Evals

john@ouachitalabs.com

About me

- Live in NLR with my wife and son
- Currently a senior software engineer on the DSML team @ Numerator
- All opinions are my own
- Connect on LinkedIn!
- Send me an email!

LinkedIn



john@ouachitalabs.com

Who this talk is for

01 Anybody interested in how model progress is tracked

02 Folks building software where AI plays a critical role

03 Creatives who want to elicit out-of-distribution behavior from LLMs



sketch.js

```
async function setup() {
  createCanvas(600, 600, WEBGL);
  angleMode(DEGREES);
  brush.scaleBrushes(3);
}

function draw() {
  translate(-width / 2, -height / 2);
  let paperColor = "#f4ebd8";
  background(paperColor);

  randomSeed(1024);
  noiseSeed(1024);

  let cx = 280;
  let cy = 260;

  let leaves = [
    { ang: 130, len: 260, spread: 45 },
    { ang: 25, len: 230, spread: 40 },
    { ang: 300, len: 210, spread: 35 }
  ];

  brush.noStroke();
  for (let i = 0; i < leaves.length; i++) {
```

[1] Flower art

john@ouachitalabs.com

What are evals?

Also known as **testing** and benchmarking

Evals involve the following:



Model or “agent”



Environment

- Could be just their own context
- May be a docker container, sandbox, etc



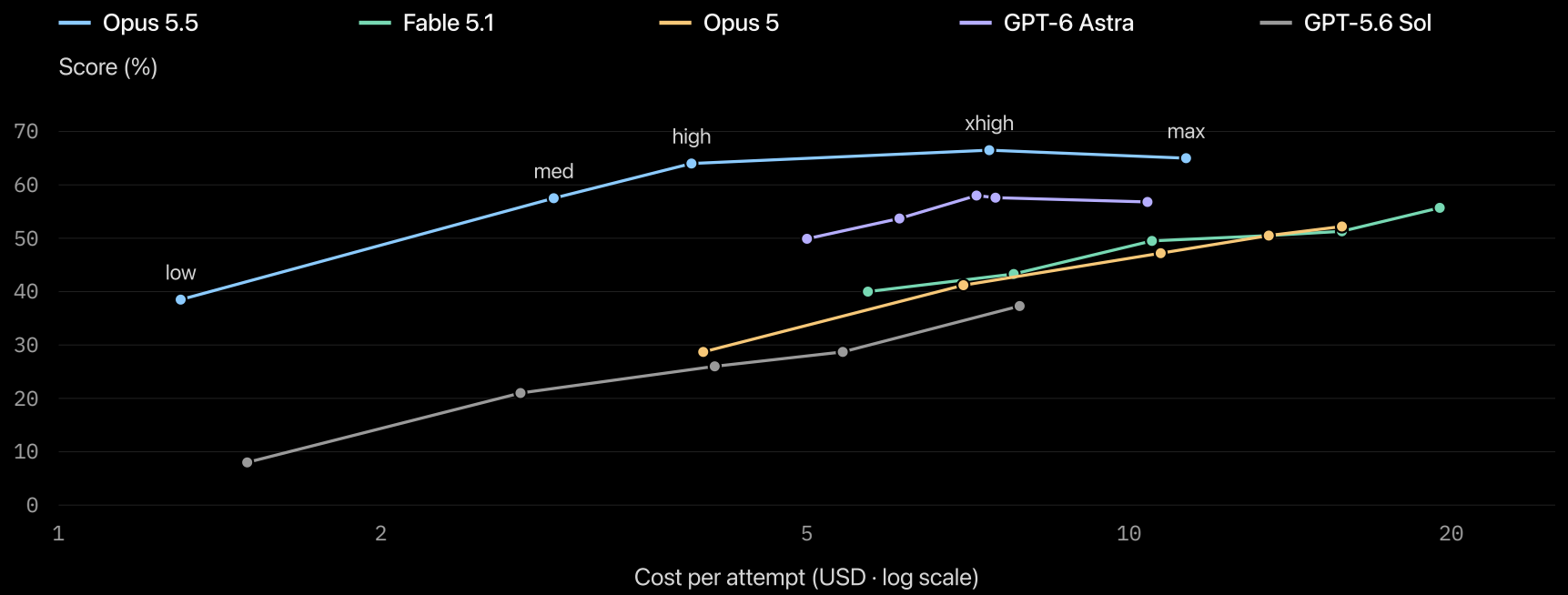
Grader

(code, LLM judge, human)

john@ouachitalabs.com

AGENTIC TERMINAL CODING

Terminal-Bench 4.0



Positions approximated from the source figure. Cost uses a logarithmic scale. [2] Source

john@ouachitalabs.com

TESTING & DESIGN

Tests shape the software

Testing is a selection pressure during the build.

