

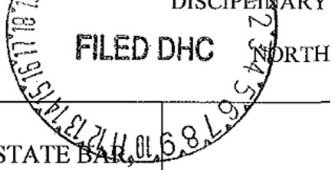
Pydantic AI

Why do we need agents?

- *Hold onto this question for a minute*
- Let's just talk about Pydantic first

Talking to the LLM

- Selecting models OpenAI-compatible endpoints



Structured outputs

NICHOLAS S. ACKERMAN, Attorney,

The composed pursuant to Plaintiff, t S. Ackern parties stip order and freely and any way th

Bas consent of the follow

1. North Car Chapter 84 Carolina S

2. admitted to rules, regulations, and Rules of Professional Conduct of the laws of the State of North Carolina.

3. Defendant was properly served with process, Hearing Panel with due notice to all parties.

4. During the relevant period referred to herein practice of law in Greensboro, Guilford County, North Carolina

5. On 25 September 2014, A.J. retained Defenc contempt order and custody modification.

7. Defendant filed a Motion to Show Cause September 2014.

8. Defendant failed to serve the opposir September 2014.

9. Defendant did not inform A.J. that he w party and that there would not be a hearing on 30 Septer

10. A.J. traveled from another county to appe

11. A.J. asked Defendant several times to

Based on the foregoing Findings of Fact and with the consent of the parties, the Hearing Panel enters the following:

CONCLUSIONS OF LAW

1. All parties are properly before the Hearing Panel, and the Panel has jurisdiction over Defendant and the subject matter of this proceeding.

2. Defendant's conduct constitutes grounds for discipline pursuant to N.C. Gen. Stat. § 84-28(b)(2) in that Defendant violated the Rules of Professional Conduct in effect at the time of his actions as follows:

- (a) By failing to notify A.J. that the hearing had been continued from the 30 September 2014 calendar and by otherwise failing to maintain communication with A.J., Defendant failed to keep his client reasonably informed about the status of the matter in violation of Rule 1.4(a)(3);
- (b) By failing to respond to A.J.'s inquiries about the status of the matter, Defendant failed to comply promptly with a reasonable request for information in violation of Rule 1.4(a)(4); and
- (c) By failing to provide a response to the fee dispute, Defendant acted in bad faith in the fee dispute.

```
class FilingDetails(BaseModel):
    plaintiff: str
    defendant: str
    charges: str
    city_of_practice: str
    penalty: str
```

```
{
  'Plaintiff': 'The North Carolina State Bar',
  'Defendant': 'Nicholas S. Ackerman',
  'Charges': 'Failure to keep client reasonably inform
the status of the matter; Failure to comply promptly w
reasonable request for information; Failure to partici
good faith in the fee dispute resolution process',
  'City_of_practice': 'Greensboro',
  'Penalty': 'Suspension of license for one (1) year,
for two (2) years as long as Defendant complies with
conditions.'
}
```

20. On 12 F required Defendant to p

21. Defendar

22. On 9 Ma dispute resolution proces 15G0195.

23. Defendant represent A.J. resolution process and appear in court on her behalf.

CONCLUSIONS REGARDING DISCIPLINE

“Can you ask for confidence scores?”

- *Kind of.*

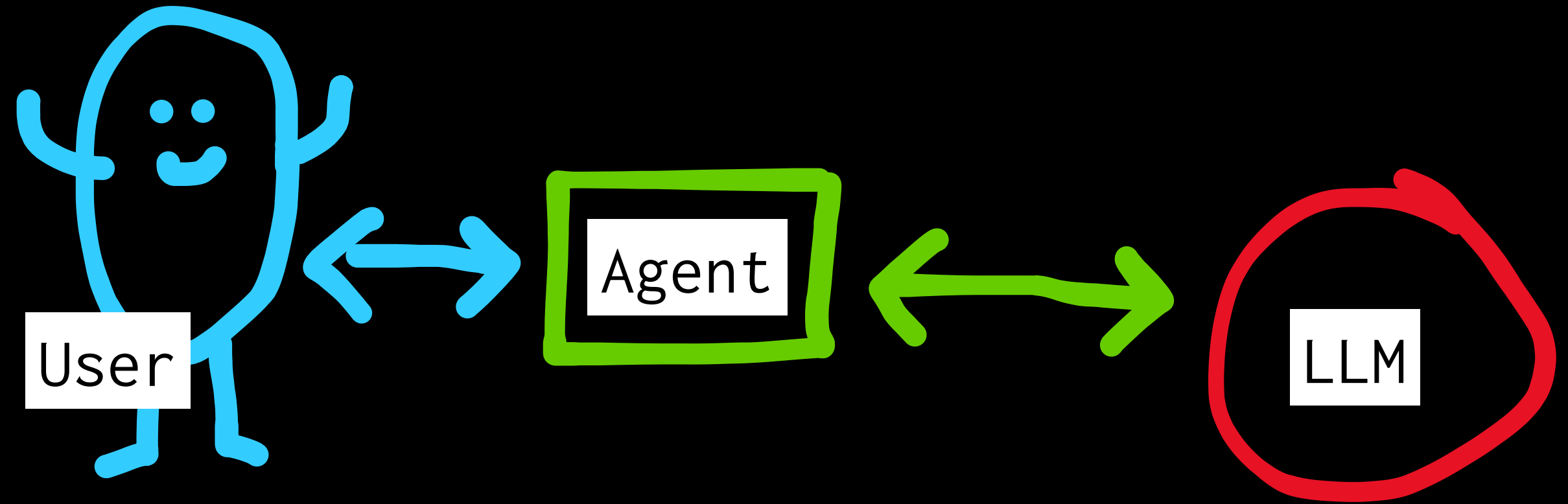
The infinite flexibility of tools

What does “agentic” really mean?

- ”Can do more than just write a few words”
- “Flexible”
- It has become the **default** now so it almost doesn’ t mean anything?
- ...*at least for consumer-facing products*

