Lecture: Multi-Agent Systems and Orchestration

Building a research harness

Boris Bolliet

NASA AI/ML Science & Technology Interest Group
Lecture Series

16 November 2026

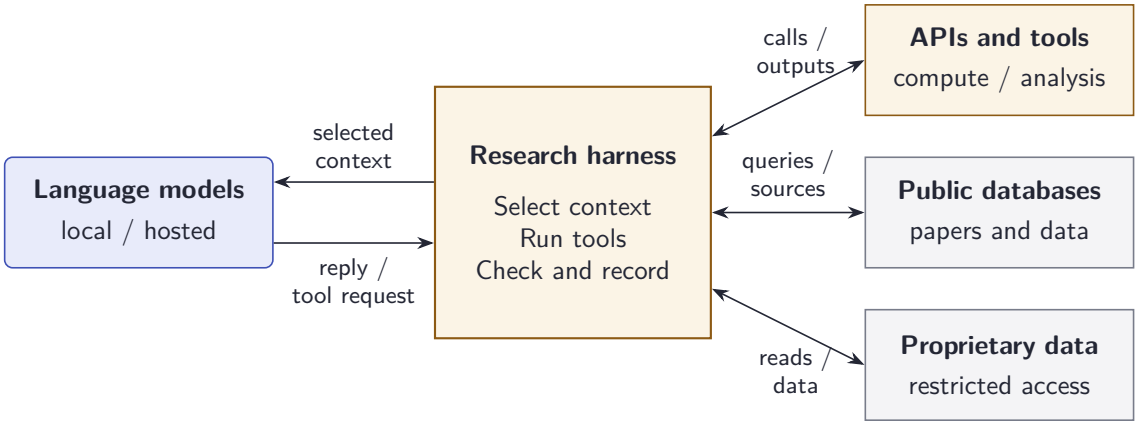


Control over the research workflow

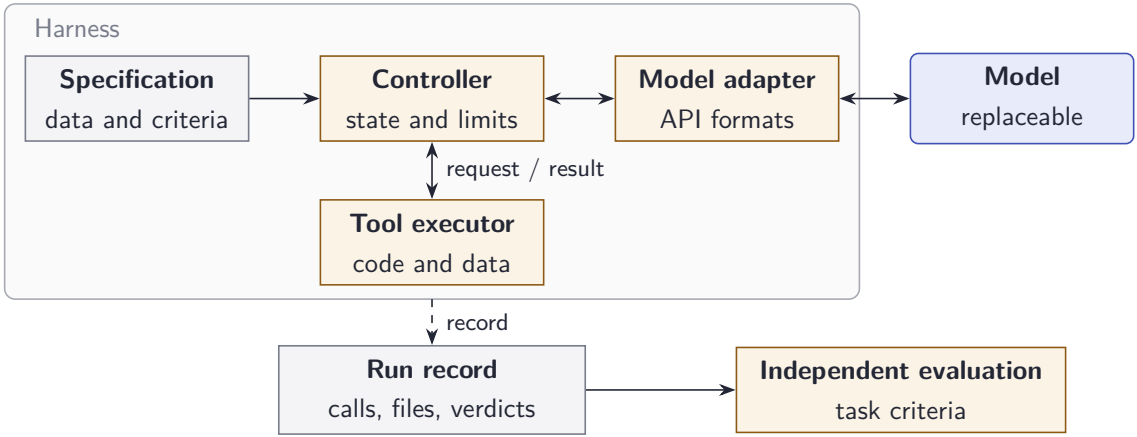
- **Traceability:** connect results to sources, code and checks.
- **Sovereignty:** control where computation runs and data go.
- **Customisation:** add domain tools and checks.
- **Portability:** replace models and services.

Evaluate gains in correctness, cost and time.

The research harness



Models, tools and recorded evidence



Solid: runtime exchanges. Dashed: recording.

Roles and context

Shared **state** stores the task, messages, files and results.
Each call receives a selected **context**.

Role	Information supplied
Engineer	Requirements, data, previous errors
Reviewer	Requirements, code, executed outputs
Writer	Accepted results and supporting evidence

Record what each call actually received.

Structured handoffs

A schema lets the controller read a review as data.

```python
from pydantic import BaseModel

class Review(BaseModel):
    approved: bool
    issues: list[str]
```

- The controller reads `approved` and routes the next step.
- The engineer receives `issues` when revising.
- Schema validation checks form; evaluation checks judgement.

Execution and termination

The executor records the script, output, exit status and saved files.

- Return errors for correction while attempts remain.
- Bound attempts, elapsed time and cost.
- Preserve an unfulfilled outcome when execution stops.
- Restrict generated code's access to files and credentials.

