INFORMAL SHARE

Models, harnesses and a notebook



model



harness



notebook
(task record)



Theme 1

Choosing the right model



Theme 2

Agent = LLM + harness

Zheng Zhang

A learning method in daily life

LLMs are changing everyday life.

A machine-learning method

`p(next token | context)`



A change in daily work

writing · coding · reading · deciding

A learning method in daily life

LLMs are changing everyday life.

A machine-learning method

`p(next token | context)`

→

A change in daily work

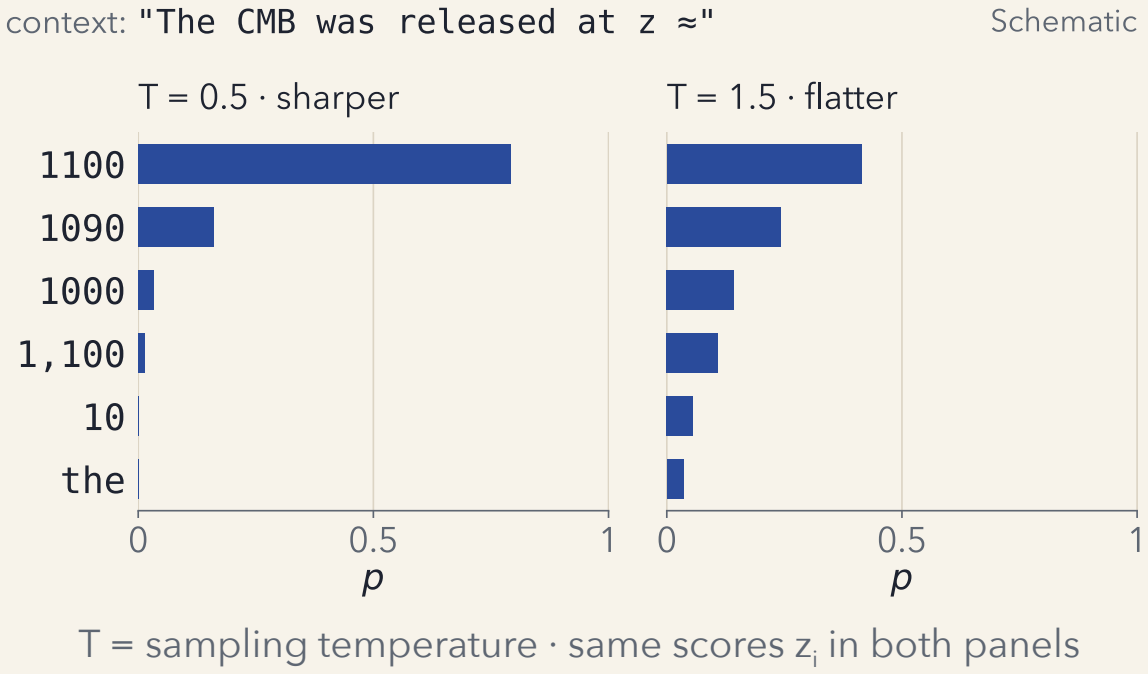
writing · coding · reading · deciding

Why?

The next-token model

An LLM predicts a continuation.

```
p(next token | context)
```

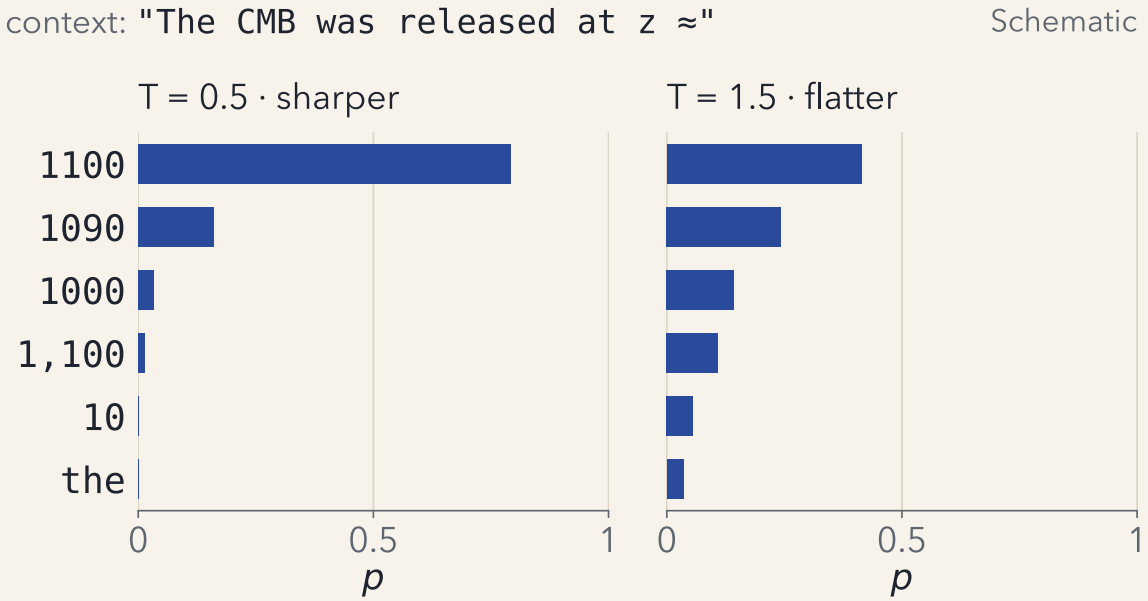


The next-token model

> An LLM predicts a continuation.

$$p(\text{next token} \mid \text{context})$$

$$p_i \propto \exp(z_i / T)$$



T = sampling temperature · same scores z_i in both panels

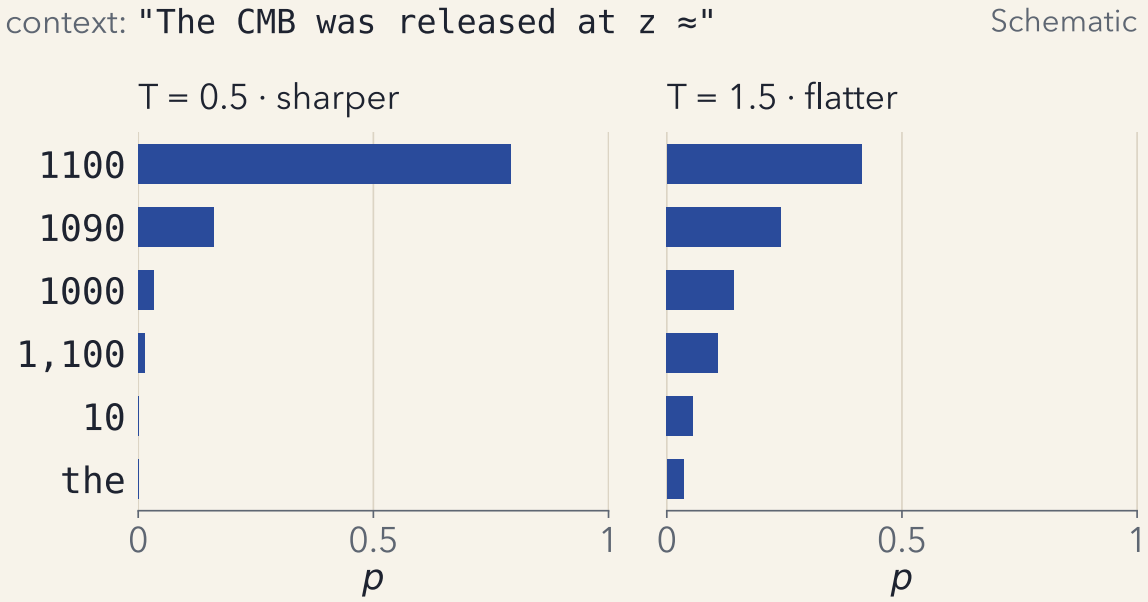
The next-token model

An LLM predicts a continuation.

$$p(\text{next token} \mid \text{context})$$

$$p_i \propto \exp(z_i / T)$$

Cross-entropy ↔ compression



T = sampling temperature · same scores z_i in both panels

The next-token model

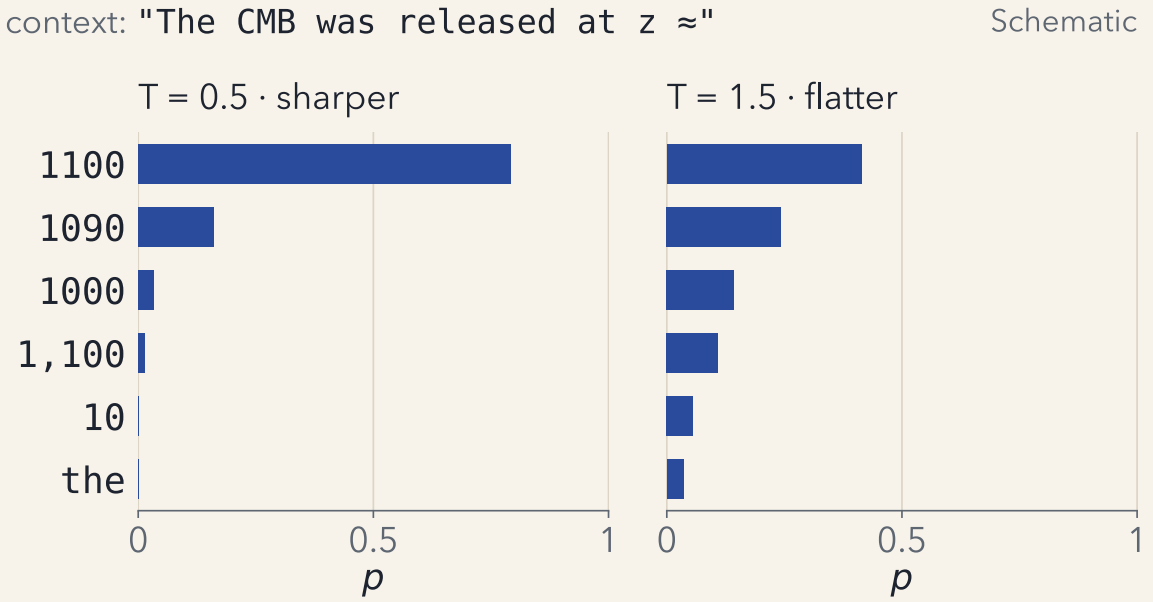
An LLM predicts a continuation.