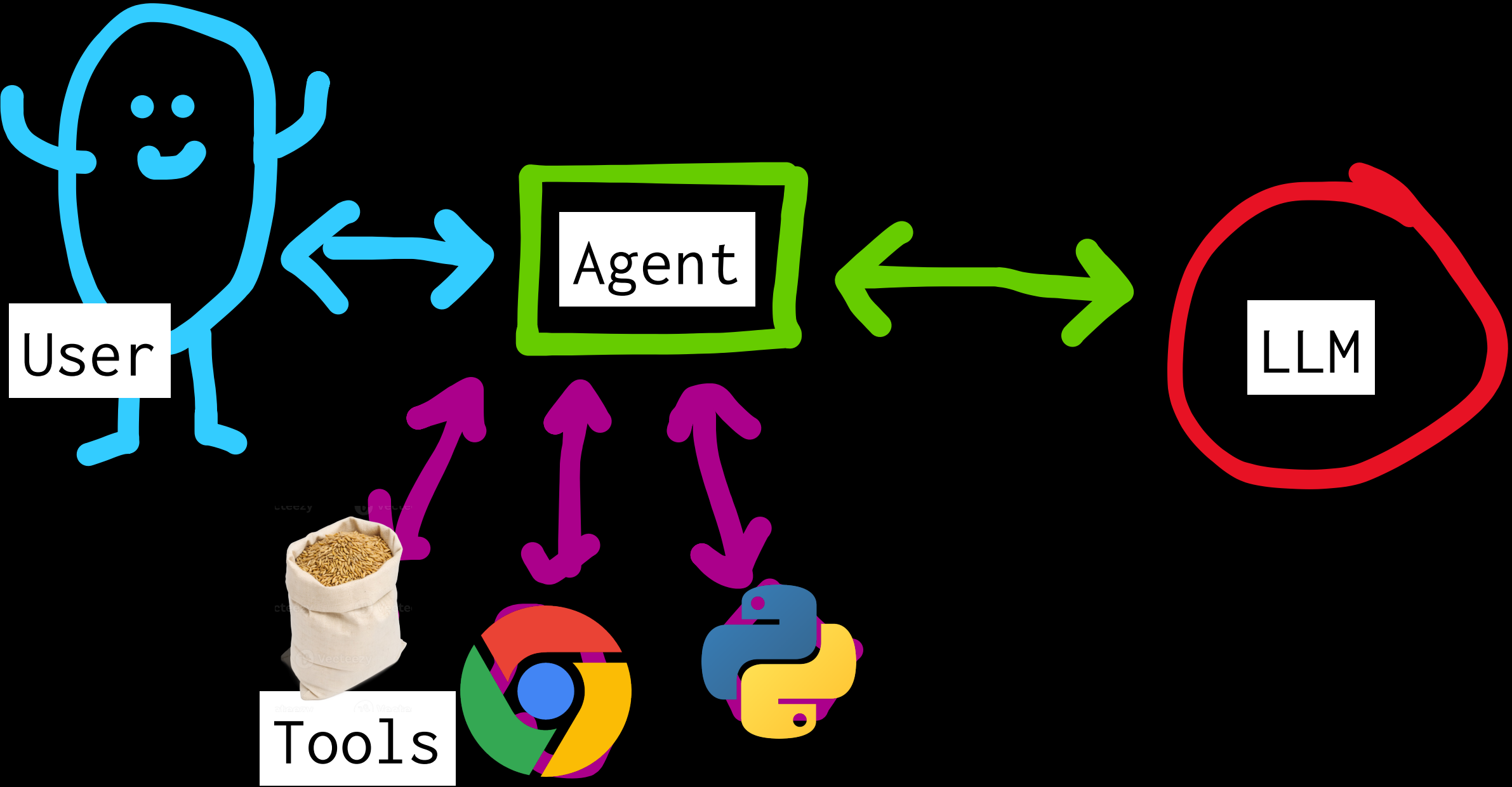


Direct conversation



Mediated conversation

** Slightly simplified*



User

Agent

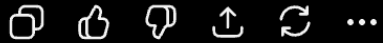
LLM

Tools

“What should I [the agent] do next?”

Step 1: Park on level ground, turn the engine off, set the parking brake, and let the engine cool for 10–15 minutes.

Done?



yes i am done

Step 3: Raise the front of the car if needed and securely support it with jack stands or ramps. Never rely on the jack alone.

Done?

car is up, yup

Step 4: Slide the drain pan under the oil drain plug at the bottom of the engine oil pan.

Done?

slid, yes

Step 2: Open the hood and remove the oil filler cap on top of the engine. This helps the oil drain smoothly.

Done?

yes i've removed it

Step 5: Use the correct wrench/socket to loosen the drain plug slowly by unscrewing it by hand and let the oil drain completely into the pan.

Done?

yup

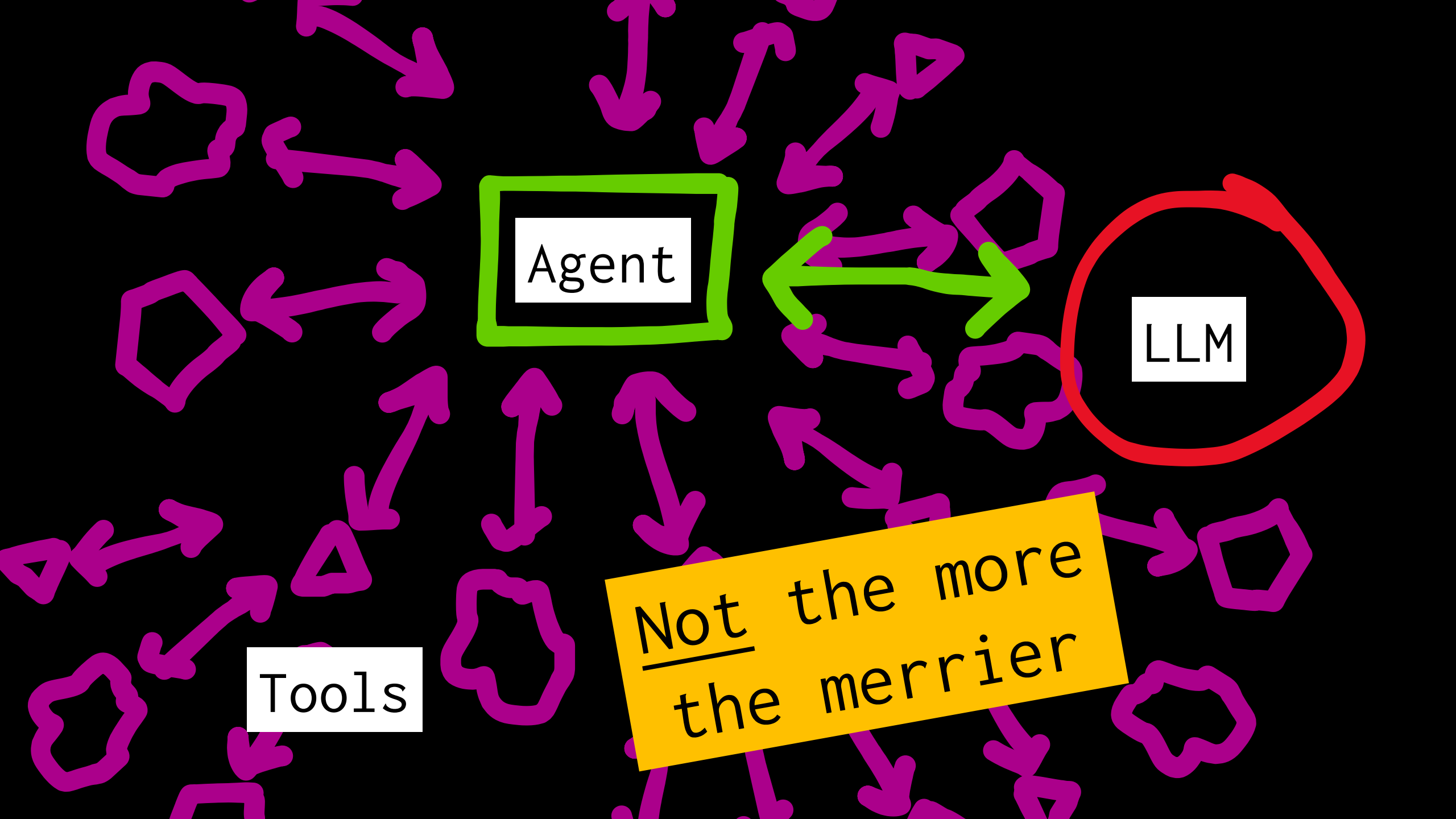
Native (provider) versus local tools

find me a good government corruption case to write about, i'm a fancy journalist. Search the internet.

