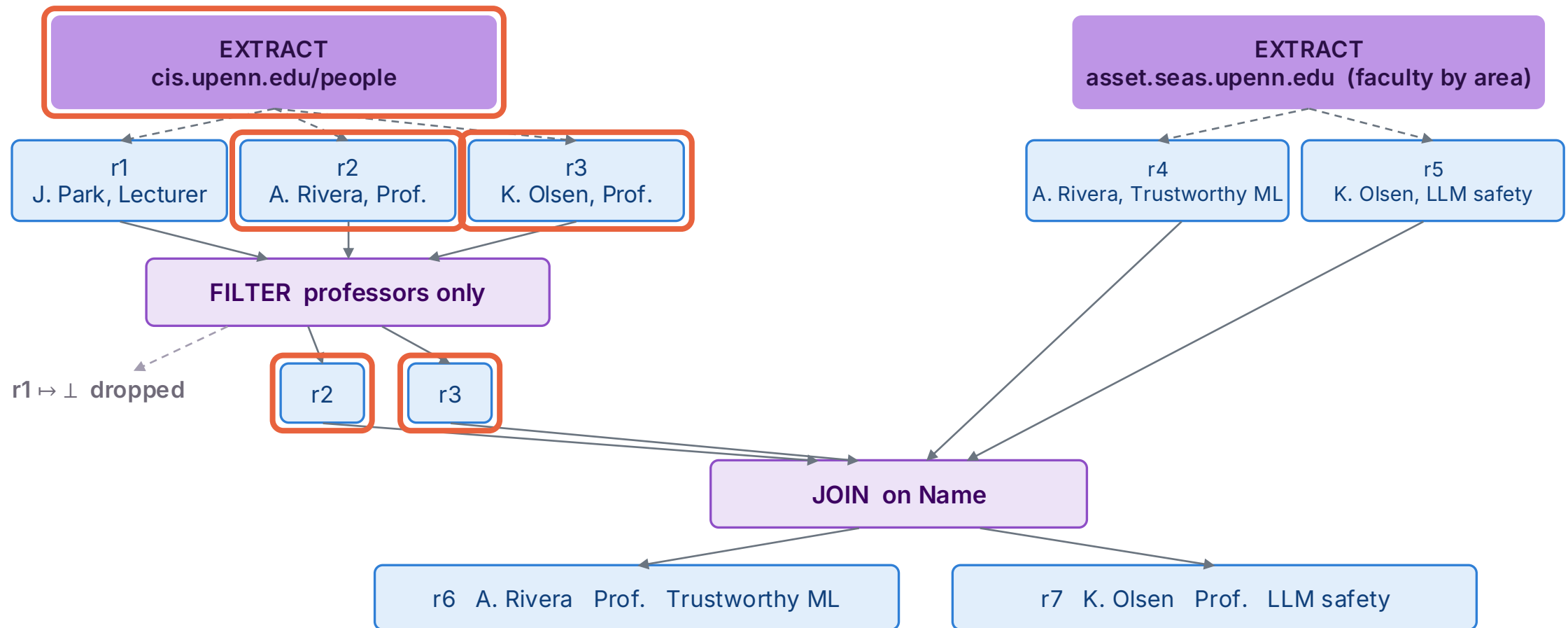
Feedback becomes an edit to the run, not to the prompt

So two questions have exact answers



Trace, backward from the wrong row r7, follow back through JOIN and FILTER to the EXTRACT call and the page it read.

So two questions have exact answers



Impact, forward from that call, the record lists every other row it produced from the same page: r2 here, and every other name taken from the Emeritus section.

What the diagnoser finds, and what it does

Extraction ambiguity the page was right, the reading was not: a heading ignored, a wrong column, ...

→ re-extract the siblings with a hint

Filter too strict a valid row was dropped

→ restore it, then re-judge that filter's other drops

Merge conflict deduplication kept the wrong value

→ re-resolve the group with the correction as tie-breaker

Wrong source the page does not describe the target population at all

→ drop that page's rows and replay downstream

Why a harness

Provenance by construction

Repair from one correction

Results

Does a stronger model solve this

Results on the WideSearch benchmark

Method	Row F1	Item F1	Entity F1	Cell acc.	Time (s)	Cost (\$)
WideSearch (GPT-5) [2]	21.8	43.4	52.6	73.3	830	1.51
Deep Research (o3) [4]	11.7	20.2	23.3	29.2	268	0.83
TaS (GPT-5) [3]	25.6	49.6	60.4	73.5	3,940	18.27
TaS (GPT-5-mini) [3]	17.1	37.3	49.3	68.8	1,835	1.93
Data Canvas (GPT-5)	26.6	52.8	63.2	74.0	2,276	1.88
Data Canvas + repair (GPT-5)	+2.3	+3.1	+3.4	+0.2	+235	+0.05

WideSearch benchmark [2], English subset, 100 queries. Repair uses 1% of the wrong cells as feedback; its time and cost are the additional cost on top of the row above.

Operators alone already beat the best baseline: +4.8 row F1 and +9.4 item F1 over WideSearch at about the same cost (\$1.88 vs \$1.51). TaS reaches similar row precision for 9.7 times the cost.

Repair adds +2.3 / +3.1 / +3.4 F1 for 5 cents and 4 minutes. Re-running with the hint in the prompt would cost \$1.88 and 38 minutes, 35 times more.

[2] Wong et al. WideSearch: Benchmarking Agentic Broad Info-Seeking. ICLR 2026

[3] Lan et al. Table-as-Search: Formulate Long-Horizon Agentic Information Seeking as Table Completion. ACL Findings 2026

[4] OpenAI. Deep Research System Card. February 2025

Why a harness

Provenance by construction

Repair from one correction

Results

Does a stronger model solve this

Does a stronger model solve this?

Yes, for accuracy better planning and search will shrink our margin over the baselines.

Not for intent half the errors are not model errors. No model knows what you meant until you see its output.

Not for attribution a million-token model can read the log and guess which cells share a cause. The graph knows.

And one thing gets worse the better the model, the less anyone checks. A silent systematic error costs more, not less.

Claude Fable 5.1, GPT-6 Astra, whatever ships next: rarer errors, not attributable ones

★ Take-aways

Users don't know what they want until they see the data.

Ambiguity in data and question. Feedback after output is necessary.

👍 **Provenance by construction.** Every operator records which page and which call produced every cell.

👍 **One correction repairs its siblings.** Trace, propagate, correct, replay only what changed.

A stronger model makes errors rarer, not attributable.

Thank you

Data Canvas: record provenance as the agent works, so one **"that's wrong"** repairs every cell that shares its cause

Zixuan Yi zixy@cis.upenn.edu zixy17.github.io



zixy17.github.io