`p(next token | context)`

$p_i \propto \exp(z_i / T)$

Cross-entropy ↔ compression

Training → post-training → generation



T = sampling temperature · same scores z_i in both panels

The capability debate

There may be a ceiling; LLMs will still change daily life.

OPEN QUESTION

Is text enough?

Other bets: world models, spatial intelligence

EVEN SO

One of the most important moments

Three reasons

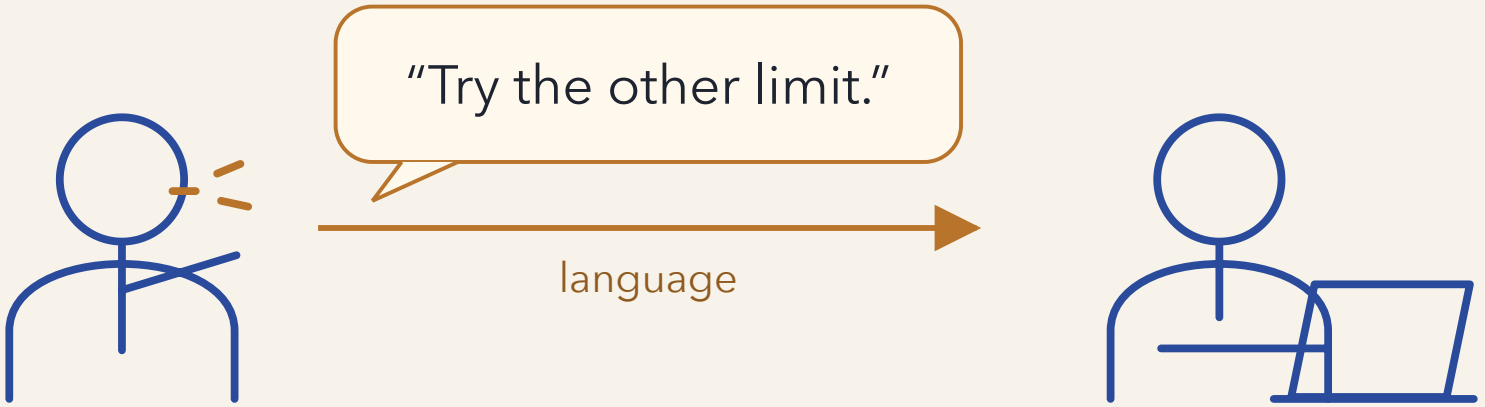
1 Language

2 Recombination

3 Action

Language connects worlds

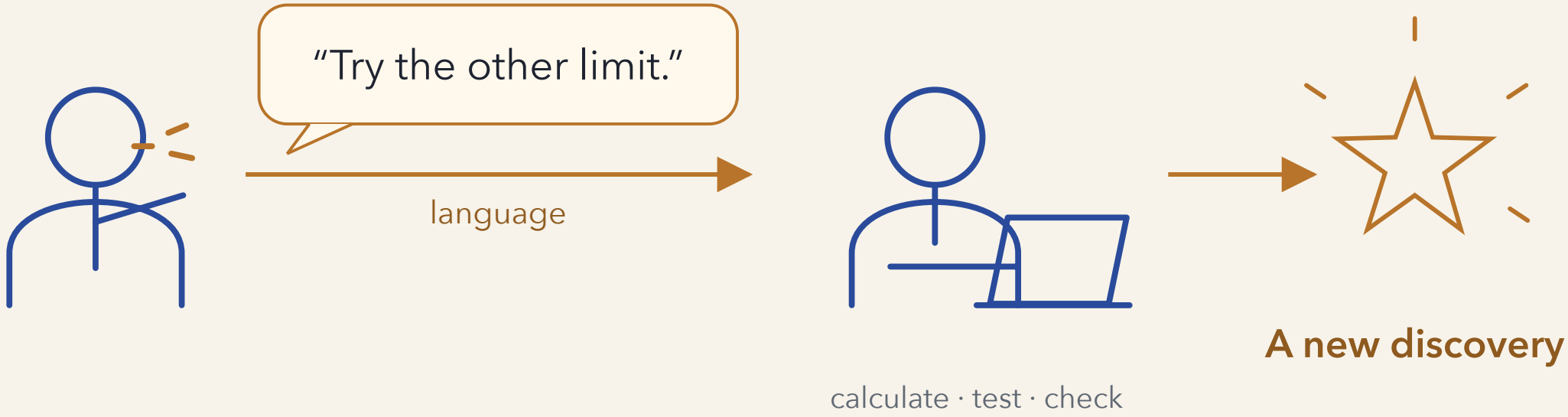
> Words connect ideas, mathematics, code and observations.



calculate · test · check

Language connects worlds

Words connect ideas, mathematics, code and observations.



What starts as words can end as something new.

Recombination

> Recombining what is already known goes further than expected.

Literature search

Erdős problems

existing solutions found in papers

Erdős episode, Oct 2025 (TechCrunch, 19 Oct 2025) · Bubeck et al., Nov 2025, arXiv:2511.16072, Sec. I.2 and III.2

Recombination

> Recombining what is already known goes further than expected.

Literature search

Erdős problems

existing solutions found in papers

Rediscovery

Kerr symmetries

Lupsasca: a known result recovered

Erdős episode, Oct 2025 (TechCrunch, 19 Oct 2025) · Bubeck et al., Nov 2025, arXiv:2511.16072, Sec. I.2 and III.2

Recombination

> Recombining what is already known goes further than expected.

Literature search

Erdős problems

existing solutions found in papers

Rediscovery

Kerr symmetries

Lupsasca: a known result recovered

New calculation

Cosmic-string loops

Scherrer: the odd-n case, checked against his even-n result

Erdős episode, Oct 2025 (TechCrunch, 19 Oct 2025) · Bubeck et al., Nov 2025, arXiv:2511.16072, Sec. I.2 and III.2

Recombination

> Recombining what is already known goes further than expected.

Literature search

Erdős problems

existing solutions found in papers

Rediscovery

Kerr symmetries

Lupsasca: a known result recovered

New calculation

Cosmic-string loops

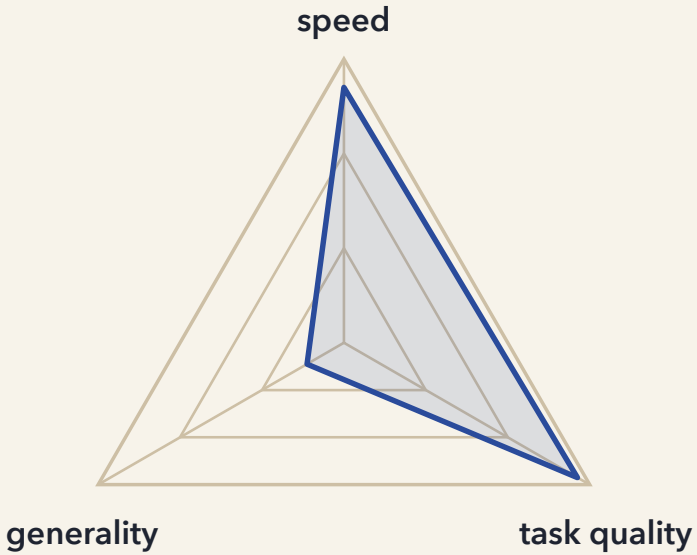
Scherrer: the odd-n case, checked against his even-n result

Reason 3: a model can also act on what it describes → Part 2

Erdős episode, Oct 2025 (TechCrunch, 19 Oct 2025) · Bubeck et al., Nov 2025, arXiv:2511.16072, Sec. I.2 and III.2

Three kinds of cook

Speed, breadth and task quality: different models, different trade-offs.