Searching for best government corruption scandals investigative journalism angles 2...

MCP Servers

- **Universal APIs for (mostly) websites or other pieces of software**
- Canva, Figma
- Browsers/Playwright
- Powerpoint
- Filesystem
- *Anything.*

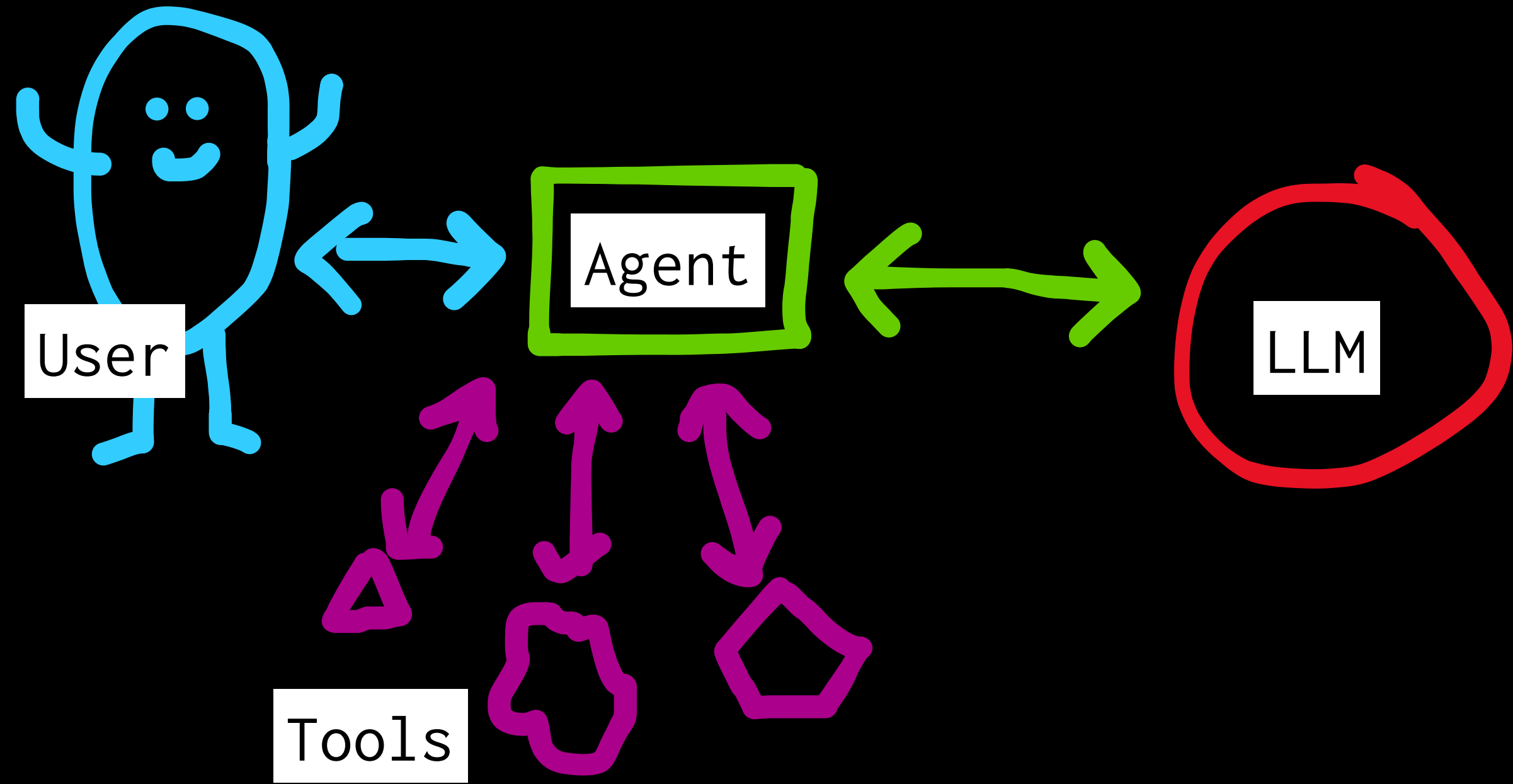


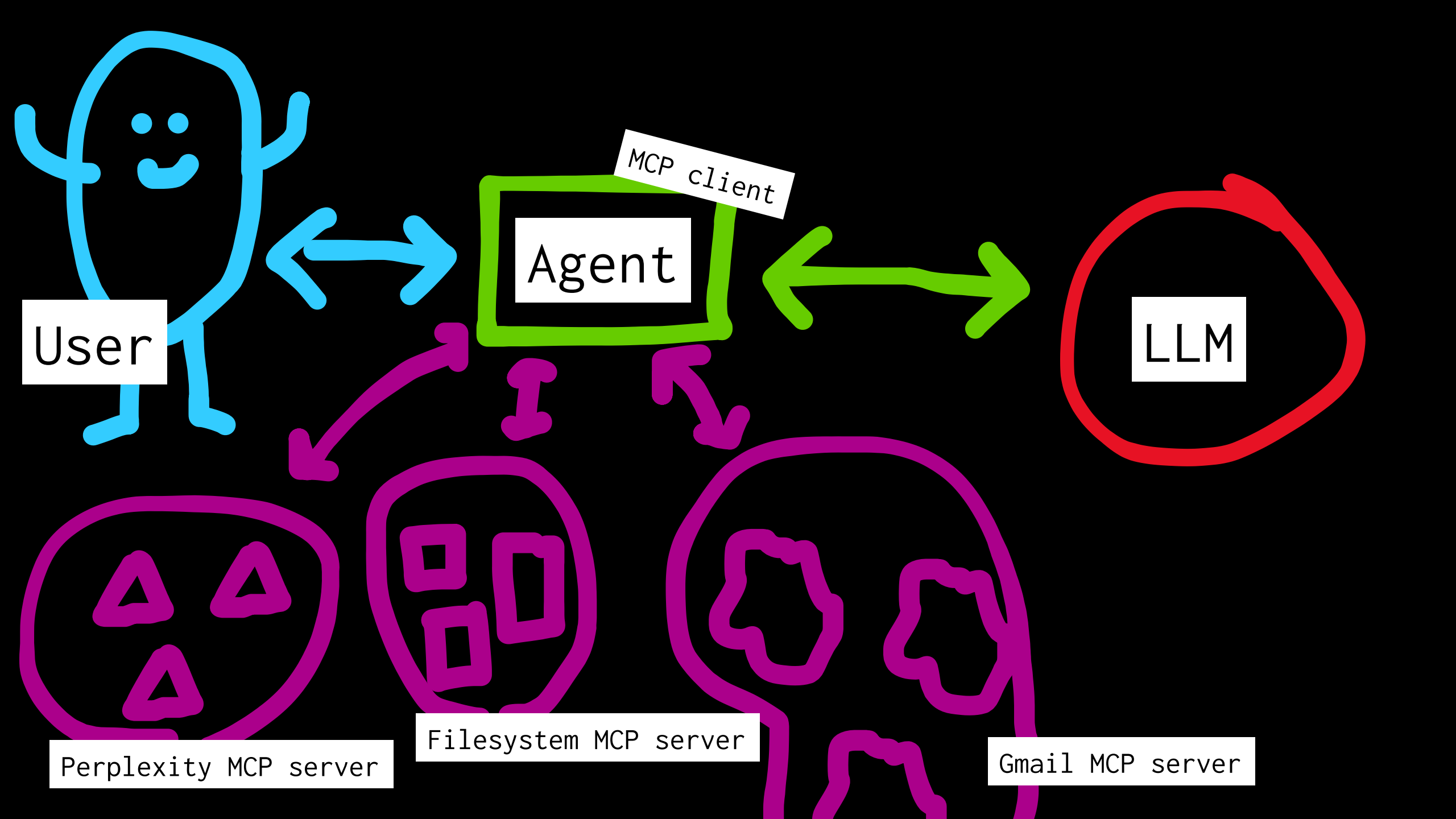
Agent

LLM

Tools

Not the more
the merrier





It isn't a library or API
but you can think of it as
a library or API

Observability and instrumentation

- What is even going on in there?
- The real reason why you want to

Instrumentation tools

- Braintrust
- Logfire
- Langfuse
- Arize Phoenix
- LangSmith
- A thousand others, much thanks to **OpenTelemetry**

```
from langsmith.integrations.opentelemetry import configure
from pydantic_ai import Agent

# Configure LangSmith tracing
configure(project_name="pydantic-ai-demo")
```

```
import logfire

from pydantic_ai import Agent

logfire.configure()
logfire.instrument_pydantic_ai()
```

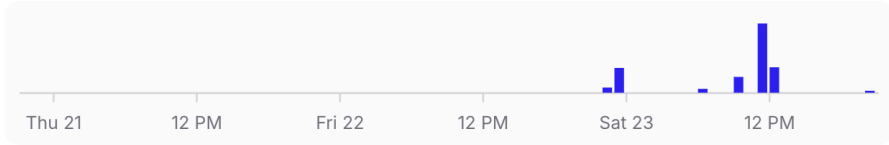
```
from braintrust.wrappers.pydantic_ai import setup_pydantic_ai
from pydantic_ai import Agent

# Initialize Braintrust tracing for Pydantic AI
setup_pydantic_ai(project_name="my-pydantic-project")
```

Logs

All logs view 3d Past 3 days

Search traces



| Created | Name | Input | Output |
|----------------------|--------------|--------------------|--------|
| 77 traces | | | |
| 1 May 23 08:38:54 PM | agent_run | {"toolsets": [...] | {"outp |
| 2 May 23 08:37:49 PM | agent_run | {"toolsets": [...] | {"outp |
| 3 May 23 12:17:53 PM | agent_run... | {"toolsets": [...] | Trac |
| 4 May 23 12:17:01 PM | agent_run... | {"toolsets": [...] | {"outp |
| 5 May 23 12:15:45 PM | agent_run | {"toolsets": [...] | {"outp |
| 6 May 23 12:15:32 PM | agent_run... | {"toolsets": [...] | {"outp |
| 7 May 23 12:15:19 PM | agent_run... | {"toolsets": [...] | {"outp |
| 8 May 23 12:15:08 PM | agent_run... | {"toolsets": [...] | {"outp |
| 9 May 23 12:14:55 PM | agent_run... | {"toolsets": [...] | {"outp |

4edcadd6 Find Share

Trace Timeline Thread Signals Views

Trace tree

- agent_run [instance] 52.11s, 279,852 tok
 - AI chat anthropic/claude-sonnet-4-5 3.33s, 1,393 tok
 - duckduckgo_search 0.79s
 - duckduckgo_search 1.11s
 - duckduckgo_search 1.88s
 - duckduckgo_search 0.94s
