THE READER OF SCIENCE IS BECOMING AN AGENT

Agent-Native Research *Artifacts*

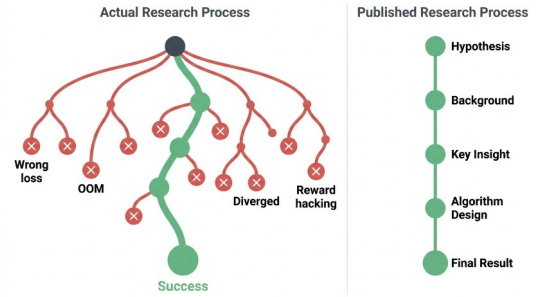
When the reader is an agent, the *paper* is the bottleneck.

Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, Carl Chen, Zhiyang Chen, Haojie Ye, Yujuan Fu, Zexue He, Zijian Jin, Zhenyu Zhang, Shangquan Sun, Maestro Harmon, Dianzhuo Wang, Qian-ze Zhu, Jianqiao Zeng, Jiachen Sun, Mingyuan Wu, Baoyu Zhou, Chenyu You, Shijian Lu, Yiming Qiu, Fan Lai, Yuan Yuan, Yao Li, Junyuan Hong, Ruihao Zhu, Beidi Chen, Alex Pentland, Ang Chen, Mosharaf Chowdhury, Zechen Zhang

A multi-institution collaboration

A paper PDF is a *lossy compression* of the research.

Storytelling Tax



The branching exploration and dead ends are compressed into a single linear story.

Engineering Tax

An agent-executable spec is compressed into reviewer-sufficient prose.

45.4%

of reproduction requirements are fully specified in the PDF; the agent guesses the rest.

The bottleneck is the artifact, not the prompt.

Human AI co-scientist

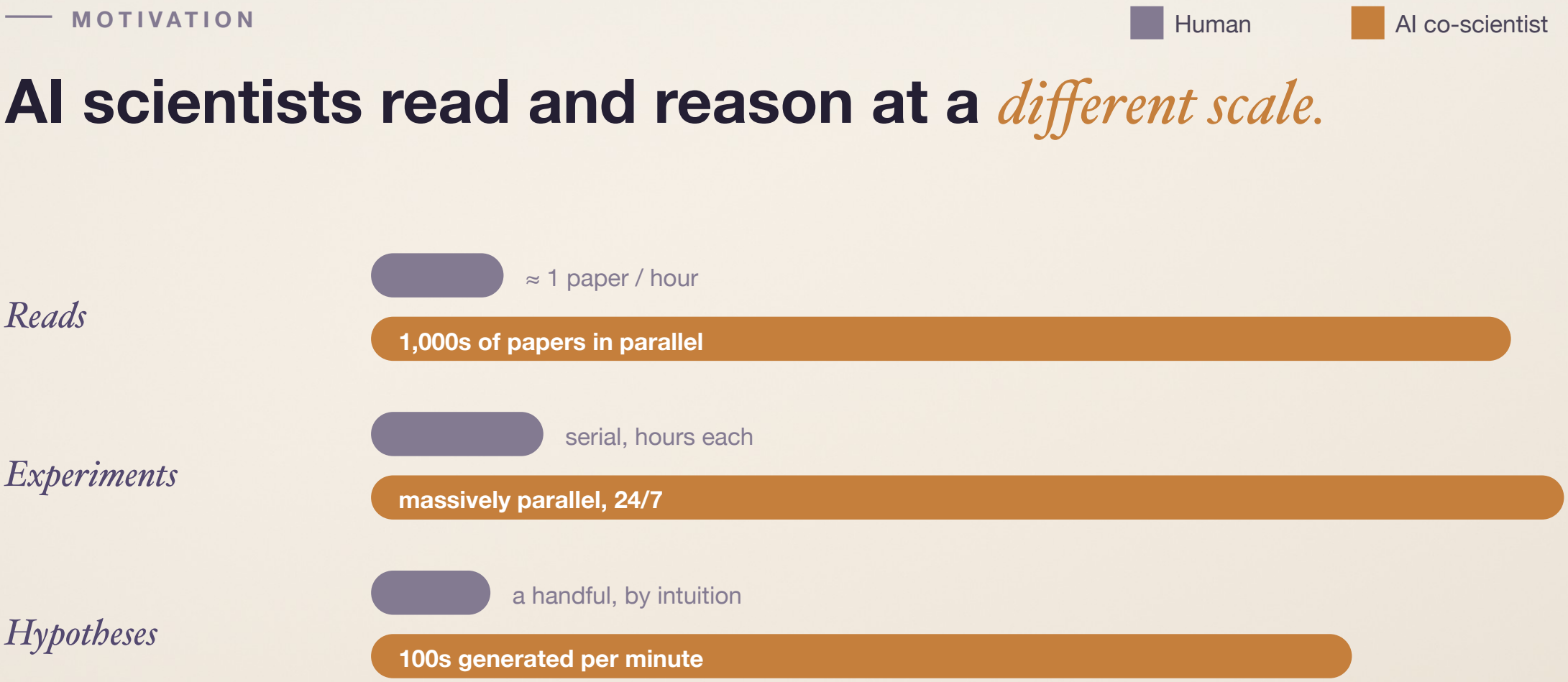
AI scientists read and reason at a *different scale.*