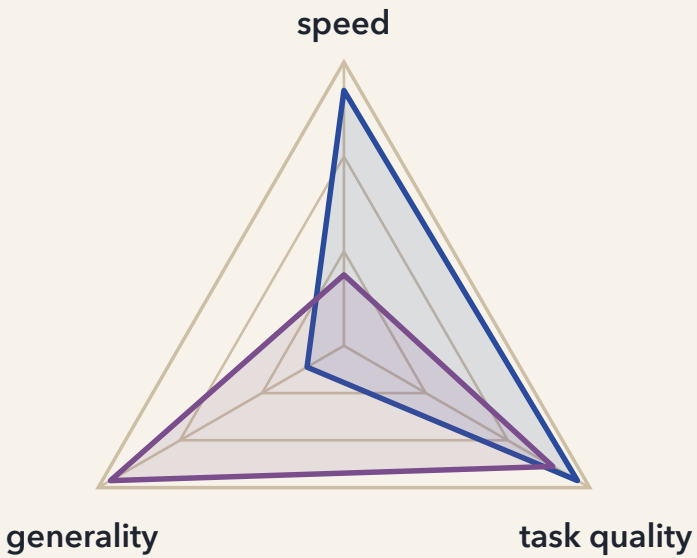


Task-specific model · the trained apprentice

Fast on a familiar menu. A new task may need retraining.

Three kinds of cook

Speed, breadth and task quality: different models, different trade-offs.



Task-specific model · the trained apprentice

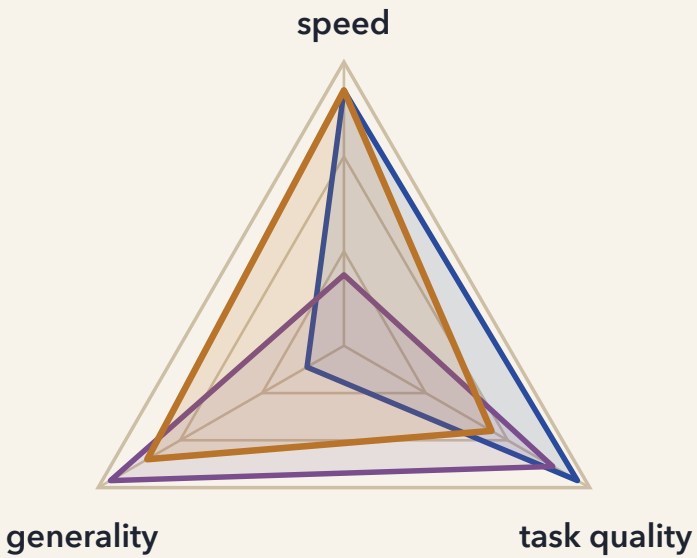
Fast on a familiar menu. A new task may need retraining.

LLM · the head chef

A wide menu. Quality, speed and cost vary by task.

Three kinds of cook

Speed, breadth and task quality: different models, different trade-offs.



Task-specific model · the trained apprentice

Fast on a familiar menu. A new task may need retraining.

LLM · the head chef

A wide menu. Quality, speed and cost vary by task.

Jev · the cook next door

New dishes without lessons, and fast.
For bounded judgments; check quality on the task.

One pass or many

A language model writes its answer; Jev picks one.

$17 \times 24 = 408$ step by step · **deliberate**

$2 + 2 = 4$ at once · **immediate**

Fast / slow thinking is an analogy; sequential generation is not itself reasoning.

One pass or many

A language model writes its answer; Jev picks one.

$17 \times 24 = 408$ step by step · **deliberate**

$2 + 2 = 4$ at once · **immediate**

LLM: generates token by token

"What colour is the sky?"

The	sky	is	blue
pass 1	pass 2	pass 3	pass 4

Fast / slow thinking is an analogy; sequential generation is not itself reasoning.

One pass or many

A language model writes its answer; Jev picks one.

$17 \times 24 = 408$ step by step · **deliberate**

$2 + 2 = 4$ at once · **immediate**

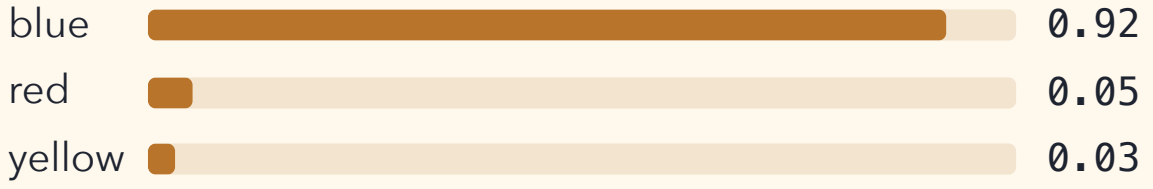
LLM: generates token by token

"What colour is the sky?"

The	sky	is	blue
pass 1	pass 2	pass 3	pass 4

Jev: one pass, every option

"What colour is the sky?" · blue / red / yellow



Fast / slow thinking is an analogy; sequential generation is not itself reasoning.

Jev, vendor-reported: 70–500 ms per call · input $0.042 / 1M tokens · output free · probabilities illustrative

Jev demo: triaging bug reports

> A typed answer, with a confidence that code can branch on.

STATE
one bug report, free text

QUESTION · SCORE
severity: 0 cosmetic → 4 critical

```
# code decides, not the model
if confidence < 0.5: a person
elif score < 1.5:    backlog
elif score < 2.5:    this sprint
else:                on-call
```

REPORT	SCORE	CONF.	OUTCOME
Typo on the settings page	0.00	1.00	backlog
PDF export fails over 50 pages; splitting works	2.00	1.00	this sprint
Checkout returns a 500 error for every customer	3.36	0.70	on-call
Notes gone after the update, or maybe deleted	3.19	0.34	a person

Recorded jev-1.14.0 answers from the TypeSafe documentation, "Routing by confidence" (accessed Nov 2026) · not a live call

Jev demo: triaging bug reports

A typed answer, with a confidence that code can branch on.

STATE
one bug report, free text

QUESTION · SCORE
severity: 0 cosmetic → 4 critical

```
# code decides, not the model
if confidence < 0.5: a person
elif score < 1.5:    backlog
elif score < 2.5:    this sprint
else:                on-call
```

"Notes gone after the update, or maybe deleted"



score **3.19**
alone, would wake on-call

confidence **0.34**
split answer → a person reads it

Recorded jev-1.14.0 answers from the TypeSafe documentation, "Routing by confidence" (accessed Nov 2026) · not a live call

Many small judgments

> Questions change daily; cheap, fast answers make small checks worth automating.

LLM · few, large jobs

a report · some code · an analysis

Jev · many small checks

every email · comment · issue · abstract

Many small judgments

Questions change daily; cheap, fast answers make small checks worth automating.

LLM · few, large jobs
a report · some code · an analysis

Jev · many small checks
every email · comment · issue · abstract

Classify 1 million abstracts, ~300 tokens each

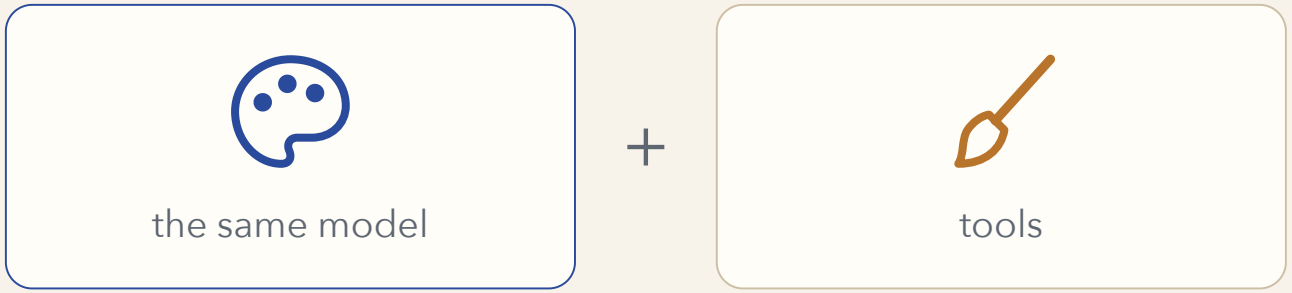
estimate · input only · accuracy not compared

Jev	Haiku 5.5	Opus 5.5
≈ $13	≈ $30	≈ $1,200

Output price per 1M tokens, one provider: Haiku 5.5 **$0.50** → Fable 5.2 **$40** (80×)

Painting and acting

> Harness changes alter the same model's behaviour.



A full description still needs a brush.

Agent = LLM + harness after LangChain, "Agent = Model + Harness" (Mar 2026)

Painting and acting

> Harness changes alter the same model's behaviour.



Critic: same model, new brief, fewer permissions

Agent = LLM + harness after LangChain, "Agent = Model + Harness" (Mar 2026)

Painting and acting

> Harness changes alter the same model's behaviour.