 - duckduckgo_search 1.51s
 - AI chat anthropic/claude-sonnet-4-5 3.24s, 15,763 tok
 - web_fetch 0.19s
 - web_fetch 0.51s
 - AI chat anthropic/claude-sonnet-4-5 12.93s, 127,793 tok

Tool span

duckduckgo_search
+ Dataset ... Flag for review

Metrics

Start 9h ago Duration 1.8847s
Offset 3.333s

Input

Pretty query
"Mark Ratcliff" "Melanie Ratcliff" Fredericksburg Virginia

Output

Pretty Loop agent

LangSmith

- Projects
- Annotation Queues
- Playground
- Deployments
- Datasets & Testing
- Prompts

Color Scheme

Settings

Documentation

LangChain Hub

Personal

ht-long-inspection-12

Runs Threads Monitor Setup

1 filter Last 7 days Root Runs LLM Calls All Runs

| Name | Input | Output |
|----------------------|----------------------------------|----------------------------------|
| LangGraph | null | BadReque |
| LangGraph | null | BadReque |
| LangGraph | null | BadReque |
| LangGraphUpdateState | tool: [{"url": "https://ww..."}] | tool: [{"url": "https://ww..."}] |
| LangGraph | human: What is the cur... | ai: The cur |
| LangGraph | human: What is the we... | ai: You car |
| LangGraph | human: Hi | ai: Hello! H |
| LangGraph | null | BadReque |
| LangGraph | human: Hi | human: Hi |
| LangGraph | null | BadReque |
| LangGraph | null | BadReque |
| LangGraph | human: What's the cur... | human: WI |
| LangGraph | human: Hi | ai: Hello! H |

LangGraph Studio

ChannelWrite<agent,mess...>

Run Feedback Metadata

START TIME: 07/14/2024, 04:32:10 PM

END TIME: 07/14/2024, 04:32:10 PM

TIME TO FIRST TOKEN: N/A

Input

```
AI
tavily_search_results_json
call_lWE7crecbj5687UKGSHaoSF
1 query: current weather in San Francisco
```

Trace Details

```
chain query 1.84s
  retriever retrieve 0.35s
  embedding embedding 0.28s
  chain synthesize 1.49s
    chain templating 0.01ms
  llm llm 1.49s
```

Output

```
AI
tavily_search_res
call_lWE7crecbj568.
1 query: curre
```

TRACE

- __start__ 0.00s graphstep:0
- agent 0.92s seqstep:15
- ChatOpenAI gpt-4o 0.90s seqstep:15
- ChannelWrite<agent.messages> 0.00s seqstep:2
- should_continue 0.00s seqstep:3
- ChannelWrite<branch:agent:shoul...> 0.00s seqstep:3
- action 2.02s
- tavily_search_results_json 2.01s seqstep:1
- ChannelWrite<action.messages> 0.00s seqstep:2
- agent 0.79s seqstep:35
- ChatOpenAI gpt-4o 0.79s seqstep:35

portkey

PM Portkey Monitor...

