

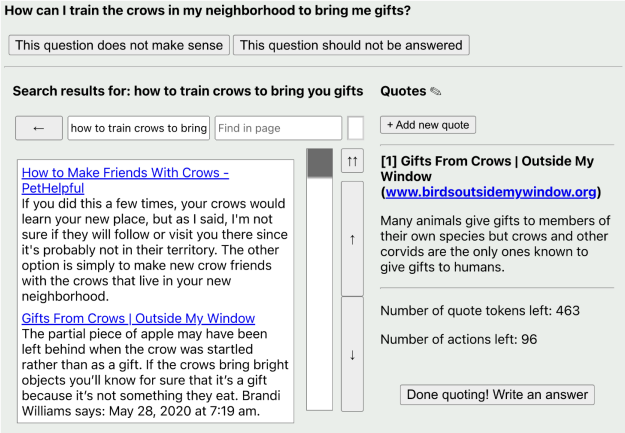
FIREFLY: ILLUMINATING LARGE-SCALE VERIFIED TOOL-CALL DATA GENERATION FROM REAL APIS

Yuxuan Lu, Ziyi Wang, Yingzhou Lu, Yisi Sang, Jiri Gesi,
Xianfeng Tang, Yimeng Zhang, Zhenwei Dai, Hui Liu, Hanqing Lu,
Chen Luo, Qi He, Benoit Dumoulin, Jing Huang, Dakuo Wang

Northeastern University | https://yuxuan.lu

Oct 2026

Agents Call Tools: WebGPT (2021)

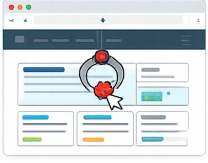


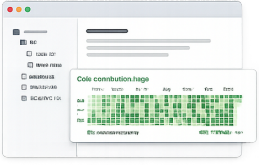
The browser is the tool: the model calls `search`, `click`, `quote`.

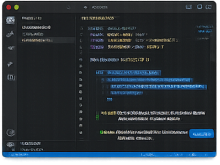
OpenClaw

opencode

Claude Code







● The opening is now in your order, the "bottleneck moved" sentence is gone, and forward generation is explained before it gets criticized. The talk compiles with no errors to 41 PDF pages.

New motivation order:

1. **Agents Call Tools: WebGPT (2021).** The browser is the tool: the model calls search, click and quote.
2. The agents overview slide, then the opencode screenshot.
3. **Where Tool Calling Comes From.** Two cards. "Emerged": prompt a large model with tool descriptions (ReAct). "Trained": fine-tune and run RL on tool-use tasks (Toolformer onward). Bottom line: "The training recipe behind the recent jump: RL with verifiable rewards (RLVR)."
4. **RLVR examples:** DeepSeekMath, o1, R1 and Reasoning Emerges. The titles now start with "RLVR:".
5. **What They Share: A Free Verifier.** It now ends with "Math has an answer key. Code has unit tests. Tool calls have neither."
6. **What One Training Example Must Contain**, using the whois example.
7. **Human-Crafted Verified Data Does Not Scale.** The left side lists Tau2-Bench (279 tasks) and API-Bank. The right side has the big numbers on cost, shown below. Then: "Big labs buy it, often exclusively at 4–5× the price; academia cannot."
8. **Automated Pipelines: Forward Generation.** This is the new slide you asked for. Three steps: an LLM writes the task, then writes the tool calls and imagines their responses, then the answer is read off that trajectory. Bottom lines: "ToolAlpaca, APIGen-MT, Toucan: cheap and scalable, up to 1.5M tasks." and "But nothing checks the label against a real API."
9. **A Wrong Answer Key Is Worse Than None.** This now reads as the failure case of the method just explained.
10. The existing datasets table, then the Pick Two image with "not scalable / not real / not verifiable".
11. One standout: "How can we get tool-call training data that is scalable, verifiable, and cheap, on real APIs?" I deleted the slide that repeated "real".
12. Key Idea: Invert the Pipeline, then continue as before.