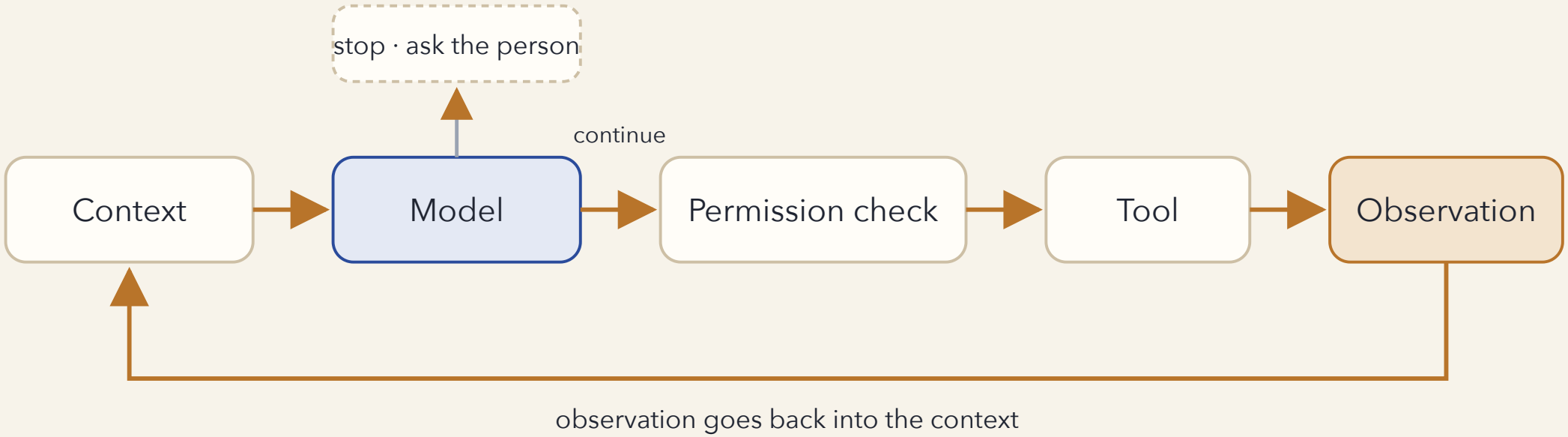


Assistant: weekly market review, memory, report files

Agent = LLM + harness after LangChain, "Agent = Model + Harness" (Mar 2026)

The tool-call loop

Tool results feed the next step.



Three harnesses

> Three working environments, three strengths.

Claude Code

subagents
session messaging

Three harnesses

> Three working environments, three strengths.

Claude Code

subagents
session messaging

Codex

long runs
scheduled jobs

Three harnesses

> Three working environments, three strengths.

Claude Code

subagents
session messaging

Codex

long runs
scheduled jobs

DSH

everything is a plugin

Three harnesses

> Three working environments, three strengths.

Claude Code

subagents
session messaging

Codex

long runs
scheduled jobs

DSH

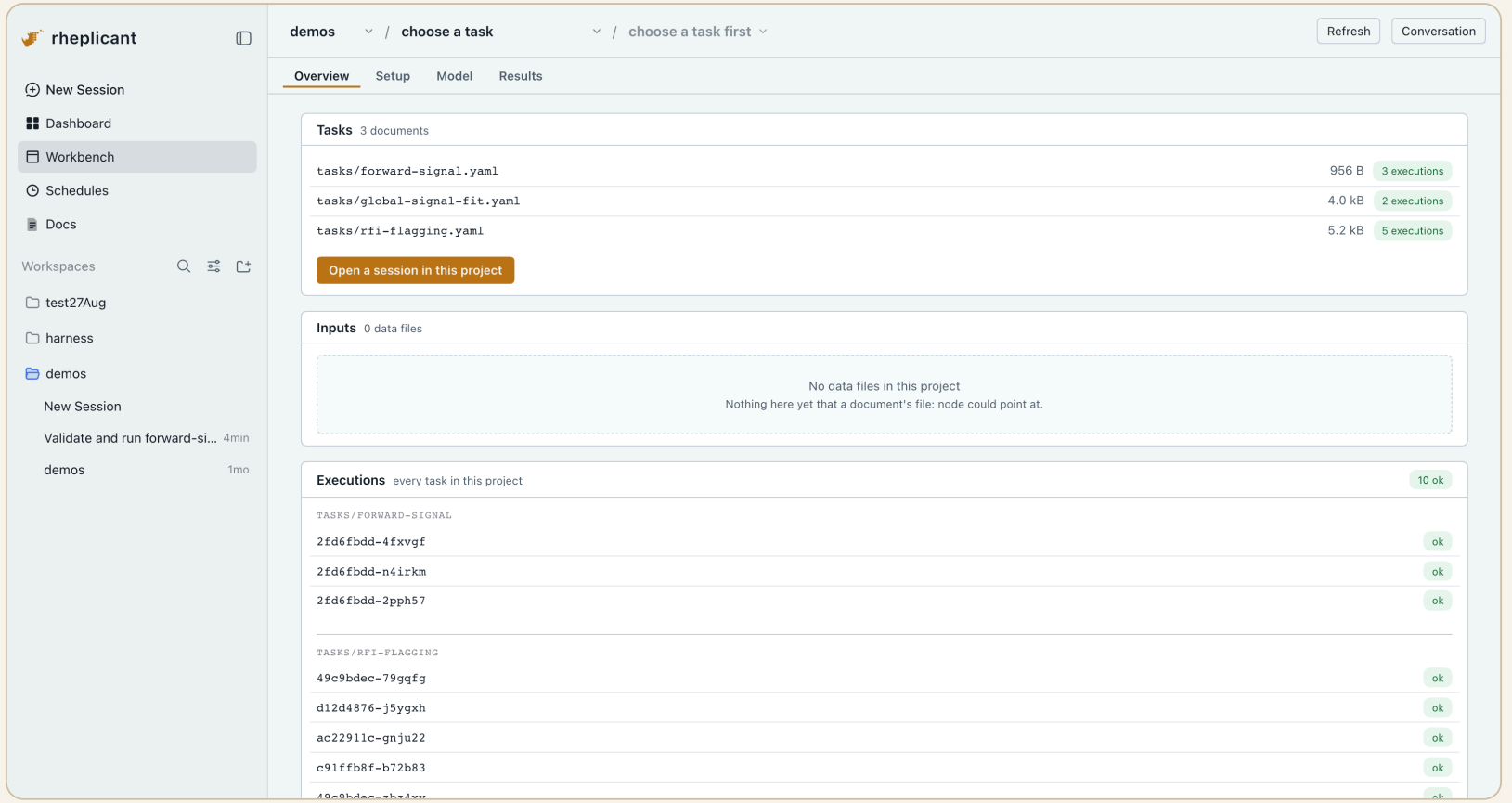
everything is a plugin

access

subscription · API key · self-hosted endpoint

rheplicant-agent: the console

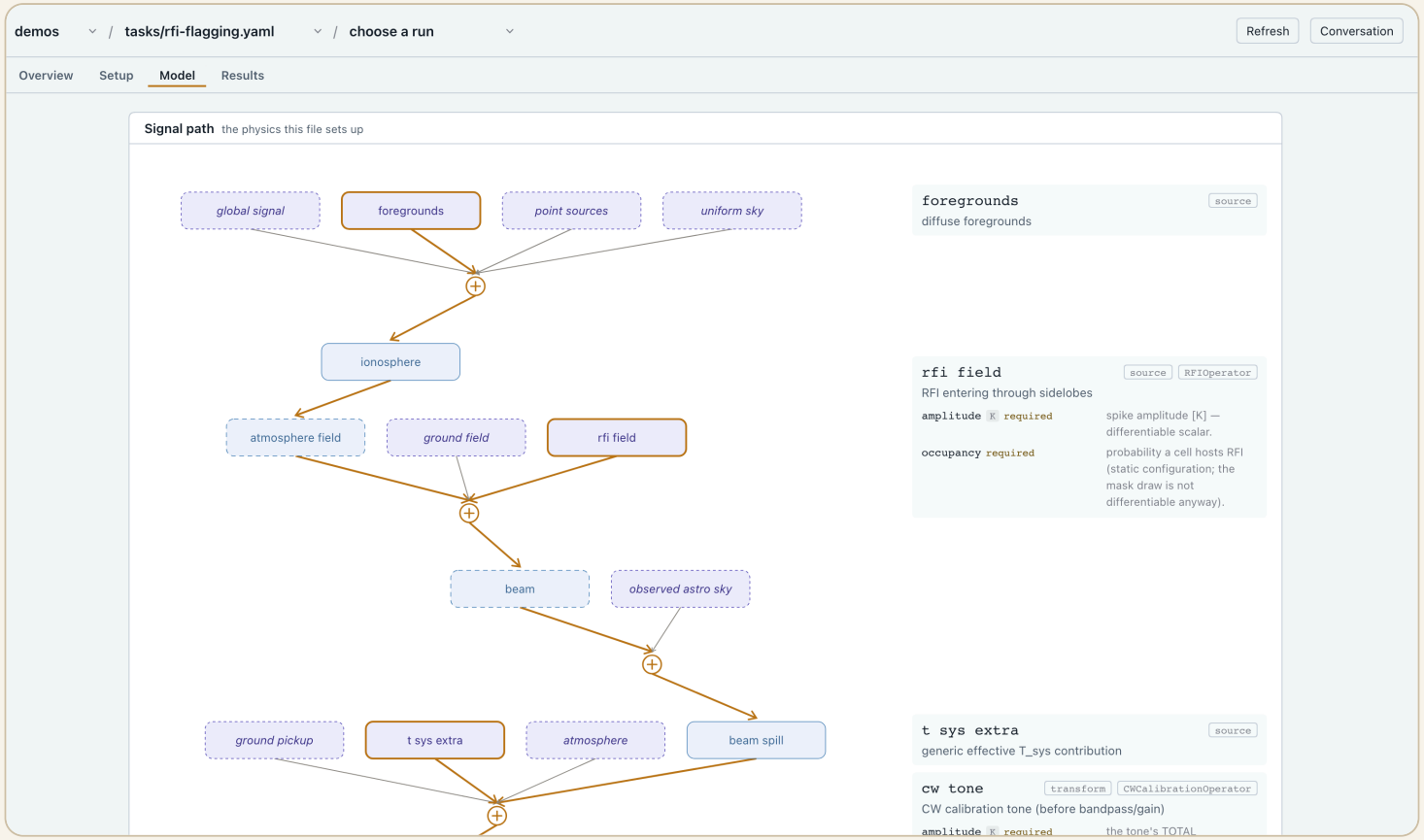
> Projects, task documents and every execution.