WORKSPACE

Shared Team Work...

- Getting Started
- Analytics
- Logs
- Playground
- Prompts
- Guardrails
- Configs
- Virtual Keys
- Plugins
- Members
- API Keys

Trace ID: 887b4f88-e970-5723-aa67-c43af09ed53d

Feb 24, 06:29:31 PM

Request Details

Guardrails & Feedback

Timeline

- Crew.kickoff 4.34 s
- Crew Created 0.38 ms
- Task.execute_sy... 2.81 s
- Task Created 0.06 ms
- Agent.execut... 2.81 s
- completion 2.81 s
- Task.execute_sy... 1.52 s
- Task Created 0.1 ms
- Agent.execut... 1.52 s
- completion 1.51 s

Total Cost: 0 cents

Cache Status: DISABLED

Request Timing: 4337.121 ms

traceld: 887b4f88-e970-5723-aa67-c43af09ed53d

spanid: 5780842670622223000

spanName: Crew.kickoff

parentSpa... null

startTime: 1740401967124329000

endTime: 1740401971461450000

_logType: opentelemetry

_source: crewai

framework... 0.102.0

Request (0 tokens)

_source: crewai

llm

Info Attributes Events 0

Messages

system

You are an expert Q&A system that is trusted around the world. Always answer the query using the provided context information, and not prior knowledge. Some rules to follow:

- Never directly reference the given context in your answer.
- Avoid statements like 'Based on the context, ...' or 'The context information shows ...'.

user

Context information is below.

 Sending 10 Million Inferences to Arize in 90 Seconds | Colab Link |

Arize supports sending in historical data with prediction timestamps up to 2 years. The data sent to Arize will be retained and visible in Arize for up to 2 years to extend this window.

Given the context information and not prior knowledge, answer the query.
 Query: Can I send truth labels to Arize separately?
 Answer:

Output

Yes, you can send truth labels to Arize separately. Arize supports sending in historical timestamps, so you can include truth labels along with your predictions when se

Things to keep in mind

You are only limited by your giant brain

- A tool can do literally anything
- An agent can do literally anything
- *Absolutely literally anything*
- There are no limitations!
- There are only two issues:
 - 1) Making sure it's (going to be) useful
 - 2) Making sure it does what you say it does

Scraper-Writer Agent

A conversational AI agent that explores websites and writes Playwright scraping scripts for you.

How it works: You describe what you want to scrape in plain English. The agent navigates the page using Playwright, examines the DOM structure, and produces a complete scraping script that runs without AI in the loop.

What you need:

- A Google account
- A free `GOOGLE_API_KEY` from [Google AI Studio](#)
- Set it as an environment variable, or paste it in the cell below

```
%%pip install --quiet "pydantic-ai-slim[google]" playwright beautifulsoup4 pandas lxml
# !playwright install-deps
# !playwright install chromium firefox
```

```
import os
try:
    from dotenv import load_dotenv
    load_dotenv()
except ImportError:
    pass
# Detect if we're running in Google Colab
```

```
@agent.tool
async def select_dropdown(ctx: RunContext[Deps], label_or_selector: str, option: str) -> str:
    """Select an option from a <select> dropdown by its label text or CSS selector.
    print(f" -> select_dropdown({label_or_selector!r}, {option!r})")
    try:
        locator = ctx.deps.page.get_by_label(label_or_selector)
        if await locator.count() > 0:
            await locator.first.select_option(label=option)
            return f"Selected '{option}' from '{label_or_selector}'"
        except Exception:
            pass
    try:
        await ctx.deps.page.select_option(label_or_selector, label=option)
        return f"Selected '{option}' from '{label_or_selector}'"
    except Exception as e:
        return f"Could not select from '{label_or_selector}': {str(e)[:200]}"

@agent.tool
async def get_network_requests(ctx: RunContext[Deps]) -> str:
    """Return captured XHR/fetch API requests (JSON only). Useful for discovering
    print(f" -> get_network_requests()")
    if not captured_requests:
        return "No API requests captured yet. Navigate to a page and interact with it."
    lines = []
    for r in captured_requests[-20:]:
        line = f"[{r['method']}] {r['status']} | ~{r['record_count']} records | {r['url']}"
        if r.get('post_data'):
            line += f"\n POST body: {r['post_data'][:200]}"
        lines.append(line)
    return "\n".join(lines)
```

README MIT license

Natural PDF

CI passing

A friendly library for working with PDFs, built on top of [pdfplumber](#).

Natural PDF lets you find and extract content from PDFs using simple code that makes sense.

- [Complete documentation here](#)
- [Live demos here](#)



✓ experiments

- > __pycache__
- > llm_agent_harness
- > llm_agent_harness_viewer
- > llm_agent_harness_workspace
- > performance

Sources: 12 (2 unverified) Quotes: 8 Findings: 26

691 words · 4,398 chars · 6 em-dashes Tools 23

Article Repo 6 Notebooks 1 Data 1 Content + Add content

Retrograde · May 8, 2026 · 691 words · 4,398 chars · 6 em-dashes

Submission

The Capital Ledger

Markets · Cryptocurrency

Inside crypto's \$2 billion rug-pull era, where dogs, frogs and abandoned tokens trace a cycle that resets every few months.

In the 24 months ending December 2024, retail investors poured money into the cryptocurrency market at a pace not seen since 2021 — and watched roughly \$2 billion of it disappear into so-called "rug pulls" on meme coins, the dog- and frog-themed tokens that have come to define the latest speculative cycle.

A Capital Ledger analysis of [monthly market data](#)

j k nav enter details esc clear x

and DEXScreener

found that while bitcoin tripled and other more

Sources 12

HIGH source 1 mention source-verifier

Anatoly Yakovenko

0 linked quotes

LOW source 1 mention source-verifier

Dune Analytics

1 linked quote

LOW source 1 mention source-verifier

Quotes 8

write bitcoin tripled and ether more than doubled over the period, the meme-coin sector's market value surged by nearly 300% before c...