About the cost evidence. The source is the Epoch AI interview piece on RL environments, published January 2026. It does not support "several thousand dollars per task" as typical, so I didn't use that number.

Claim on the slide	What the source says
$200–$2,000 per task	"I've seen $200 to $2000 mostly. $20k per task would be rare but possible."
$20K per website replica	Website replicas cost about $20k each
$1B+	Anthropic discussed spending over $1 billion on RL environments in a year
4–5× for exclusivity	Exclusive deals cost roughly 4–5× more than non-exclusive ones

The source does not say anything about academic access. "Academia cannot" is our inference from the price and the exclusive deals. Mechanize argues labs will soon spend a few thousand dollars per task. That is a forecast, not an observed price, so I left it out.

Sources:
- An FAQ on Reinforcement Learning Environments, Epoch AI
- Cheap RL tasks will waste compute, Mechanize

✳ Baked for 4m 23s · done 0:00

18 files changed +852 -270

files/talks/3_papers/figures/firefly/challenge-drifting.png
files/talks/3_papers/figures/firefly/challenge-too-many-tools.png
files/talks/3_papers/figures/firefly/firefly_icon.png
files/talks/3_papers/figures/firefly/firefly.pdf
files/talks/3_papers/figures/firefly/generation.pdf
files/talks/3_papers/figures/firefly/pass_at_k.pdf
files/talks/3_papers/figures/firefly/pick-two.png
files/talks/3_papers/figures/papers.png
↓ 10 more below (opt+↓ to scroll)

files/talks/3_papers/figures/firefly/challenge-drifting.png (untracked)

New file not yet staged.
Run `git add :/files/talks/3_papers/figures/firefly/challenge-drifting.png` to see line counts.

files/talks/3_papers/figures/firefly/challenge-too-many-tools.png (untracked)

New file not yet staged.
Run `git add :/files/talks/3_papers/figures/firefly/challenge-too-many-tools.png` to see line counts.

files/talks/3_papers/figures/firefly/firefly_icon.png (untracked)

New file not yet staged.
Run `git add :/files/talks/3_papers/figures/firefly/firefly_icon.png` to see line counts.

files/talks/3_papers/figures/firefly/firefly.pdf (untracked)

New file not yet staged.
Run `git add :/files/talks/3_papers/figures/firefly/firefly.pdf` to see line counts.

files/talks/3_papers/figures/firefly/generation.pdf (untracked)

New file not yet staged.

Update available! Run: brew upgrade claude-code@latest

❯

Where Tool Calling Comes From

Northeastern University

Emerged

Models trained only on language tasks can already call tools when prompted (ReAct).

A sign of big potential.

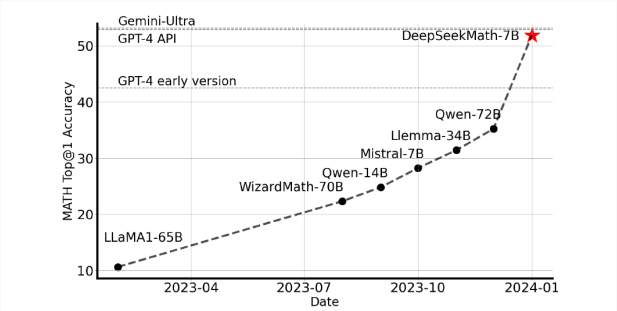


Trained

So tool use is now trained directly, with fine-tuning and RL (Toolformer).

The training recipe behind the recent jump: RL with verifiable rewards (RLVR).