rheplicant-agent console, captured 13 Nov 2026 · demo task with placeholder physics; numbers are not predictions

A task document sets up the physics

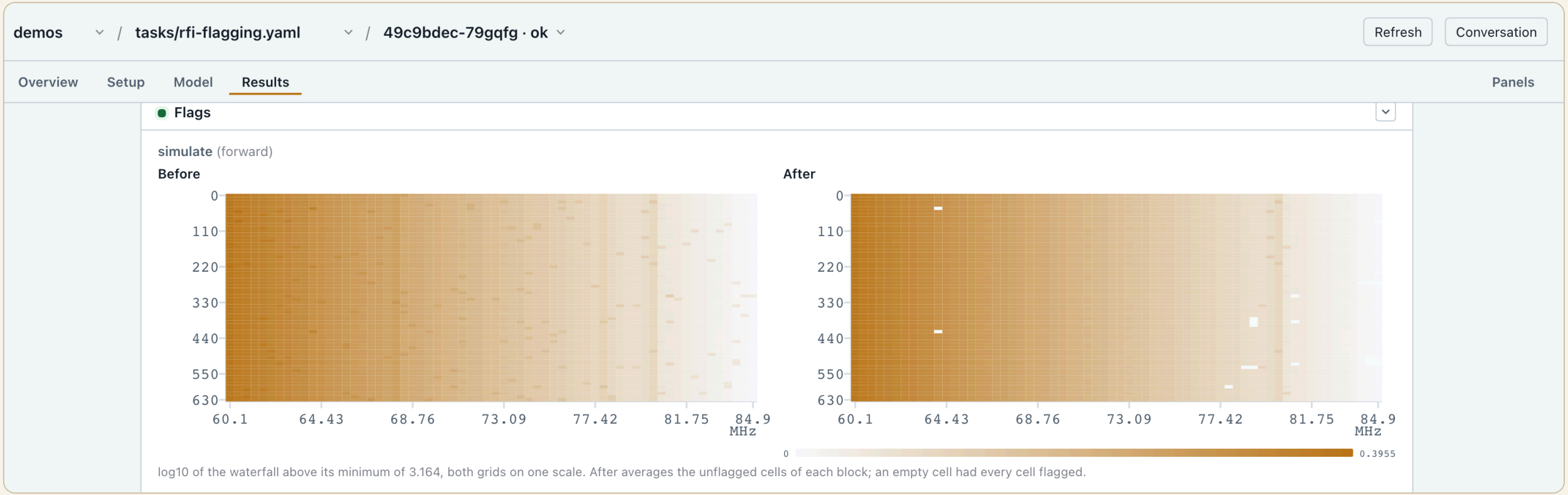
> The signal path, drawn from the task file.



rheplicant-agent console, captured 13 Nov 2026 · demo task with placeholder physics; numbers are not predictions

Results from the run

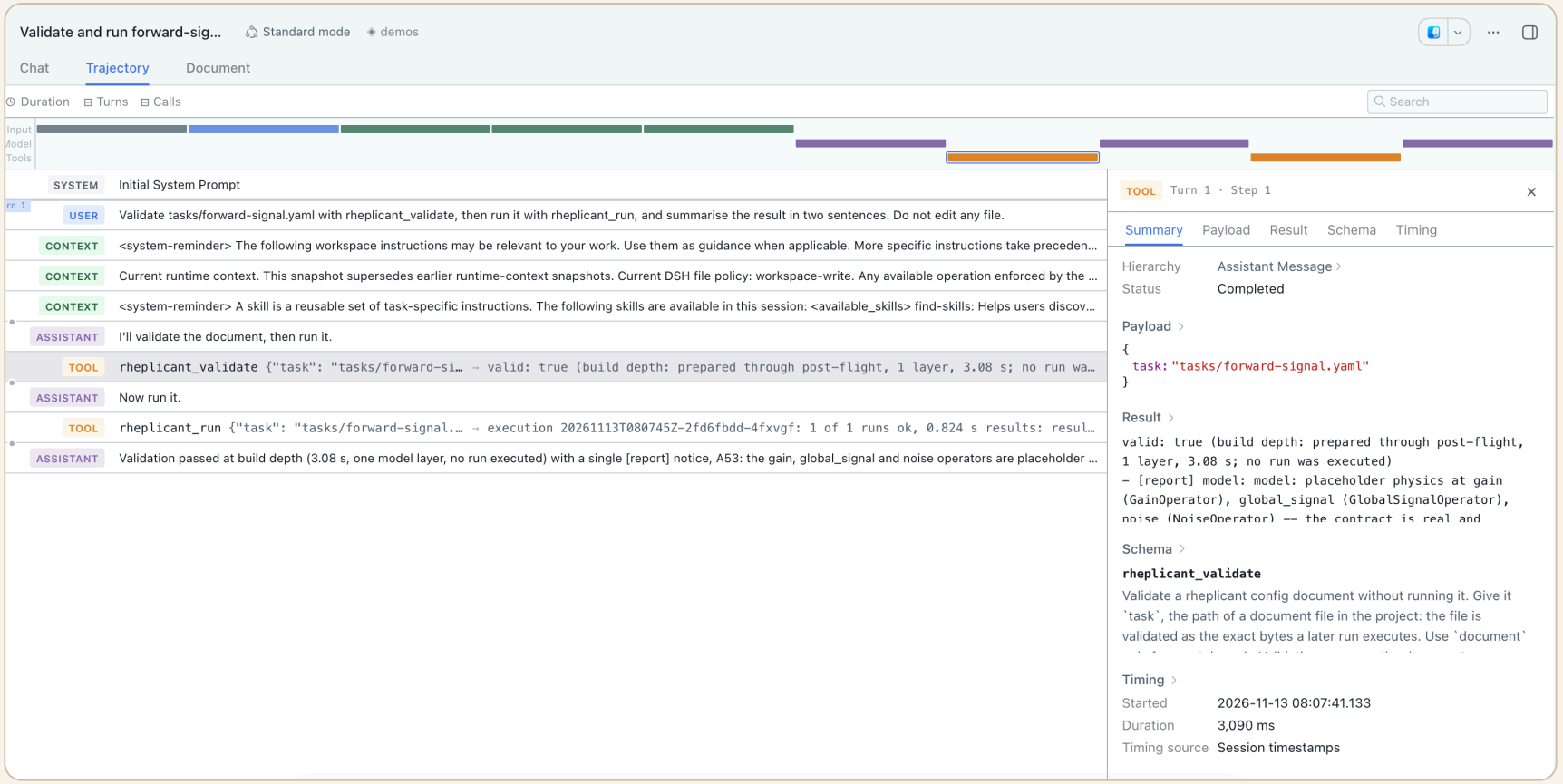
Waterfall before and after flagging, read from saved output.



rheplicant-agent console, captured 13 Nov 2026 · demo task with placeholder physics; numbers are not predictions

Trajectory: one recorded tool run

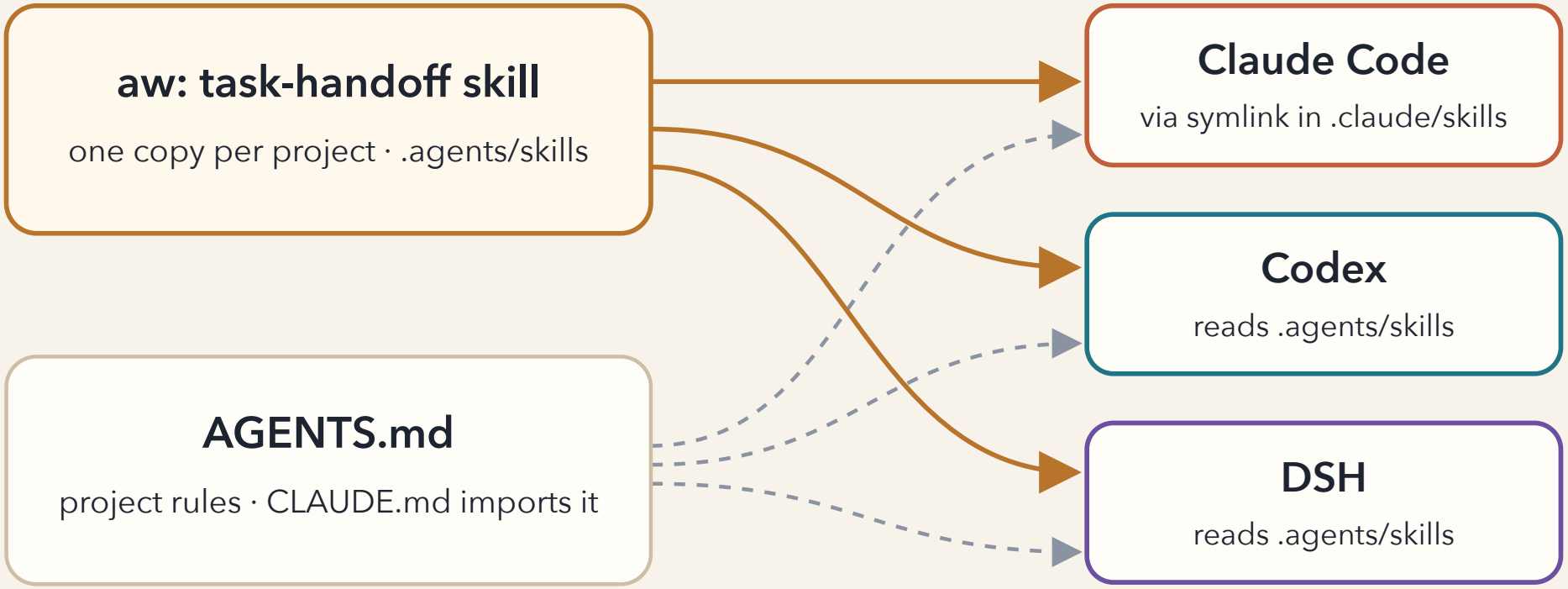
Prompt → validate → run → results · 9 seconds.