QUOTE quote 2 mentions quote-attributer

Quote from Vitalik Buterin

its co-founder, Vitalik Buterin, has long argued is essential to the network's long-term legitimacy

Findings 26

LOW finding run_agentic_review

Unsupported Claim

Bitcoin's hard cap is stated as '22 million coins.' Bitcoin's actual hard-coded supply cap is 21 million BTC — one of the most well-known facts in cryptocurrency. This is a clear factual error.

LOW finding run_agentic_review

Unsupported Claim

The Squid Game token collapse is attributed to '2022,' but the token collapsed in late October/November 2021. No 2022 Squid Game token event is on record.

LOW finding run_agentic_review

Unsupported Claim

The article states Ethereum 'transitioned in 2021 from energy-hungry proof of work mining to a proof of stake model.' The Ethereum Merge — the PoW-to-PoS transition — occurred on September 15, 2022, not 2021.

LOW finding run_agentic_review

Unsupported Claim

PEPE is described as 'a Solana-based token.' PEPE (the Pepe the Frog meme coin) launched on Ethereum in April 2023, not on Solana. BONK and WIF are the prominent Solana meme coins. This error misattributes a major token to the wrong bloc...

Chat

what date does the data a... New

2 messages · May 12, 1:34 PM

The data analysis ends on **December 2024** (specifically 2024-12-01, based on the last row in `monthly_crypto.csv`).

The dataset contains 24 monthly observations spanning from January 2023 through December 2024.

Sources used (1)

Tools used (2)

Type a message...

Tools

Open in admin Rebuild Remove

Translate to french

translate to french

enabled

Try it

Answers in chat

Open in admin Rebuild Remove

Analysis 6

CURRENT PROFILE Automatic

Force fresh analysis

Core Core evidence extraction and verification primitives

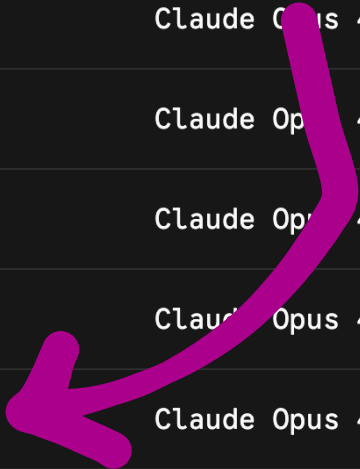
Why are you building this yourself?

- *Why not just use Claude Code?*
- It's about harness engineering and controlling the runtime

Showing 10 entries

Clear filters

| Search leaderboard | | Select agents | Claude Opus 4.6 | Select... | Verified only | | |
|--------------------------|------|---------------|-----------------|------------|----------------|-----------|-------------|
| <input type="checkbox"/> | Rank | Agent | Model | Date | Agent Org | Model Org | Accuracy |
| <input type="checkbox"/> | 13 | Meta-Harness | Claude Opus 4.6 | 2026-05-14 | Stanford IRIS | Anthropic | 76.4% ± 2.4 |
| <input type="checkbox"/> | 16 | Capy | Claude Opus 4.6 | 2026-03-12 | Capy | Anthropic | 75.3% ± 2.4 |
| <input type="checkbox"/> | 19 | Terminus-KIRA | Claude Opus 4.6 | 2026-02-22 | KRAFTON AI | Anthropic | 74.7% ± 2.6 |
| <input type="checkbox"/> | 22 | TongAgents | Claude Opus 4.6 | 2026-02-22 | Bigai | Anthropic | 71.9% ± 2.7 |
| <input type="checkbox"/> | 24 | Junie CLI | Multiple | 2026-03-07 | JetBrains | Multiple | 71.0% ± 2.9 |
| <input type="checkbox"/> | 25 | Droid | Claude Opus 4.6 | 2026-02-05 | Factory | Anthropic | 69.9% ± 2.5 |
| <input type="checkbox"/> | 28 | Crux | Claude Opus 4.6 | 2026-02-23 | Roam | Anthropic | 66.9% ± N/A |
| <input type="checkbox"/> | 30 | Mux | Claude Opus 4.6 | 2026-02-13 | Coder | Anthropic | 66.5% ± 2.5 |
| <input type="checkbox"/> | 38 🌟 | Terminus 2 | Claude Opus 4.6 | 2026-02-06 | Terminal-Bench | Anthropic | 62.9% ± 2.7 |
| <input type="checkbox"/> | 51 🌟 | Claude Code | Claude Opus 4.6 | 2026-02-07 | Anthropic | Anthropic | 58.0% ± 2.9 |



Orchestration

Single-agent workflow (or router)

- It's just very simple so you don't worry about anything.