RLVR: DeepSeekMath (2024)

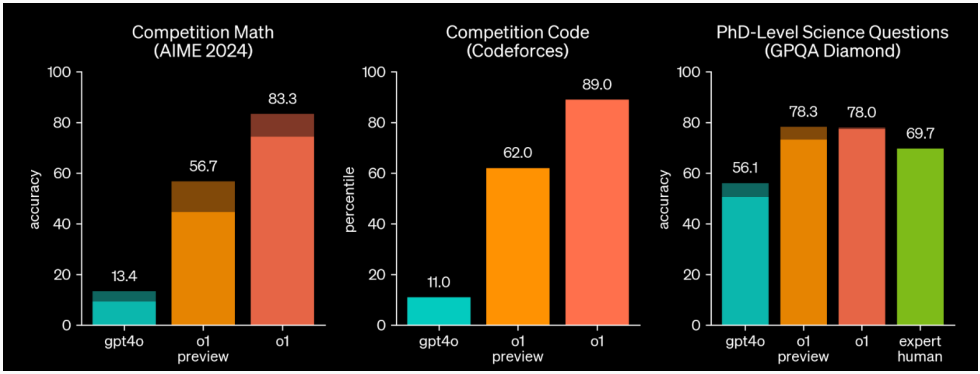


51.7%

on competition-level MATH with a 7B model, approaching Gemini-Ultra and GPT-4

Introduced **GRPO**: RL where the reward is simply whether the final answer matches.

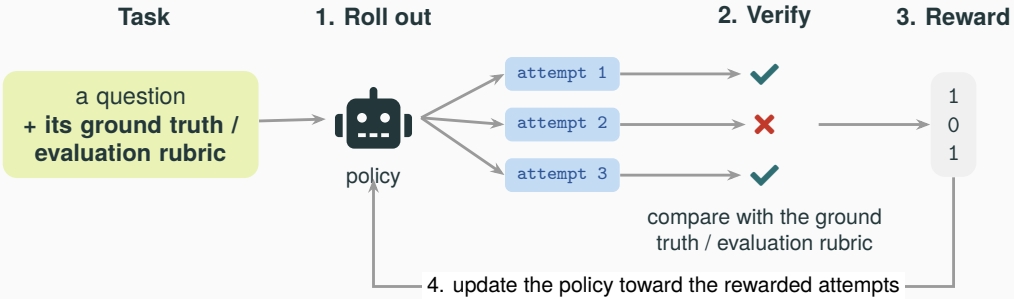
RLVR: OpenAI o1 (2024)



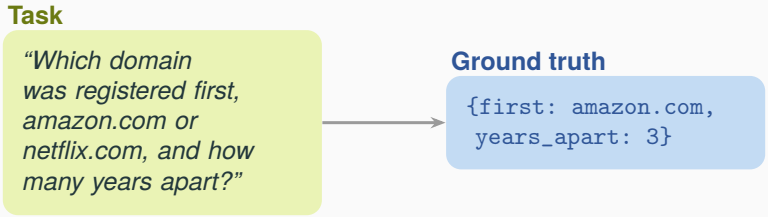
Large-scale RL teaches the model to think before it answers.

AIME 13.4 $\rightarrow$ 83.3, Codeforces percentile 11 $\rightarrow$ 89 over GPT-4o.

What RLVR Needs



RLVR for Tool Calling Needs



✔ a natural task ✔ an answer that is actually correct

For thousands of tasks, across hundreds of APIs.

Human-Crafted Verified Data Does Not Scale

Northeastern University

Open, human-curated tool-use data

Tau2-Bench: 279 tasks over 68 APIs, each with a hand-checked goal state.

API-Bank: 1,888 annotated tool-use dialogues.

Carefully verified, but small, and built one task at a time.

$200–$2,000
typical price of one RL task bought by AI labs

$20K
for one replicated website to train in

$1B+
Anthropic discussed spending on RL environments in a year

Big labs buy it, often exclusively at 4–5× the price; academia cannot.
Verified, but every new task and API costs more money.

Automated Pipelines: Forward Generation

Northeastern University

1. Write the task

Input: all tool descriptions.

An LLM writes a task that uses these tools.

2. Solver rollout

A separate solver model rolls out a trajectory of tool calls for that task.

3. Ground truth

The solver's final answer becomes the ground truth.

Automated Pipelines: Forward Generation

Northeastern University

1. Write the task

Input: all tool descriptions.

An LLM writes a task that uses these tools.

2. Solver rollout

A separate solver model rolls out a trajectory of tool calls for that task.