- An executable check makes intent concrete
- Hard-to-test behavior pushes you toward smaller parts, clearer interfaces, and isolated dependencies
- Run, change, repeat: feedback shapes the implementation while you're building it

---

The evals we build today exert the same pressure.

john@ouachitalabs.com

Evals in AI powered software

Writing evals helps you

01

Define the performance axes you care about

02

Break up the problem into smaller ideas

03

Understand where complexity helps and hurts

04

Build more malleable systems

Still not perfect

Demystifying evals (Anthropic)



john@ouachitalabs.com

Evaluate RAG retrieval

Semantic similarity is an engineering decision. Back it with data.

- **Embedding space**
  Test paraphrases, domain terms, exact identifiers, and chunk boundaries against labeled relevance
- **Vector search**
  Compare approximate search with exact neighbors. Check filters and missed matches
- **Keyword baseline**
  Use the same queries and corpus. Compare relevance, p95 latency (including query embedding), and cost

MEASURE RETRIEVAL QUALITY

`Precision@k`	What fraction of the top k chunks is relevant?
`Recall@k`	What fraction of all relevant chunks appears in the top k?
`F1`	The harmonic mean of precision and recall at a chosen cutoff.

Use representative queries and labeled relevant chunks. Nearest in vector space does not guarantee relevance.

[8] Introduction to Information Retrieval

Choose semantic, keyword, or hybrid retrieval with data. Then measure answer quality across the complete RAG pipeline.

john@ouachitalabs.com

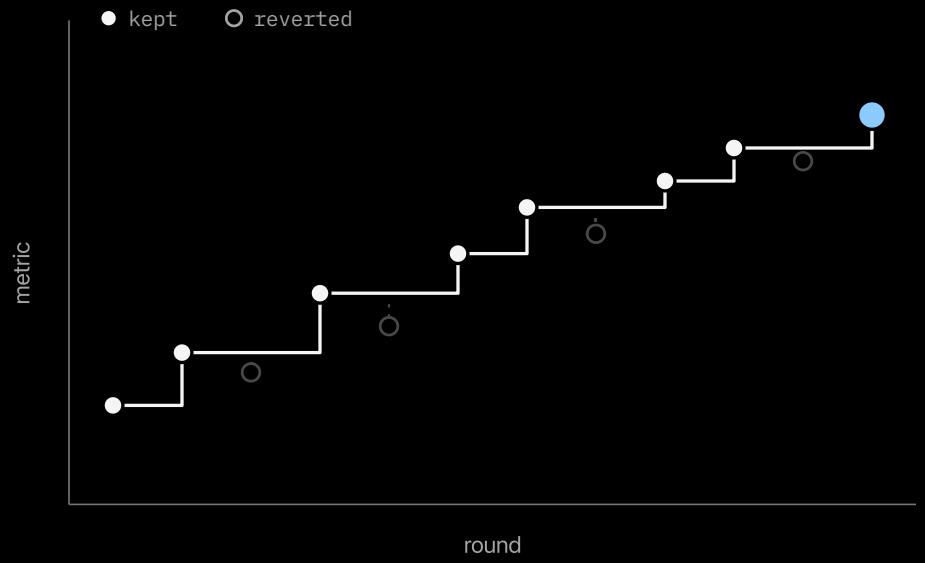
Optimization

john@ouachitalabs.com

Hillclimbing starts with a metric

A number you can score after every change

- Accuracy
- Pass rate
- Cost per ticket
- Trigger rate



john@ouachitalabs.com

The hillclimbing loop

One change per round. Keep improvements; revert overfit or regression.

- Choose one thing to edit, such as the prompt. Hold the model and the rest of the harness fixed.
- Repeat baseline runs to understand normal score variation.
- Read failures on the train split, form a hypothesis, and propose one change.
- Score the change on train and held-out test. Keep it when both improve beyond the noise.
- Revert regressions or gains that do not generalize. Repeat, keeping test examples out of the analyzer.

john@ouachitalabs.com

HILLCLIMBING

GEPA (prompt optimization)

Use eval failures to propose a better prompt. Then measure it.

Inspect failures

Give an LLM the failed runs, grader feedback, and context.

Reflect

Identify what went wrong and what the prompt needs to change.

Revise the prompt

Turn that feedback into a candidate prompt.