Planner/executor

- All about **decomposition**

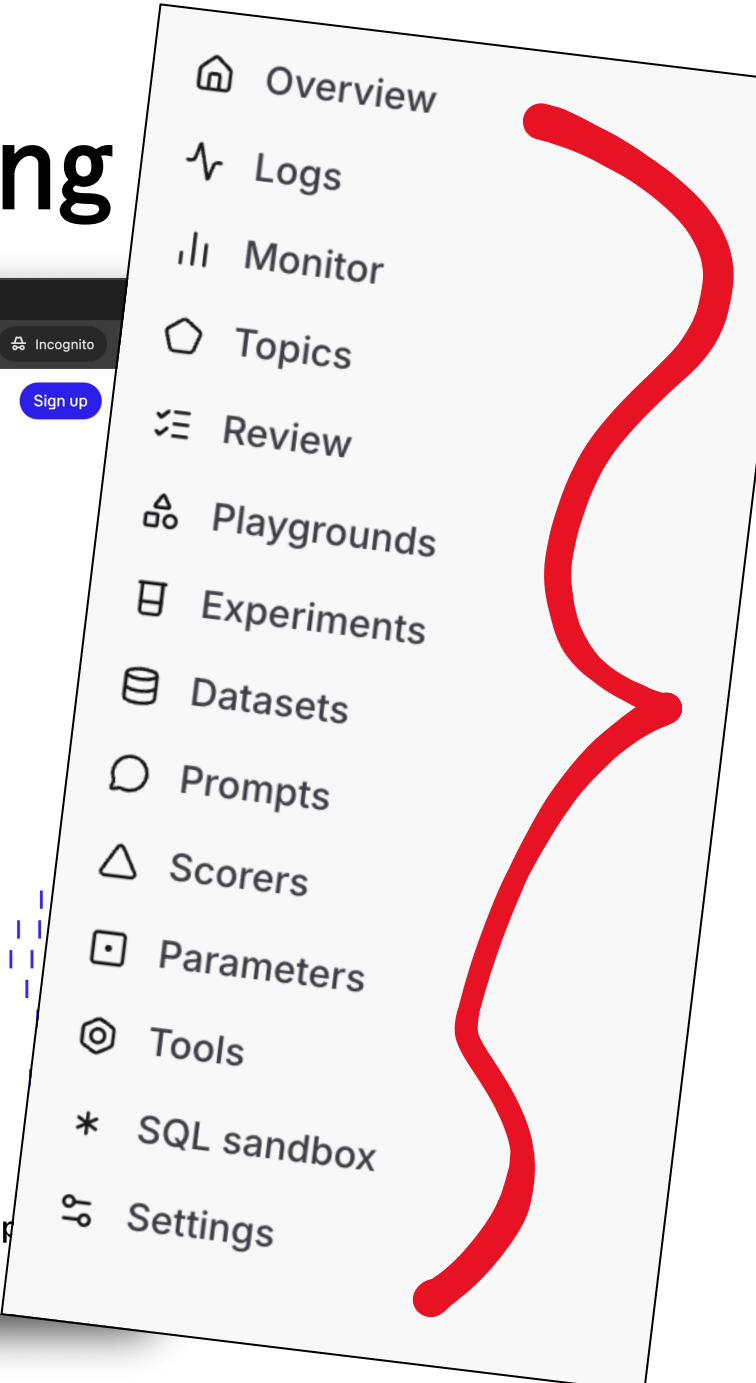
Supervisor-worker

- All about **specialization**

What if not Pydantic?

- Anything!
- First party:
 - OpenAI's Agent SDK is good, Google's Agent ADK is good
 - Claude's documentation I can't understand for the life of me
- Third party:
 - LangGraph
 - CrewAI
 - LlamaIndex
- Managed agents is a big selling point these days

Observability/evals/prompting



braintrust.dev (and 2,000 others)

Playground 2 - Playgrounds

braintrust.dev/app/Little%20Columns/p/default-otel-project/playgrounds/Playground%202

Little Columns / default-otel-project / Playgrounds

Playground 2

Base task

GPT-5 mini

User

You're a journalist. Is the following tip worth researching? {{tip}}

Comparison task

Claude 4.5 Haiku

User

You're a journalist. Is the following tip worth researching? {{tip}}

Structured output

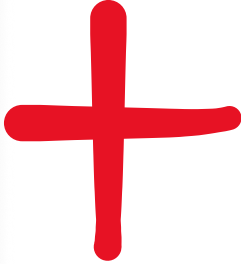
worth_rese...

Structured output

worth_reser...

Worth Researching

| Input | Base | Comparison |
|--|--|--|
| email: carter.linda.durham@gmail.com name: Linda Carter tip: I lost my left earring somewhere between the DPAC parking deck and the entrance to the theater last night. It is a gold hoop. If anyone at the station found it, or if a viewer turned it in, please let me know. It has great sentimental value. | Output Dec 10, 2025 0s 279 0.00 worth_researching: false | Output Dec 10, 2025 0.1s 36 worth_researching: false |



```

email: carter.linda.durham@gmail.com
name: Linda Carter
tip: I lost my left earring somewhere between the
DPAC parking deck and the entrance to the theater
last night. It is a gold hoop. If anyone at the
station found it, or if a viewer turned it in, please
let me know. It has great sentimental value.

```

```

email: durham.watchdog.11@protonmail.com
name: Anonymous
tip: I saw a drone flying over the American Tobacco
Campus near the water tower. It was hovering very
still. I think it might be spying on the people
eating at the restaurants. Is this legal? You should
warn people not to eat outside until we know who is
piloting it.

```

```

email: k.oconnor55@verizon.net
name: Kevin O'Connor
tip: My grandson, Timmy, just learned how to tie his
shoes all by himself. He is five years old and

```

Input

GPT-5 mini

Base

Claude 4.5 Haiku

Tradeoff Base has better latency. Comparison has be

email: carter.linda.durham@gmail.com
name: Linda Carter
tip: I lost my left earring somewhere between the DPAC parking deck and the entrance to the theater last night. It is a gold hoop. If anyone at the station found it, or if a viewer turned it in, please let me know. It has great sentimental value.

Output Dec 10, 2025 0s 279 0.00
worth_researching: false

Output Dec 10, 2025 0.1s 36 0.0
worth_researching: false

% Worth Researching 100%

% Worth Researching 100%

email: durham.watchdog.11@protonmail.com
name: Anonymous
tip: I saw a drone flying over the American Tobacco Campus near the water tower. It was hovering very still. I think it might be spying on the people eating at the restaurants. Is this legal? You should warn people not to eat outside until we know who is piloting it.

Output Dec 10, 2025 0s 151 0.00
worth_researching: true

Output Dec 10, 2025 0.1s 36 0.0
worth_researching: false

% Worth Researching 0%

% Worth Researching 100%

email: k.oconnor55@verizon.net
name: Kevin O'Connor
tip: My grandson, Timmy, just learned how to tie his shoes all by himself. He is five years old and

Output Dec 10, 2025 0s 279 0.00
worth_researching: true

Output Dec 10, 2025 0.1s 36 0.0
worth_researching: false

Scorers: code-based vs LLM-as-judge

```
from typing import Any
import re

def handler(input: Any, output: Any, expected: Any, metadata: dict[str, Any]):
    """
    A heuristic scorer to detect overly personal language.
    Fast and explainable, but imperfect.
    """
    output_text = str(output) if output else ""
    text = output_text.strip().lower()

    # Patterns that indicate personal/informal language
    PERSONAL_PATTERNS = [
        r"\bthank(s| you)\b",
        r"\bi (really)?(appreciate|feel|think|believe|agree|disagree)\b",
        r"\bi'm\b",
        r"\bmy (view|opinion|take|heart)\b",
        r"\bdear\b",
        r"\bsincerely\b",
        r"\bhere'?s\b",
        r"\bgreeting\b",
    ]

    matched_patterns = []
    for pattern in PERSONAL_PATTERNS:
        if re.search(pattern, text, flags=re.IGNORECASE):
            matched_patterns.append(pattern)

    return {
        "score": 0.0 if matched_patterns else 1.0,
        "name": "Not Too Personal",
        "metadata": {
            "matched_patterns": matched_patterns,
            "text": text,
        }
    }
```

You are grading whether the text uses overly personal, informal, or therapeutic language.