3. Ground truth

The solver's final answer becomes the ground truth.

✖ The task may be impossible: its final state is unreachable.

✖ The solver may get it wrong: then the ground truth is wrong.

ToolAlpaca, APIGen-MT, Toucan: scalable, up to 1.5M tasks.

Nothing guarantees the ground truth is correct.

Scalable, Real, Verifiable: Pick Two

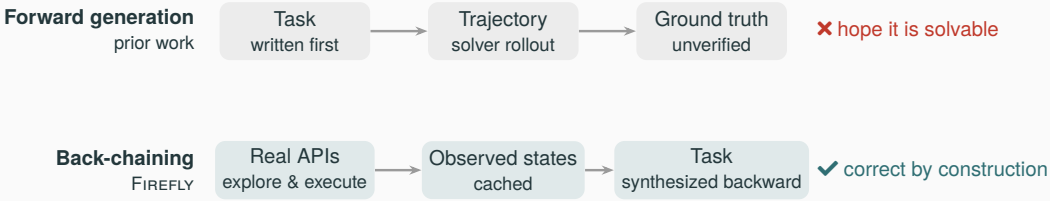
Northeastern University



✗ not scalable ✗ not real ✗ not verifiable

How can we get tool-call training data that is scalable, real, and verifiable?

Key Idea: Invert the Pipeline



Explore first, record what happened, then write the task whose answer you hold.

Method

Northeastern University

Dataset	Tasks	Servers/APIs	Multi-step	Verifiable	Real APIs
API-Bank	1,888	2,138	✓	✓	×
ToolAlpaca	3,938	400+	×	×	✓
APIGen-MT	5,000	27	✓	✓	×
Tau2-Bench	279	68	✓	✓	×
Toucan	1.5M	2000+	✓	×	✓
FIREFLY	**5,144**	**993**	✓	✓	✓

The only dataset that is multi-step, verifiable, and grounded in real APIs.

The Firefly Dataset

Northeastern University

5,144
verified tasks

240 servers **993** tools

3.0 tool calls per task on average (1–10)

38.5% of tasks span *multiple* servers

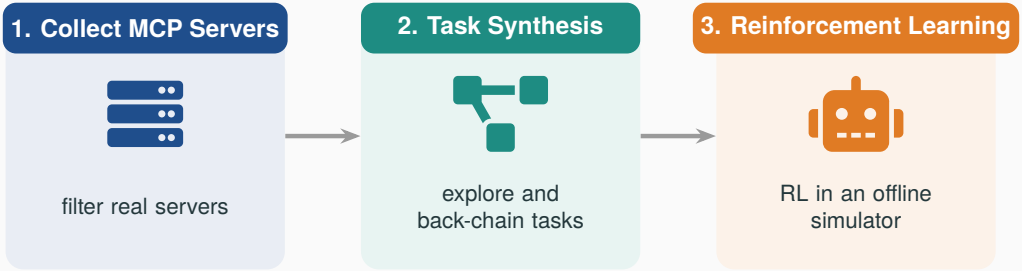
4.6 answer fields per task

7.8 / 10 realism, LLM-judged

Each task ships with its trajectory, answer schema, and all replay responses.