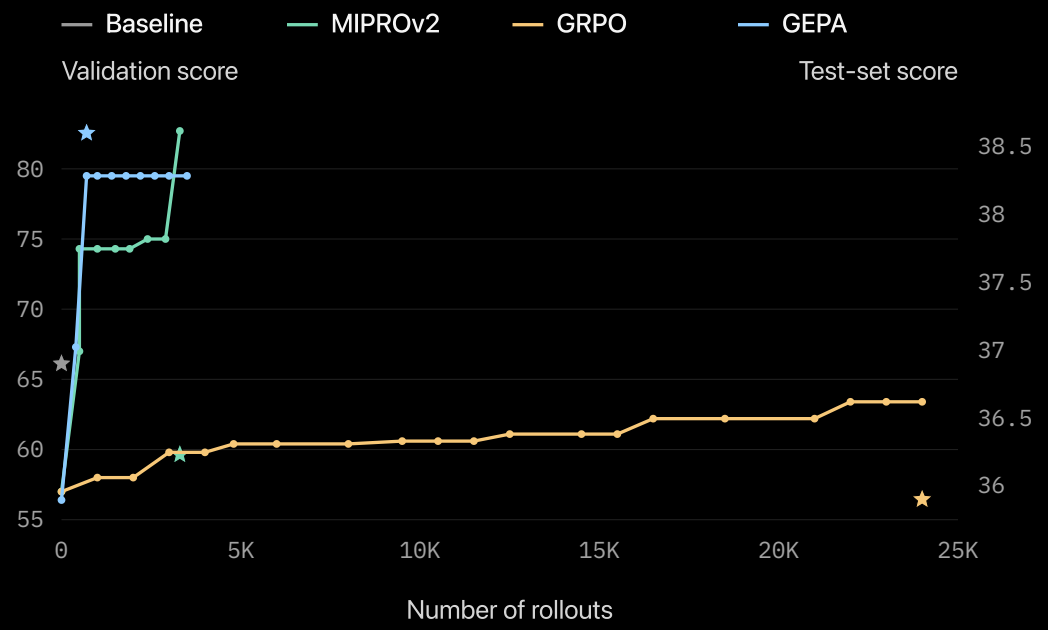
Run evals again

Compare results and keep the improvements.

Repeat with the next round of failures

john@ouachitalabs.com

GEPA (prompt optimization)



Lines / circles: validation (left). Stars: test set (right).
Curves and points approximated from the source figure. [4] GEPA · [6] IFBench

GEPA

Big jump in under 1K rollouts

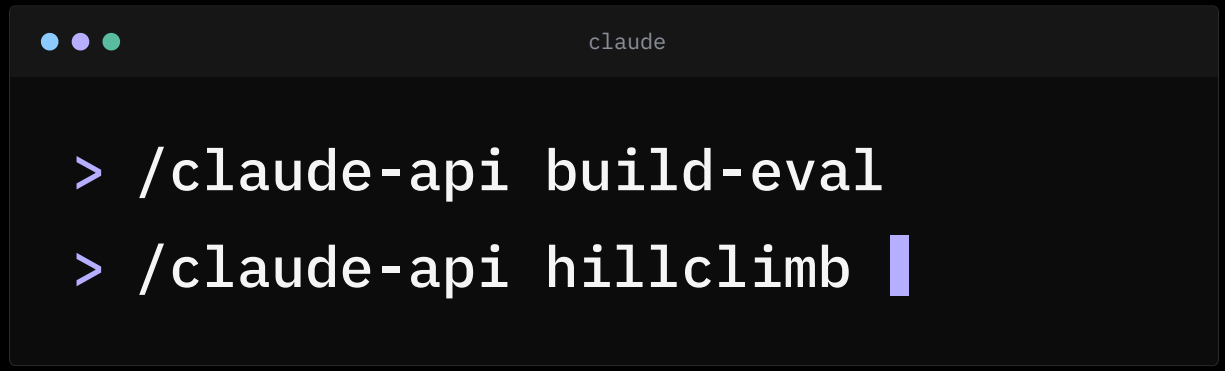
GRPO

Still lower after 24K

john@ouachitalabs.com

HILLCLIMBING

There's a skill for that now



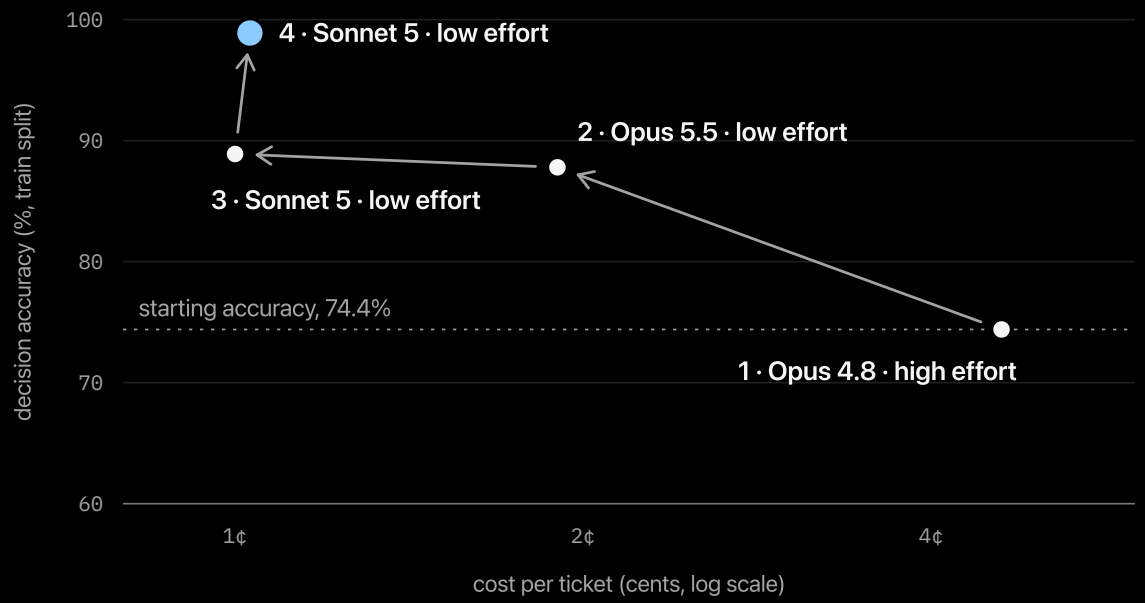
→ Your competitors are already running evals