Reads — ≈ 1 paper / hour

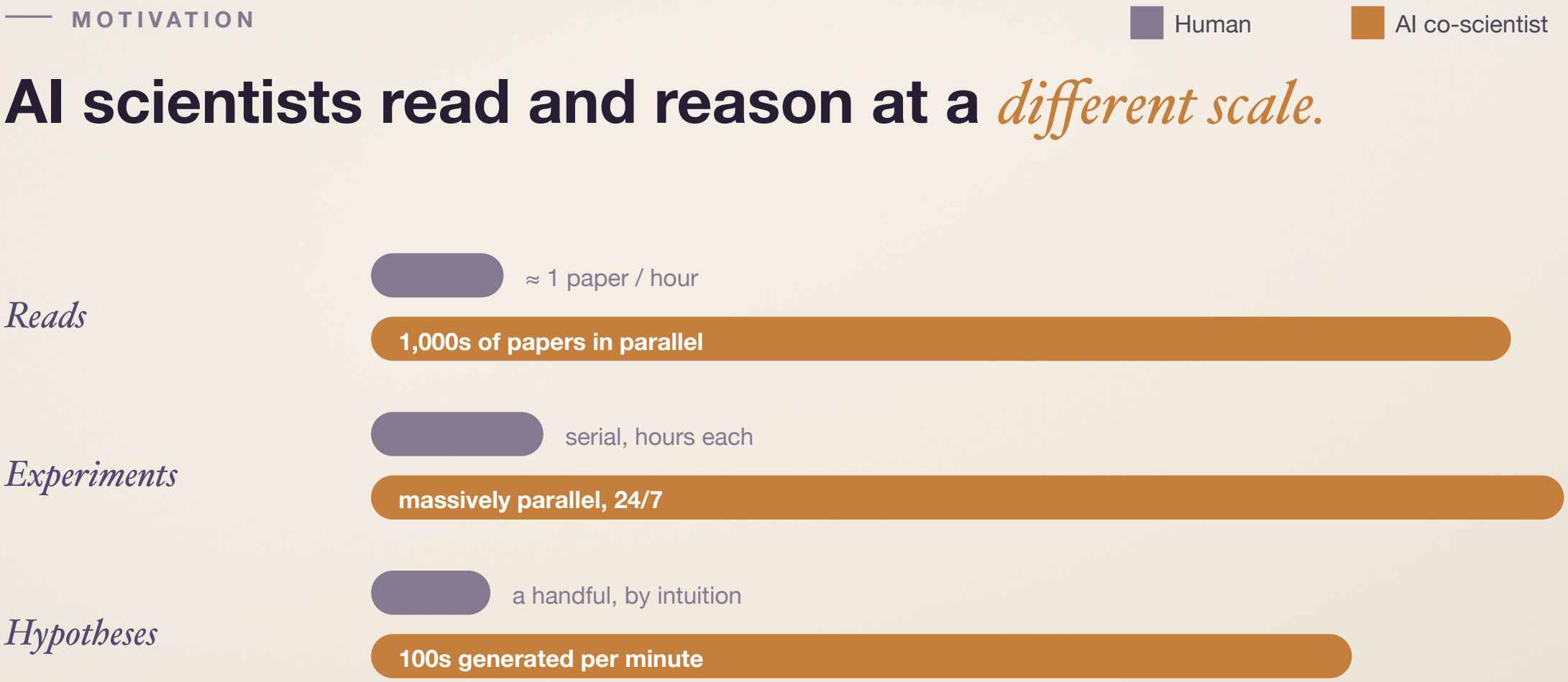
Experiments — serial, hours each

Hypotheses — a handful, by intuition

AI scientists read and reason at a *different scale.*



AI scientists read and reason at a *different scale.*



Different scale, different modality: AI wants structured, executable artifacts, not narrative.

Introducing ARA, *a research package built for agents.*

~/research/ara

/logic *the why*

Claims, hypotheses, lineage; typed and machine-readable.