rheplicant-agent console, captured 13 Nov 2026 · demo task with placeholder physics; numbers are not predictions

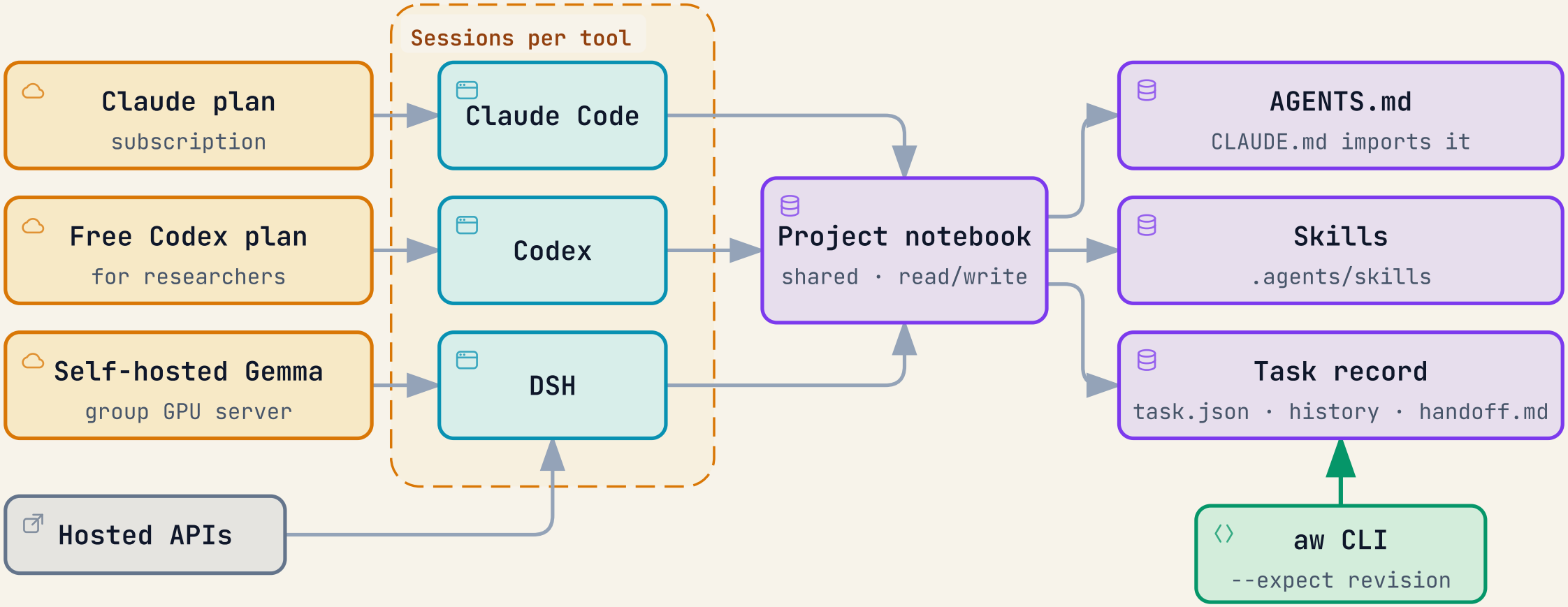
Shared instructions and skills

> The aw skill, task-handoff, travels with the task.



Provider independence

> One notebook connects providers.



This talk's task record

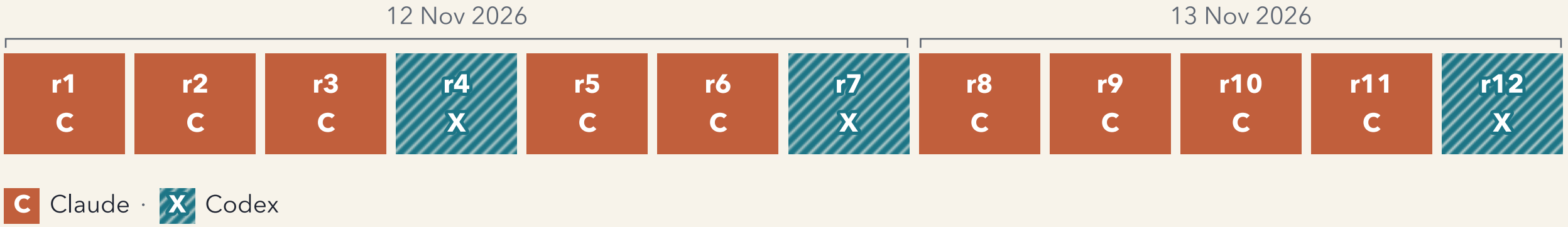
> Claude and Codex took turns preparing this talk.

```
$ aw tasks | cut -f1-3
T-001   ready   r12
```

This talk's task record

> Claude and Codex took turns preparing this talk.

```
$ aw tasks | cut -f1-3
T-001   ready   r12
```

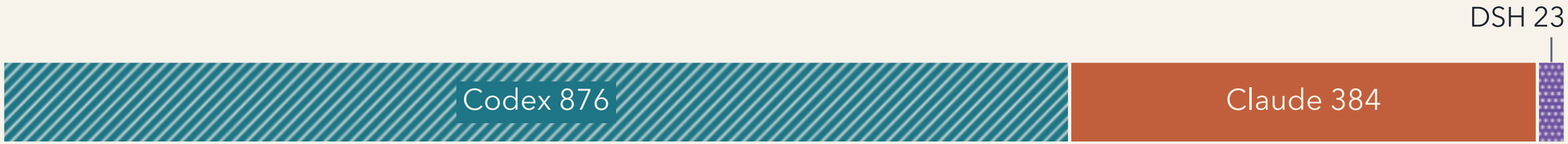


Recorded handoffs

1,283 handoffs in 91 tasks, across three agents.

24 of 91

tasks used two or more agents



aw task records, 12 Sep – 12 Nov 2026 (snapshot) · labels name the harness, not the model behind it

Completion and acceptance

> Done is a claim.

sft-wick T-001 · a public package · one block per revision



Summary

Task-sized models, one shared notebook.

Theme 1

Choosing the right model

task-sized models

Theme 2

Agent = LLM + harness

task record kept outside any one provider


model


harness


notebook

BACKUP

For questions

System 1 terminology

Access and payment paths

aw limits and adoption

DSH extension layers

Routing and self-hosting

Sessions and access

Shared interfaces

A researcher-use survey

Weekly handoffs

Emulator runtime

Domain harness architecture

Building a harness

Self-hosting and memory

Harness references

GPU energy estimate

Physical-model acceptance

System 1 terminology

> Terms depend on context.

TypeSafe	"System One": typed decisions (Jev)
LeCun	autoregressive language modelling counts as System 1
Labs	non-thinking vs thinking mode

Shared interfaces

> These interfaces do different jobs.

AGENTS.md
project instructions

Agent Skills
task procedures (SKILL.md)

MCP
tool connections

Harness support varies.

BACKUP ↩

Building a harness

> A harness needs execution and recovery.

Model side

loop · context · model adapter

Execution side

tools · permissions · sandbox

State side

session log · failures · recovery · checks

BACKUP ↩

Access and payment paths

> Access and task records are separate.

Subscription login

check each vendor's terms

API credentials

pay per token

Self-hosted endpoint

OpenAI-compatible server

Check support and terms for each tool.

A researcher-use survey

> Survey results describe their respondents.

10,480

respondents · Springer Nature · Jul 2026

≈ 50%

frequent use for information gathering and editing

Self-reported use

Self-hosting and memory

> Context is shared across slots.

Model	Gemma 4 31B · QAT Q4_0
Runtime	llama.cpp · one A100 40 GB
Capacity	8 slots · 224K context, one shared pool
Rule of thumb	weights ≈ parameters × bits ÷ 8, plus cache and runtime

aw limits and adoption

> A task record reports work.