Tracing with Langfuse

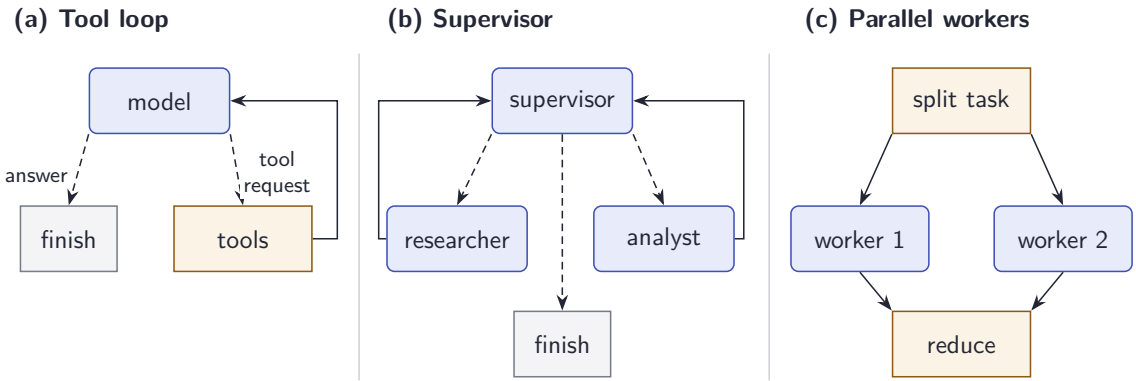
A **trace** nests operations; model calls are **generations**. A **session** groups traces.

```
from langfuse.langchain import CallbackHandler

graph.invoke(inputs, config={
    "callbacks": [CallbackHandler()],
    "metadata": {"langfuse_session_id": "research-run"}})
```

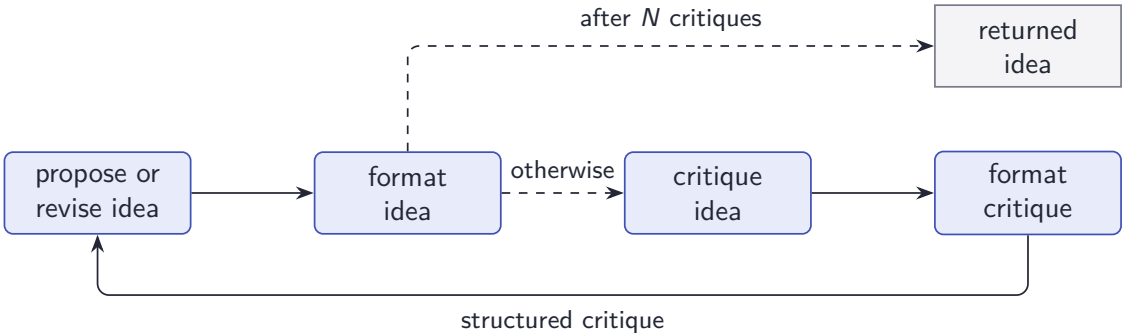
Label calls by agent; record inputs, outputs, usage and tool outcomes.

Control in LangGraph



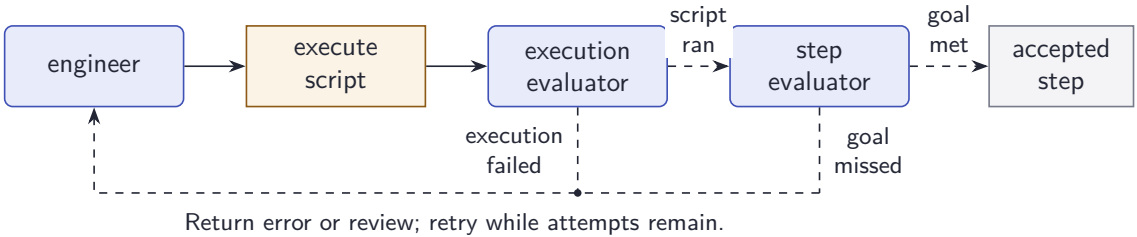
Rounded: model calls. Solid: fixed paths. Dashed: alternatives.