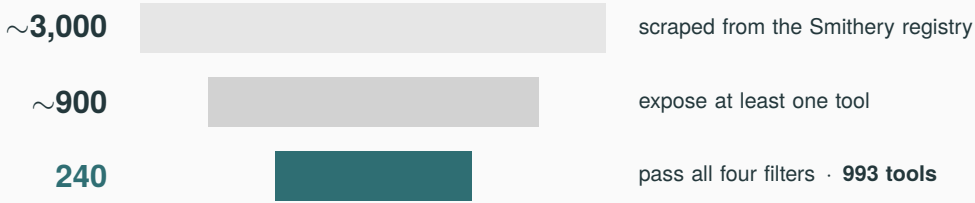
Firefly Pipeline

Northeastern University



Server Collection

Northeastern University



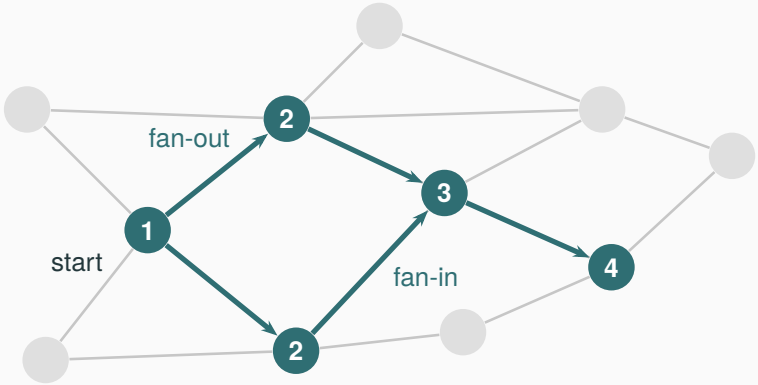
An LLM screens every server; a server must satisfy all four, spot-checked by hand.



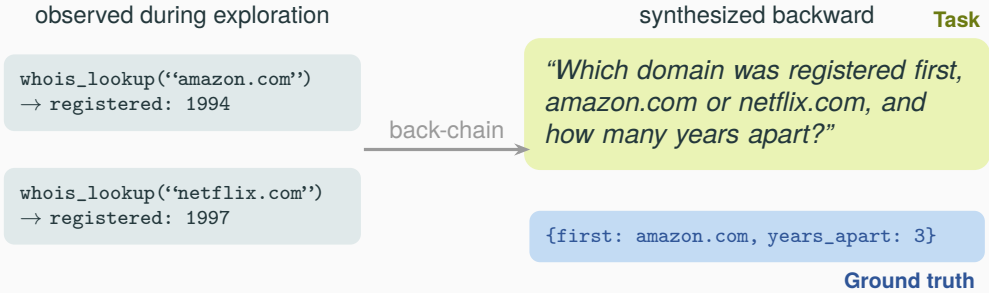
993 tools **83K** edges: “can *A*’s output feed *B*?” **88** successors per tool

Graph-Guided Exploration

Northeastern University



Back-Chaining: From Observed States to a Task



The task embeds the chosen nodes’ inputs; the answer is read off their outputs. Unused nodes become distractors. Several tasks per DAG from different subsets.

Task Validation

Northeastern University



Verifiable
answer is objectively checkable



Well-specified
exactly one correct answer



Interpretable
no opaque IDs



Realistic
realism $\geq$ 5 of 10



Calibrated
difficulty matches trajectory

9,275 candidate tasks $\rightarrow$ **5,144** pass all five, judged by an LLM

Why Not Train on Live Servers?

Northeastern University



Servers go down

`search("hotels")`

→ `503 Unavailable`



Not reproducible

`get_weather("Boston")`

Mon → `18°C, sunny`
Tue → `9°C, rain`



Random by design

`fortune_telling("Leo")`

1st → `"great luck"`
2nd → `"beware Tuesday"`

The same call must return the same response, or the reward means nothing.

Replay Simulator: How It Works

Northeastern University

Cached from exploration: `search(query="hotels", num=5, format="csv")`

exact match

`query="hotels"` ✓
`num=5` ✓
`format="csv"` ✓

return the **real** cached response

fuzzy match

`query="hotels"` ✓
`num=5` ✓
`format="json"` ✗

an LLM **simulates** it: the cached result as JSON

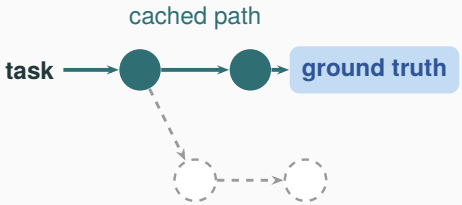
no data

tool never called during exploration

return an **error**

No live server is touched: the same call always gets the same response.