/src *the how*

Code, configs with rationale, environment pins, seeds.

/trace *what was tried*

DAG of decisions, dead ends, pivots; the git log of research.

/evidence *the raw numbers*

Logs and results; separable for access control and verification.

Understandable

Hierarchical structure that preserves the full exploration trajectory.

Executable

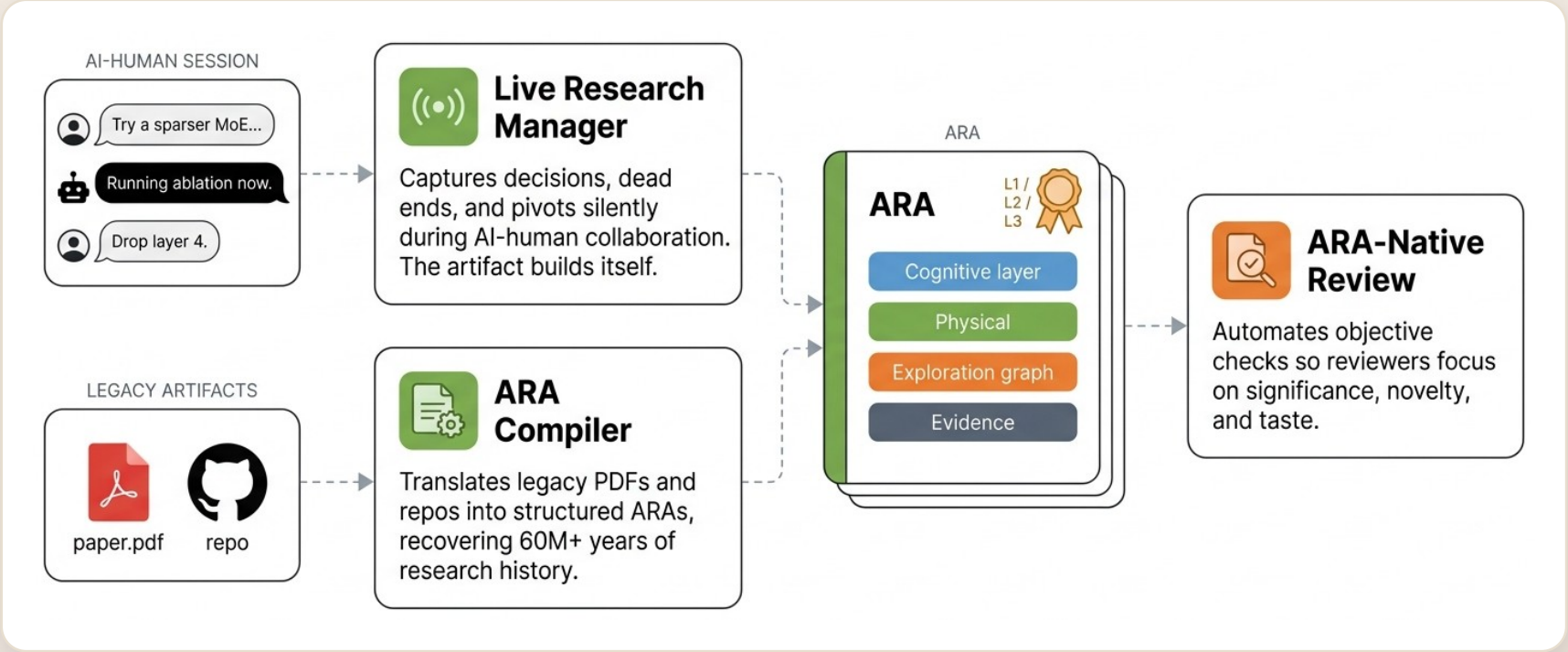
An agent reproduces the claim zero-shot, not reimplements it.