Propose–critique



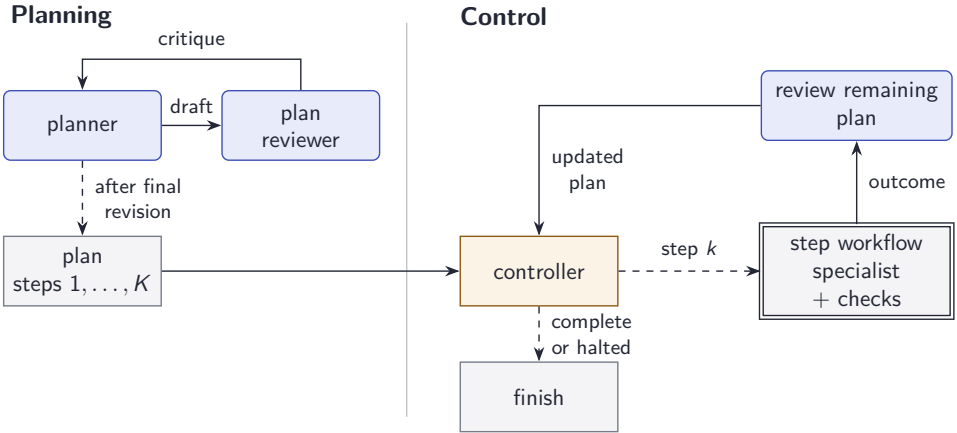
The final critique receives a revision and a formatting pass.

Self-debugging



Both evaluators share one attempt limit. Exhaustion leaves an unfulfilled step.

Planning and control

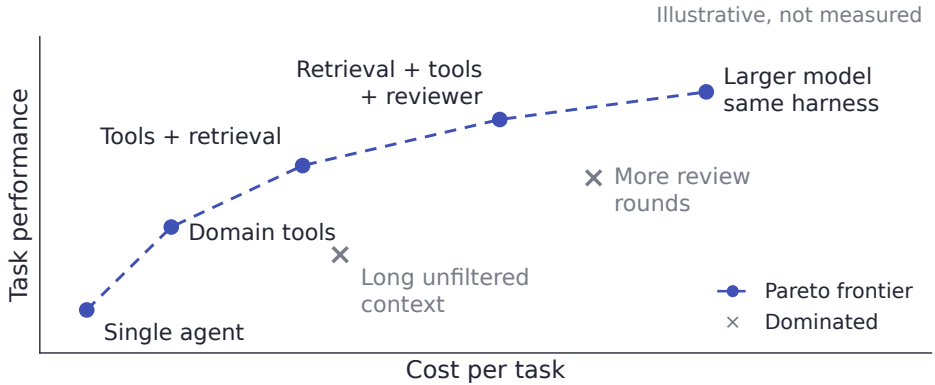


Test cases and checks

Proposed test	Expected behaviour
Reference fit	Check result and method
Wrong covariance	Reject script
Missing input	Record unresolved step
Unsupported claim	Reject or flag report
Interrupted run	Preserve outputs and failure

Fault injection: deliberately wrong scripts or reports test the checker.

Comparing configurations



Repeat held-out cases; count failed attempts in the cost.

Datasets in Langfuse

Each item stores:

- **Input:** task and data locations.
- **Expected output (optional):** independently checked reference.
- **Metadata:** case type and provenance.

Create in the interface or SDK. Keep the dataset version fixed across comparisons.
Reference values go only to the evaluator.

Experiments in Langfuse

`task` runs the harness; `fit_match` scores numerical agreement.

```
from langfuse import get_client

client = get_client()
dataset = client.get_dataset("pantheon-checks")
result = dataset.run_experiment(
    name="reference-fit-v1", task=task,
    evaluators=[fit_match], max_concurrency=1)
client.flush()
```

Compare scores, cost and time. Inspect individual traces.
Method checks need separate evaluators.

CMBAgent

- Introduced for cosmological parameter analysis.
- Extended to autonomous planning and control beyond cosmology.
- First-place FAIR Universe award: autonomous exploration with human intervention.
- LangGraph implementation: checked steps, bounded retries, replanning and report audits.

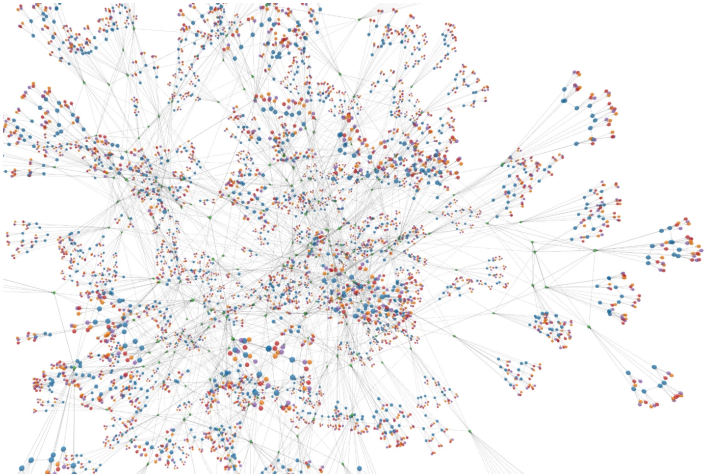
Denario

Ideas → literature → methods → analysis → papers.