Adoption (snapshot)

76 registered projects · 21 with tasks · 55 without tasks

Native-session bindings used: 0

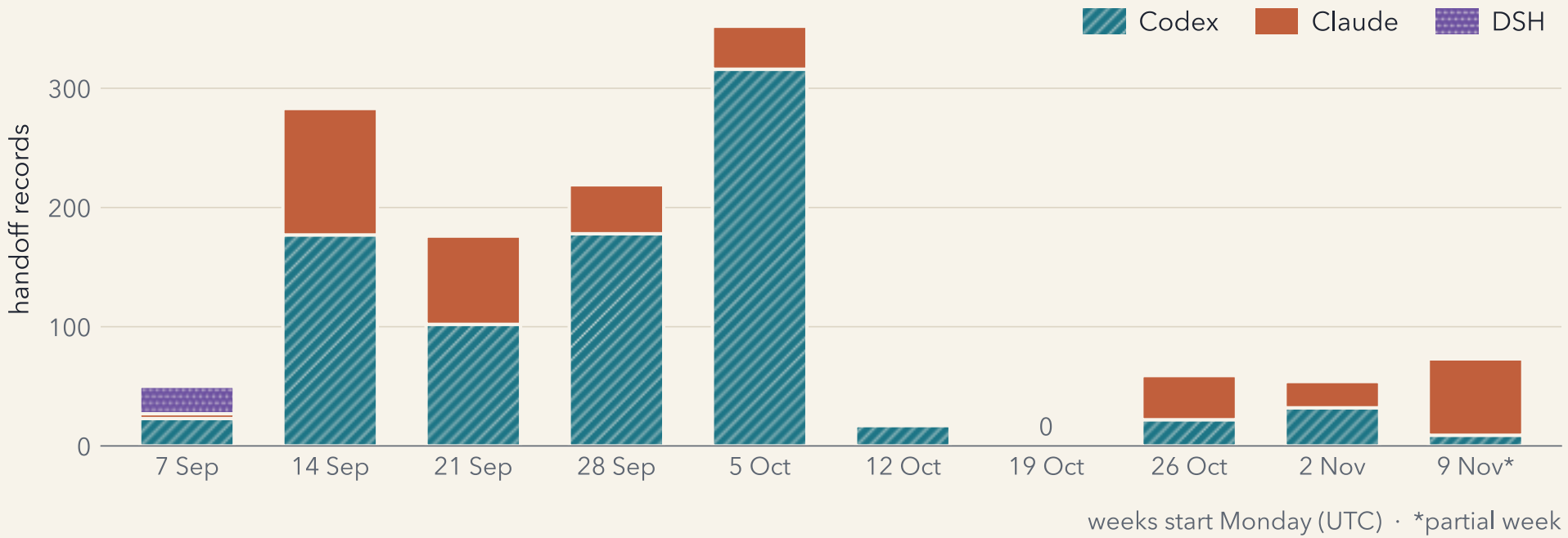
Limits

no automatic verification · local, not encrypted or redacted · revision checks do not lock code

BACKUP ↩

Weekly handoffs

> Activity came in bursts.



Harness references

> Several sources shaped this framing.

Weng, 2023	planning, memory, tools around an LLM
Anthropic, 2025	harnesses for long-running work
Hashimoto, 2026	"harness engineering"
LangChain, 2026	Agent = Model + Harness

Weng, "LLM Powered Autonomous Agents", 23 Jun 2023 · Anthropic, "Effective harnesses for long-running agents", 26 Nov 2025 · Hashimoto, "My AI Adoption Journey", 5 Feb 2026 · Trivedy, "The Anatomy of an Agent Harness", LangChain, 10 Mar 2026

DSH extension layers

> Plugins compose the working environment.

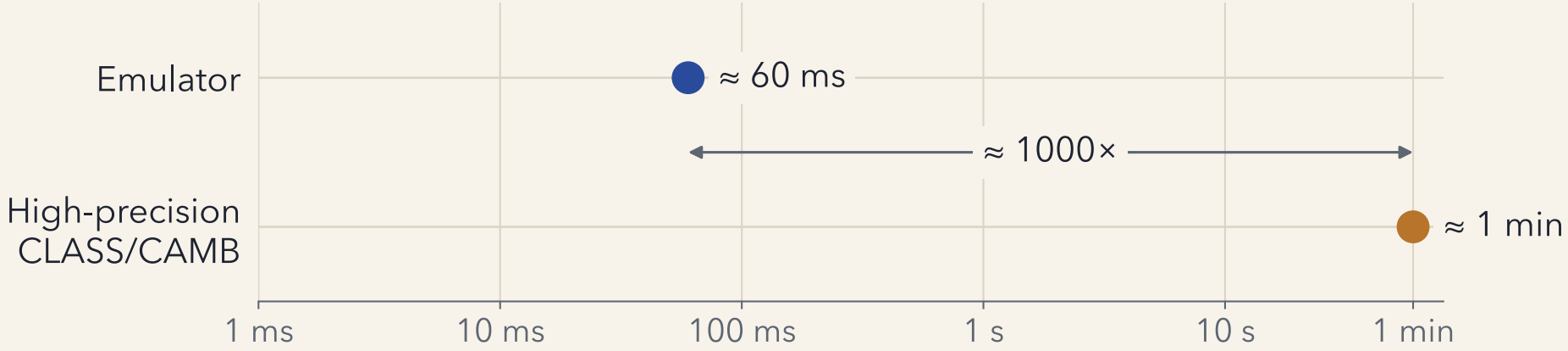
Profile	bundles → patch layers → user overrides
Plugins	tools · prompt sections · UI · model adapters
Domain service	rheplicant compute, over JSON-RPC

The assembled configuration still needs checking.

Emulator runtime

> A task-specific model trades generality for speed.

Runtime per call: one set of CMB and matter spectra



Bolliet et al. 2023, CosmoPower (arXiv:2303.01591)

GPU energy estimate

> About a quarter watt-hour per 500 tokens, GPU only.

Measured · Sep 2026 deployment test

8 concurrent image+text requests

GPU ≈ 234 W

≈ 130 tokens/s in total

Estimate · GPU only

≈ 1.8 J per token

500 tokens ≈ **0.25 Wh**

≈ a 10 W LED for 90 seconds

Routing and a self-hosted model

> Name model tiers, not versions; some jobs run on the group GPU.

Rule for subagents

least expensive model that reliably does the bounded task

It named versions; they are now "older generation".

DSH

DeepSeek Harness

→ OpenAI-compatible endpoint

Group GPU server

Gemma 4 31B · llama.cpp
one A100 40 GB

Group deployment record, 11 Sep 2026 · energy per token is an estimate from that test

Routing and a self-hosted model

> Name model tiers, not versions; some jobs run on the group GPU.

Rule for subagents

least expensive model that reliably does the bounded task

It named versions; they are now "older generation".

DSH
DeepSeek Harness

→ OpenAI-compatible endpoint

Group GPU server

Gemma 4 31B · llama.cpp
one A100 40 GB

Deployment test, 8 concurrent requests · GPU only · energy estimated

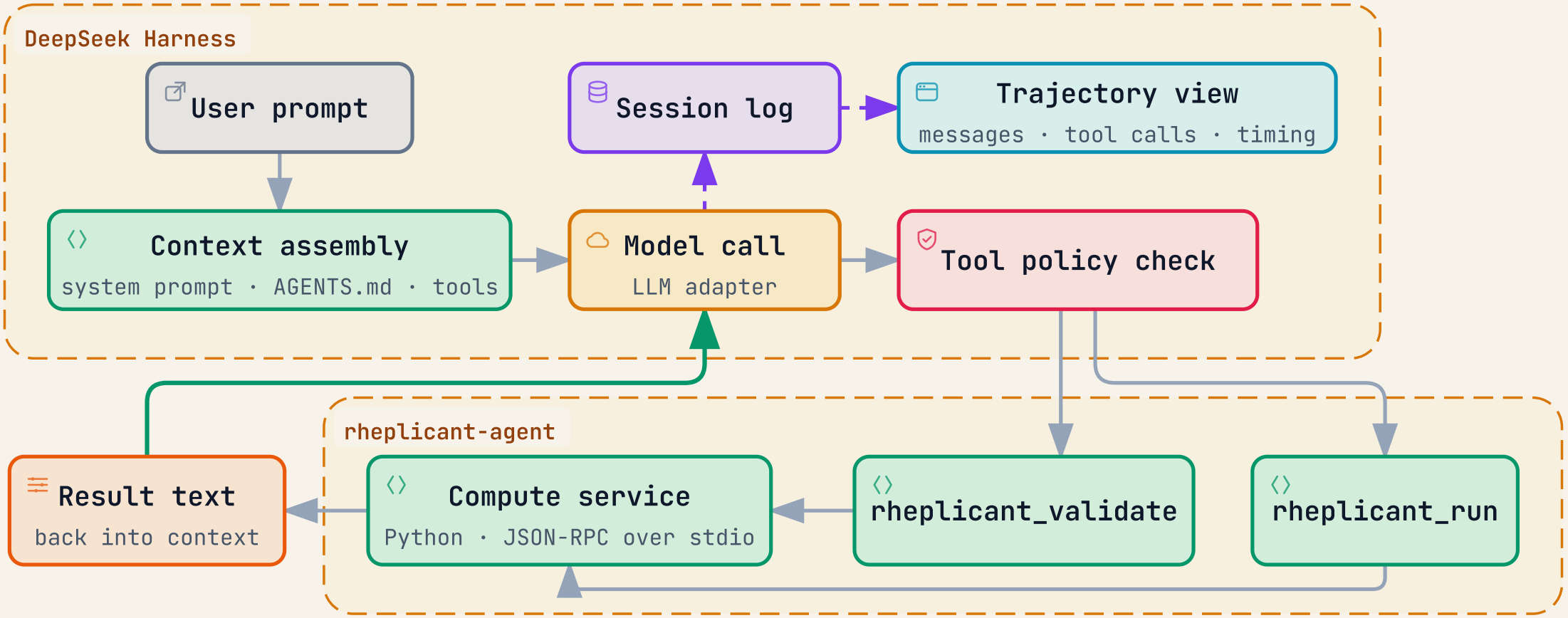
≈ 234 W at ≈ 130 tokens/s in total → 500 tokens ≈ **0.25 Wh**

a 10 W LED for 90 s

Group deployment record, 11 Sep 2026 · energy per token is an estimate from that test