Verifiable

Every ARA is auditable and re-runnable by the community.

Agents author, verify, and *extend one artifact.*

Too rich to fill by hand, so agents author, verify, and render it as a byproduct of research.

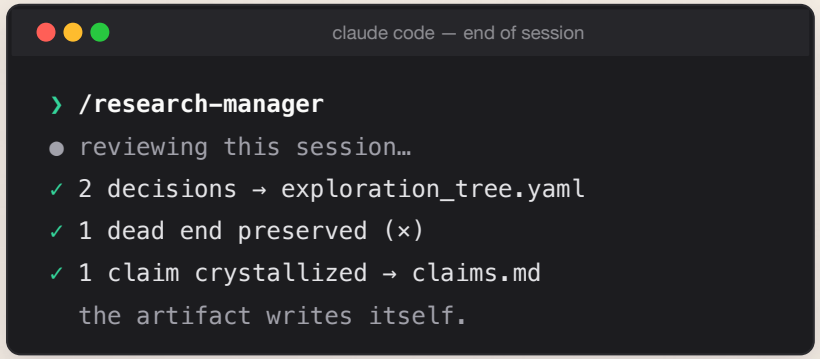


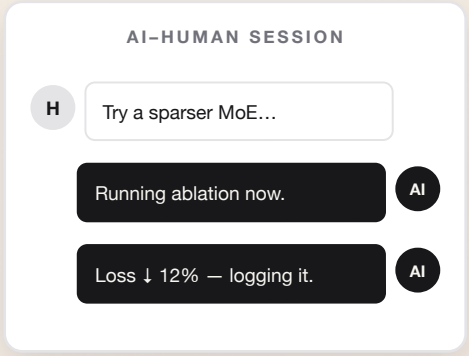
One loop, one persistent state: the ARA is the only thing humans and research agents share.

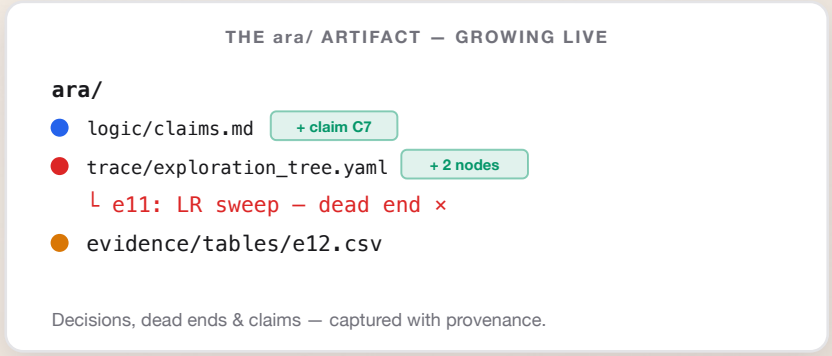
Capture live, or *compile the past.*

1 /research-manager

Rides along as an observer, documenting the ARA while the AI does the research.