CMBAgent supplies analysis.

Repeated research cycles build knowledge graphs linking projects and their results.

Ecology graph: detail.



Papers, ideas, methods, results, evaluations, datasets.

OpenCode

- A coding agent reads, edits and executes files.
- The model delegates through a task tool; subagents have separate contexts.
- Session exports record tools, usage and cost.

ACT DR6 reanalysis: Qwen and DeepSeek served locally on one RTX 6000.[1]

[1] ACT DR6 replication: Licong Xu.

Skepthical

Parallel reviewers check general issues, mathematics, numerics and citations.

- Tools: computer algebra, executed code and retrieved papers.
- Benchmark: 357 inserted errors across 30 papers, plus clean copies.
- Highest astrophysics citation detection among tested systems.
- No established citation advantage over the best single-model reviewer on ML.

CLArXivR

Paper recommendations grounded in a reader's interests and library.

- Rank candidates before fetching full papers.
- Summarise findings and explain relevant connections.
- Evaluate support, coverage and numerical fidelity separately.
- Corrupt passages to test graders; evaluate prompts on held-out papers.

Work led by G. Farren.

The Pantheon+ task

Fit matter density Ω_{m} in flat ΛCDM.

- $z_{\mathrm{HD}} > 0.01$: 1590 light curves, 1473 supernovae.
- Full statistical and systematic covariance.
- Profile over absolute magnitude; derive uncertainty from $\Delta\chi^2 = 1$.

Independent reference, withheld from agents:

$$\Omega_{\mathrm{m}} = 0.3316 \pm 0.0182.$$

Agents receive the complete analysis specification.

Three harness configurations

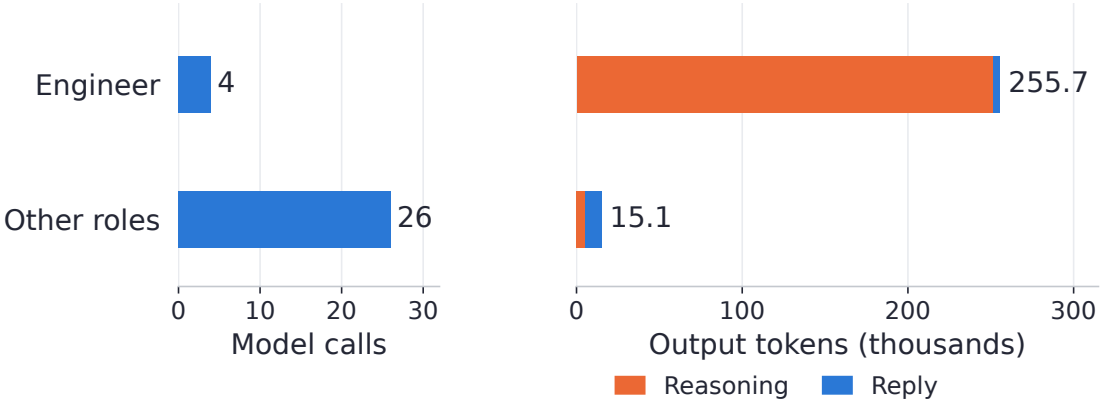
All roles: `gemini-3.1-flash-lite`.

	Matches	Calls	Time (s)	Cost (US$)
LangGraph single agent	8/10	3–10	7.3–25.8	0.003–0.022
CMBAgent	2/2	30–31	558–588	0.426–0.440
OpenCode + reviewer	2/2	17–23	28–50	0.025–0.045

Matches: estimate and uncertainty within 0.005.

Different price sources and reasoning; no general ranking.

Reading the trace: calls and reasoning



CMBAgent run 1: four engineer calls produce 94% of output tokens.

Checking the reported result

The notebooks grade different outputs:

Harness	Graded output
Single agent	Final answer
CMBAgent	Script's last successful `RESULT`
OpenCode	`report.md`

A numerical match does not establish that the method or report is correct.

Long-horizon research

- ACT DR6 reanalysis: maps, spectra, systematics and cosmological inference.[2]
- OpenAI reports a Navier–Stokes Millennium Prize solution, with a Lean formalisation.
- Multi-day work needs persistent artefacts, checks and recoverable execution.
- Parallel Science: infrastructure for papers and reviews.

[2] ACT DR6 replication: Licong Xu.

Evolving research harnesses

- Preserve tasks, code, traces and evaluations alongside papers.
- Denario v2 (in development): evaluate adaptations learned from past trials.
- Impact-aware research selects questions by expected scientific contribution.
- A science of science studies discovery and influence.[3]

[3] Citation-network animation by Licong Xu.