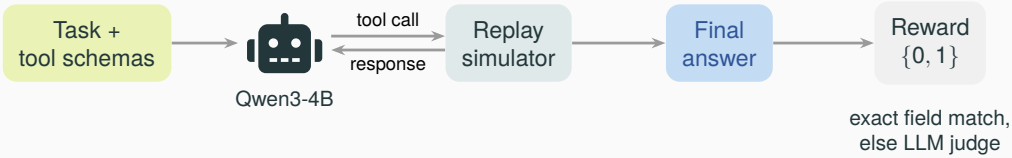
Why Replay Is Enough



✓ **Correct agent**: stays on the path. Every call is cached, so it gets real responses and reaches the ground truth.

✗ **Wrong agent**: leaves the path and may hit missing data. That does not matter: it is already wrong, and its reward is 0 either way.

Key assumption: every call on the path to the ground truth is cached.
Back-chaining guarantees it: the task was written from those very calls.



GRPO with dynamic filtering

4,944 training tasks **200** held-out test tasks

Evaluation

Does RL on FIREFLY help in distribution, and does it transfer?

Model

Qwen3-4B-Thinking-2507, trained with RL on FIREFLY.

Benchmarks

FIREFLY test set: 200 held-out tasks, pass@k over 16 rollouts.

Tau2-Bench: Retail, Airline, Telecom.

MCP-Atlas and MCPMark.

Baselines

The base checkpoint.

Claude Haiku 4.5, Sonnet 4.6, Opus 4.7 under the identical protocol.

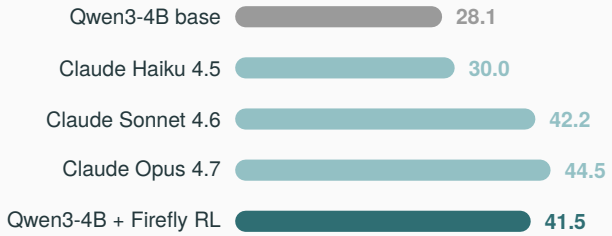
Results: A 4B Model Matches Claude Sonnet

Northeastern University

41.5

pass@1 on held-out Firefly tasks
up from 28.1 before RL

Qwen3-4B after GRPO on Firefly
matches **Claude Sonnet 4.6** (42.2)



At pass@8 it reaches 52.8, above Sonnet (50.3) and close to Opus (54.4).

Results: It Transfers

Northeastern University

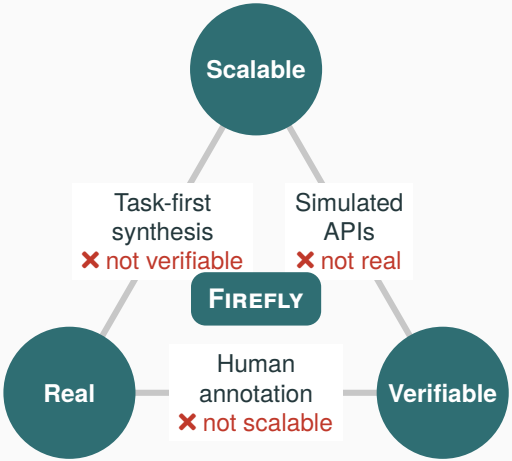
Model	Tau2-Bench			MCP-Atlas	MCPMark		
	Retail	Airline	Telecom		File Sys.	PG Easy	PG Std
Qwen3-4B	49.1	36.5	18.9	19.4	40.0	70.0	9.5
+ Firefly RL	**62.7**	**52.5**	**20.4**	**26.0**	**60.0**	**80.0**	**13.3**

Improves on every public benchmark, **multi-turn** and **real MCP execution** included.

None of these tools or environments appear in Firefly.

Conclusion

Conclusion



Invert the pipeline. Explore real APIs first, back-chain tasks from observed states: the ground truth is correct by construction.

Scale it. A tool graph steers exploration across 993 tools; a replay simulator removes drift from RL.

It works. 5,144 verified tasks take a 4B model to Claude Sonnet level, and the gains transfer.

Thank you!
Questions?