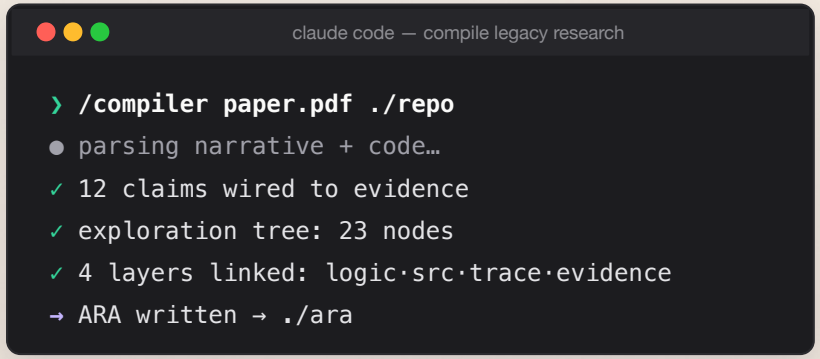


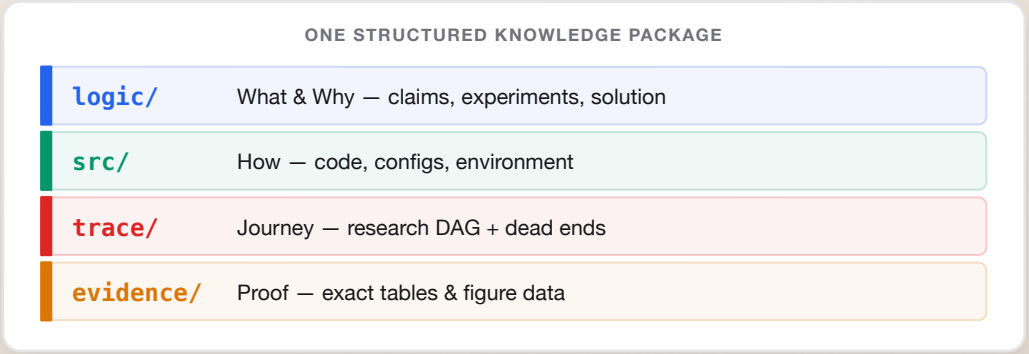




2 /compiler

Compile any paper, repo, or notes into a structured ARA.





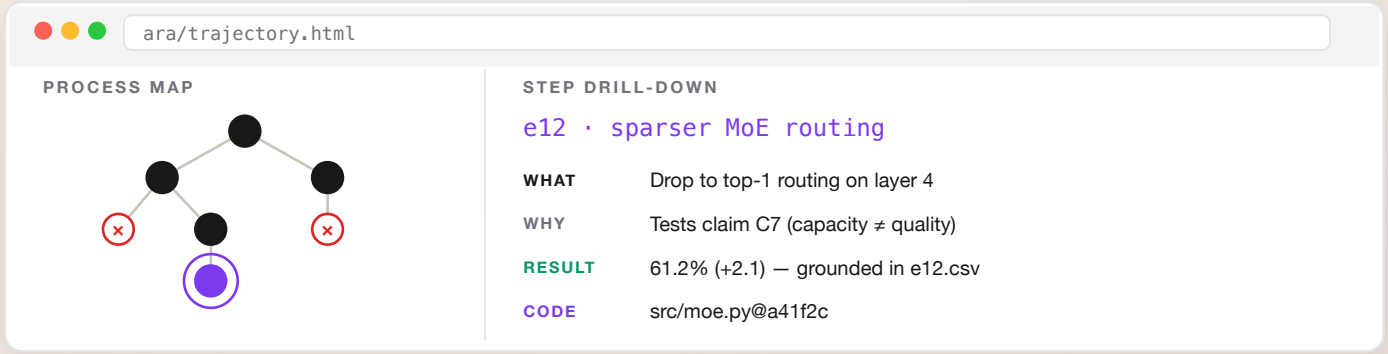
Replay it, then *audit its rigor.*

3 /research-visualizer

Watch the whole trajectory, every branch and dead end, in one interactive map.

```
claude code — render trajectory

❯ /research-visualizer ./ara
● reading exploration graph…
✓ 23 steps · 4 dead ends mapped
✓ trajectory.html — one file, interactive
→ open in browser to replay the research
```