[9] claude.dev · Sep 28, 2026 [10] claude-api skill

john@ouachitalabs.com

Hillclimbing for cost reduction

44 support tickets · 14 held out



74.4% → 98.9%
Accuracy

4.6¢ → ~1¢
Cost per ticket

78.6% → 90.5%
Held-out accuracy

[9] Automating eval design and hillclimbing

john@ouachitalabs.com

Supervised fine tuning

- Still used extensively
- Evals produce training data
- Dataset of example inputs and outputs

Some of the issues you may run into

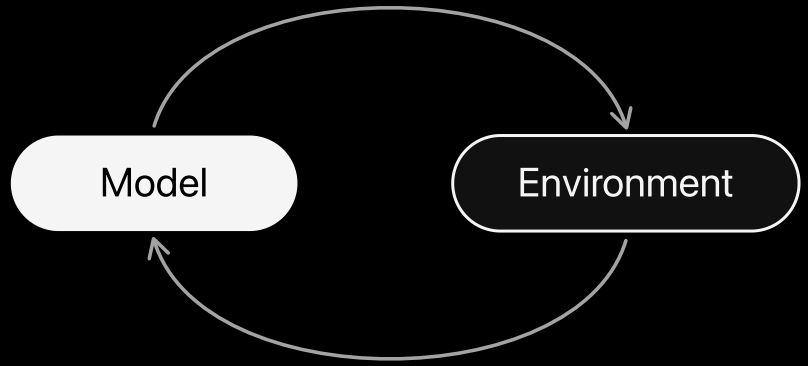


- General intelligence drops off a cliff
- Overfitting concerns and training distribution issues

john@ouachitalabs.com

Reinforcement learning

- Used for adapting a model to an environment
- Difficult to get right
- Can be extremely effective for agents



john@ouachitalabs.com

OPTIMIZATION

RL at Ramp

Ask a question, an agent executes tools over a spreadsheet to find the answer.

- Well thought out harness design
- Synthetic eval set was generated
- More accurate than Haiku

Eval leaderboard

Model	Accuracy	Time vs Haiku
Qwen3.5-35B-A3B (Fast Ask)	66.25%	1.05×
Claude Opus 4.6	61.88%	1.44×
Claude Sonnet 4.6	59.38%	1.60×
Qwen3.5-35B-A3B (base)	56.25%	1.34×
Claude Haiku 4.5	51.88%	1.00×

Accuracy: exact task completion. Time: mean elapsed time per rollout, relative to Haiku 4.5. All models evaluated in non-thinking mode.

john@ouachitalabs.com

Thanks!

If your team is

- Needing help building evals
- Wanting a bit of guidance on a new project
- Looking into training something

Send me a message →

EMAIL

Send me a message!



john@ouachitalabs.com

john@ouachitalabs.com

Appendix

01	Flower art	https://surya.website/rling-qwen-to-paint-with-code
02	Terminal-Bench 4.0	https://www.anthropic.com/claude-opus-5-5
03	Anthropic - Demystifying Evals	https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents
04	GEPA	https://arxiv.org/html/2507.19457v2
05	HotpotQA	https://arxiv.org/pdf/1809.09600
06	IFBench	https://github.com/allenai/IFBench
07	Ramp RL	https://labs.ramp.com/research/prime-rl-post-training/
08	Information retrieval evaluation	https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-in-information-retrieval-1.html
09	Automating eval design and hillclimbing	https://claude.dev/blog/automating-eval-design-and-hillclimbing/
10	claude-api skill	https://github.com/anthropics/skills/tree/main/skills/claude-api
11	Test-driven development	https://martinfowler.com/bliki/TestDrivenDevelopment.html

john@ouachitalabs.com