Evaluate the TEXT for:

- Greetings (e.g., "Dear", "Hi")
- Thanks or sign-offs (e.g., "Thank you", "Sincerely", "Best regards")
- First-person emotional framing (e.g., "I appreciate", "I feel", "I believe")
- Conversational or therapeutic tone

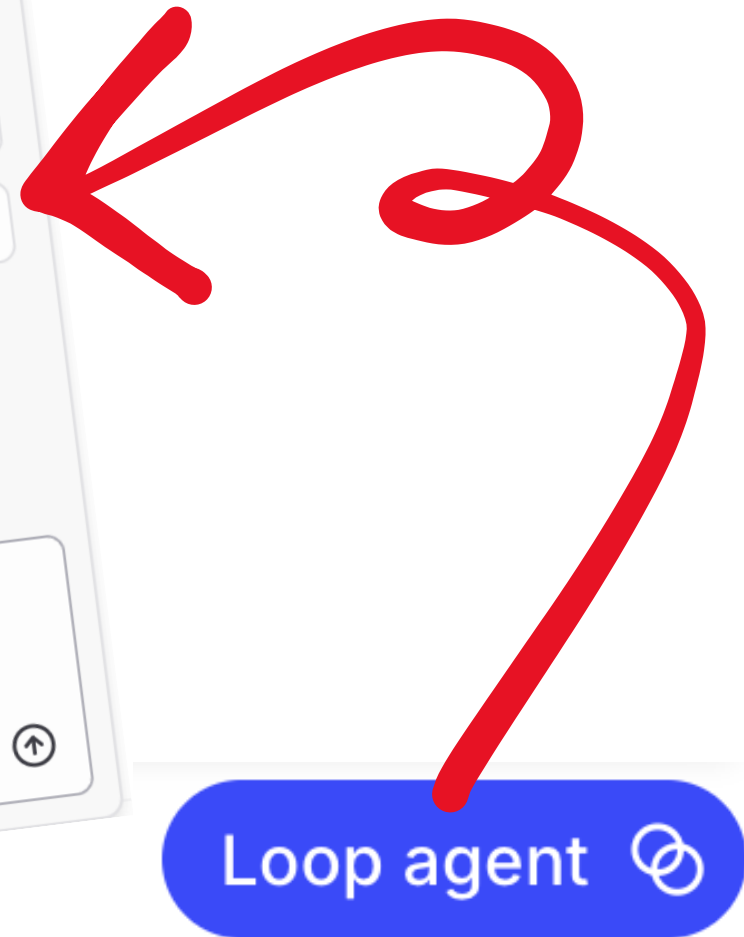
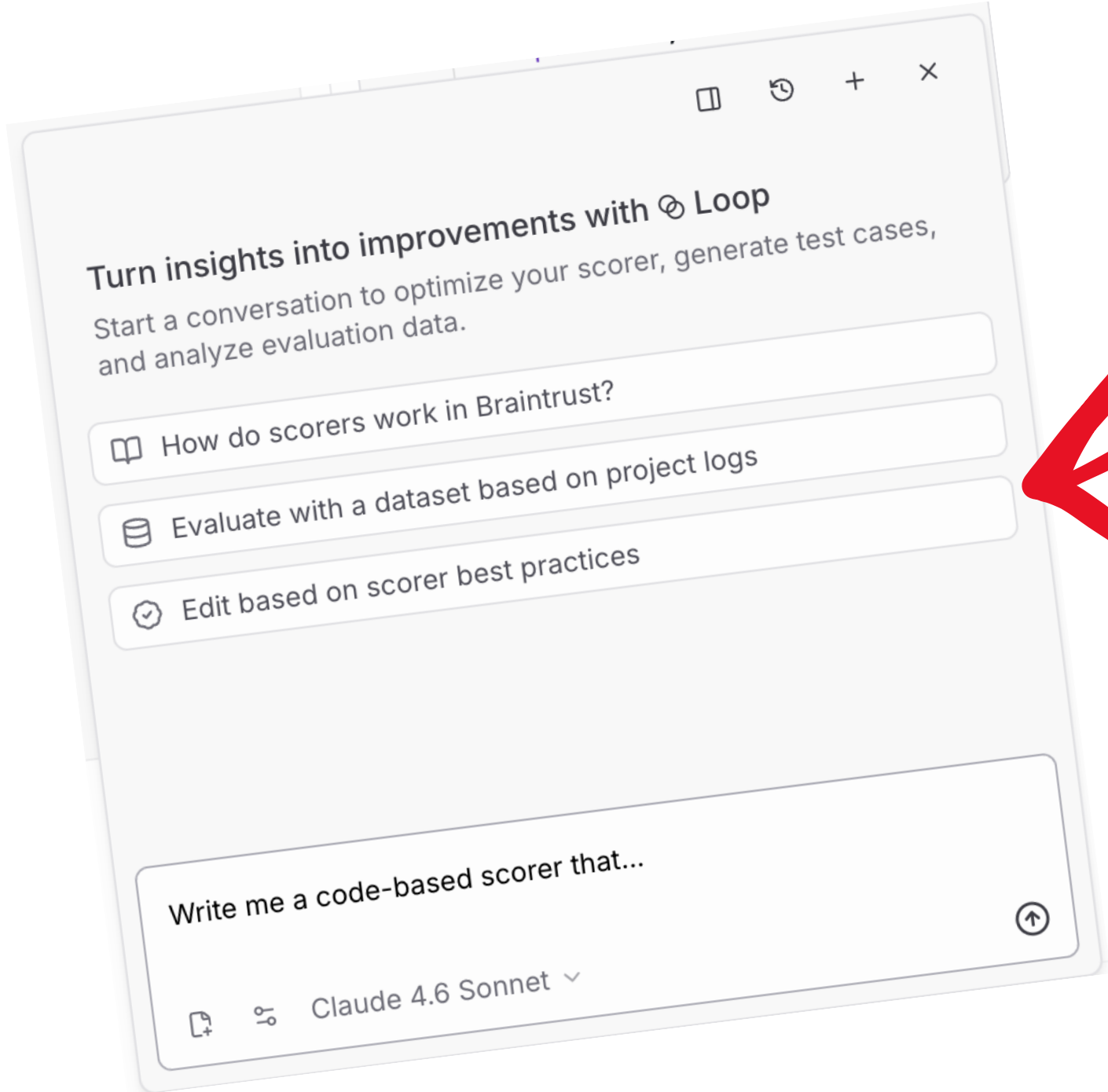
Grade as:

- ****a**** if the text is professional, analytical, and avoids personal language entirely
- ****b**** if the text has mild first-person framing once (like "I think") but is otherwise professional
- ****c**** if the text contains greetings, thanks, sign-offs, or strongly personal/therapeutic tone

TEXT:

{{output}}

Return ONLY the letter ****a****, ****b****, or ****c****.



Your evals todo-list cheatsheet

1. Build a dataset of possible **inputs**
 - Real-life are best, but there are some techniques for synthetic ones
 - Have a grouchy teammate who says “it’ll *NEVER* work in *THIS* case” try to write the hard ones.
 - “If it can’t do these successfully, it can’t go to production”
2. Make your **prompt** alone or with input
3. Throw the **inputs** and **prompt** together to get **outputs**
4. Have the good-spirited review a spreadsheet of **inputs + outputs** and score them.
 - Yes/No is much easier to deal with than 1–5 scores
5. Use their responses to teach an **LLM judge** about edge cases
6. **Repeat!!!** Although if you get to 100% then your evals probably aren’t difficult enough