One turn in rheplicant-agent

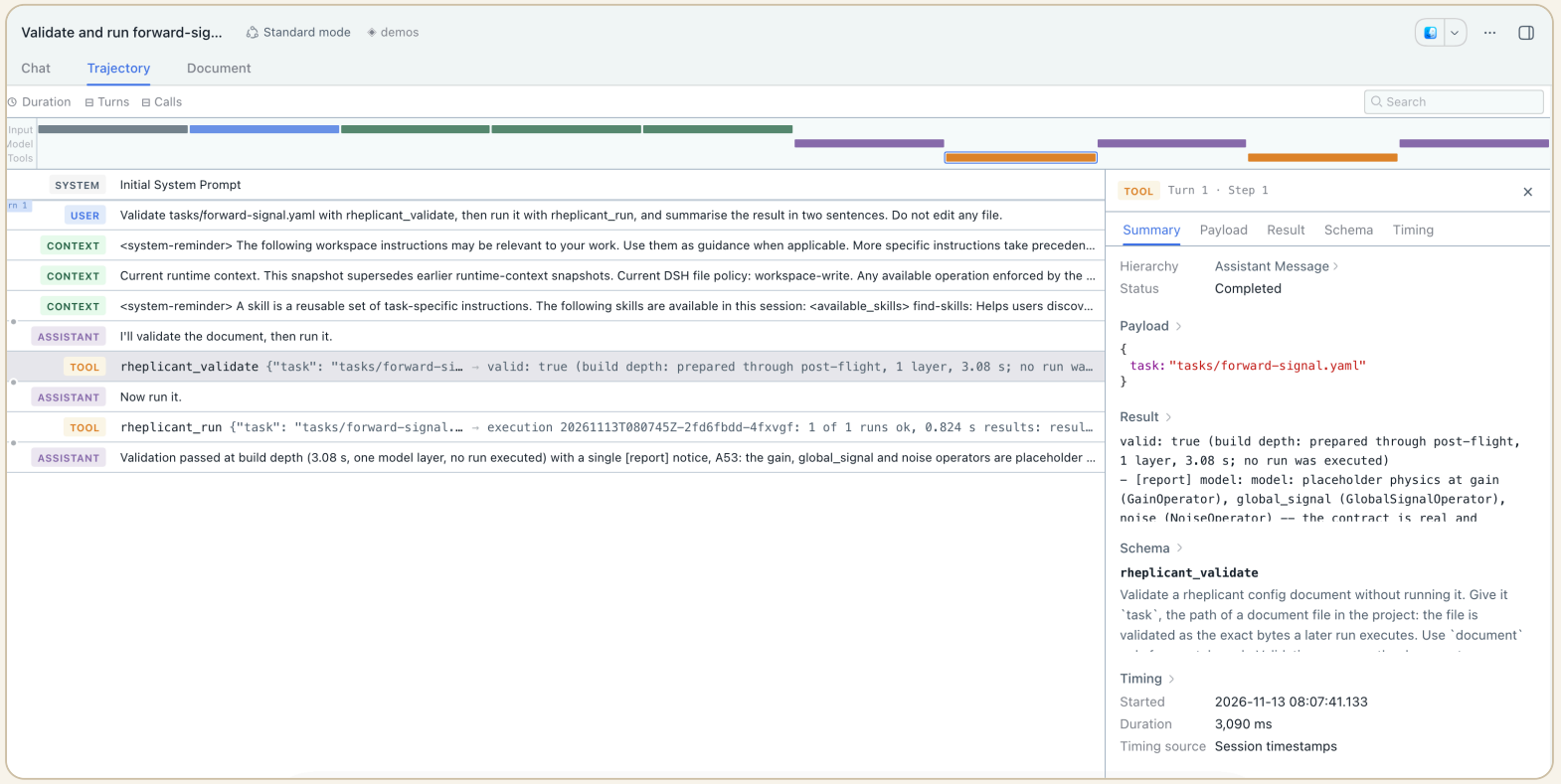
> A domain harness connects the model to scientific tools.



Recorded run, 13 Nov 2026 (DSH Trajectory view) · rheplicant-agent: 26 package rows on DSH 0.1.5-rc.2; no fork, four hash-checked adaptor patches

One turn in rheplicant-agent

> A domain harness connects the model to scientific tools.



Recorded run, 13 Nov 2026 (DSH Trajectory view) · rheplicant-agent: 26 package rows on DSH 0.1.5-rc.2; no fork, four hash-checked adaptor patches

BACKUP ↩

One turn in rheplicant-agent

> A domain harness connects the model to scientific tools.

26
packages on DSH

0
forks of DSH

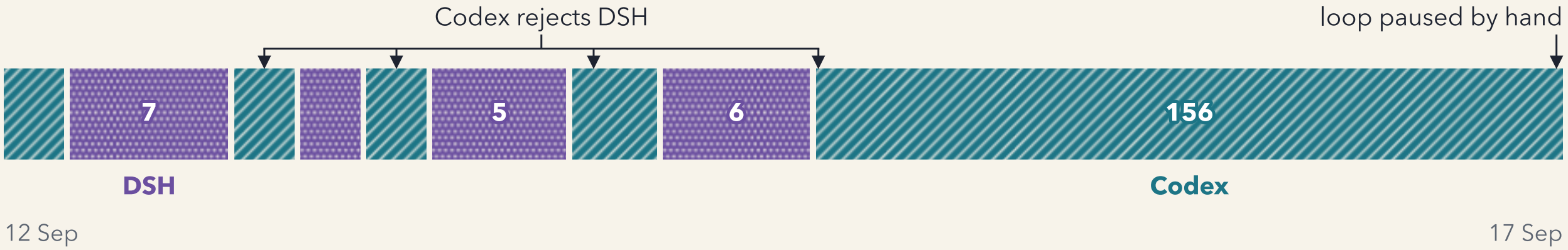
YAML
model route, tools, prompts,
UI: config rows

rheplicant-agent: five model-facing tools in front of rheplicant, a radio-telescope digital twin

Recorded run, 13 Nov 2026 (DSH Trajectory view) · rheplicant-agent: 26 package rows on DSH 0.1.5-rc.2; no fork, four hash-checked adaptor patches

Physical-model acceptance

> The physical model still needed questioning.



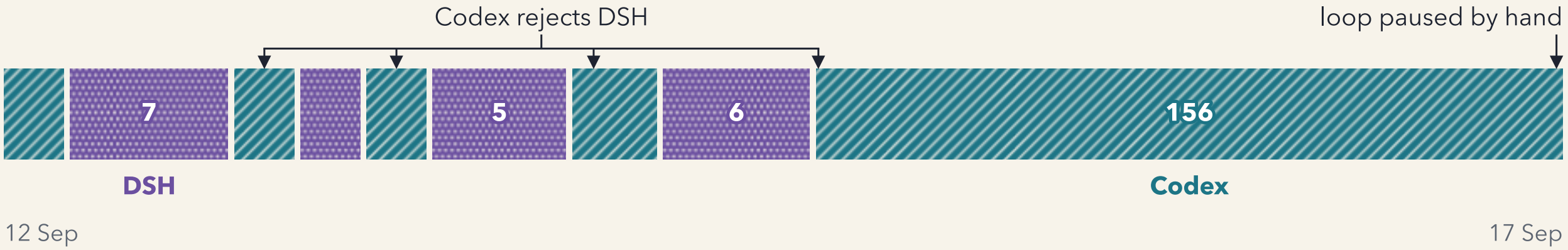
Codex rejects **DSH** ×4

implementation errors found and fixed

Numbers in the strip count revisions; block width grows with the square root of the count

Physical-model acceptance

> The physical model still needed questioning.



Codex rejects **DSH** ×4

implementation errors found and fixed

Continue → **loop paused by hand**

and question the physical model

Numbers in the strip count revisions; block width grows with the square root of the count

BACKUP ↩

Sessions and access

> Sessions stay native; the task record moves.

aw moves no sessions

386

Claude Code sessions Codex imported by itself

Sessions and access

> Sessions stay native; the task record moves.

aw moves no sessions

386

Claude Code sessions Codex imported by itself

Access is separate

Anthropic consumer logins

Free · Pro · Max

only in claude.ai and Claude Code

Anthropic consumer-terms clarification, Feb 2026 (The Register, 20 Feb 2026; Claude Agent SDK documentation)