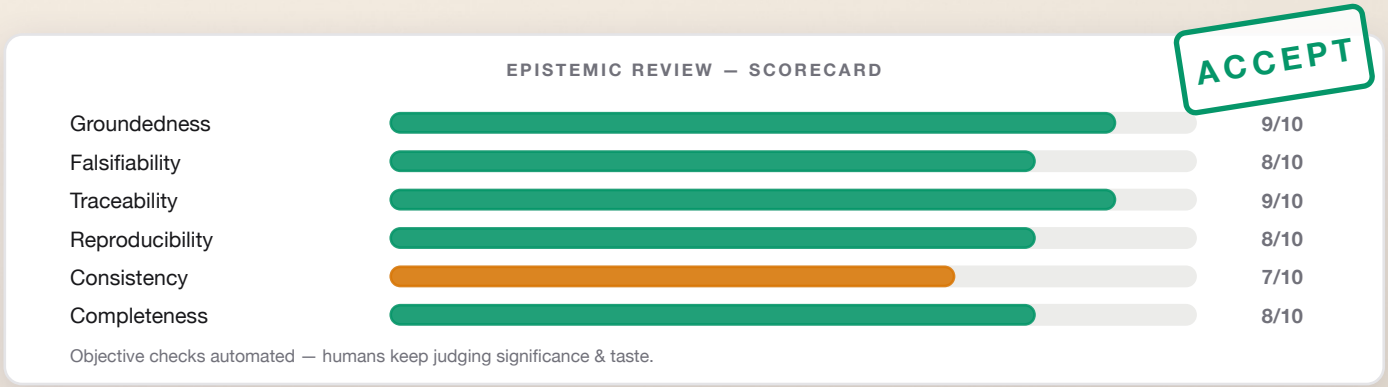


4 /rigor-reviewer

Audit an artifact's epistemic rigor before you trust or publish it.

```
claude code — seal level 2

❯ /rigor-reviewer ./ara
● level-2 semantic review…
✓ 6 epistemic dimensions scored
⚠ 2 findings (minor severity)
→ recommendation: Accept
```



How we test it: *three layers, two benchmarks.*

TWO BENCHMARKS SUPPLY WHAT PDFs LACK

PaperBench

Fine-grained rubrics decomposing paper reproduction into gradable requirements.
54.6% are underspecified in the source PDF.

RE-Bench

Close-ended AI R&D hill-climbing tasks on well-defined objectives, spanning engineering and research.
90.2% of agent effort is spent on discarded dead ends.

THREE LAYERS WE TEST

I	*Understanding* Can an agent extract knowledge?	`450 paired QA`
2	*Reproduction* Can it execute the research?	`PaperBench · 1,743 reqs`
3	*Extension* Can it build on prior work?	`RE-Bench · 5 tasks`

Every comparison is ARA vs. PDF + repo with agent, task, and ground truth held fixed · Claude Sonnet 4.6 · blinded judge · ARAs auto-compiled.

ARA wins on *all three layers.*

UNDERSTANDING

72.4 → 93.7%

QA accuracy

+65.7% on failure knowledge, where the PDF baseline has no source

REPRODUCTION

57.4 → 64.4%

success rate

lead grows monotonically with task difficulty

EXTENSION

5 / 5

faster first move

reaches a useful first move earlier on all five; wins 3/5 on final score

THE TAKEAWAY

Don't write the paper. Ship the *artifact.*

Research that agents can read, run, and verify.

`github.com/ARA-Labs/Agent-Native-Research-Artifact`

Scan: code + spec