



From Data to Discovery

Automating the model adaptation loop

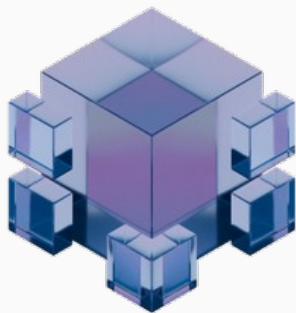
Carson Rodrigues · Adaption Labs Ambassador

1. Why the dataset is the bottleneck
2. What Adaption automates
3. What I measured
4. Demo
5. Your turn

1

Why the dataset
is the bottleneck

Fine-tuning is one line. The research process is not.



The command

```
adaption datasets adapt  
--dataset seed.jsonl  
--base llama-3.3-70b
```

It genuinely is one line. Everything
on the right is the part nobody funds.

Collect

Find it, scrape it, or pay
for it.

Clean

Dedupe, decontaminate,
drop the broken rows.

Label

Or hire people to. This is
the term that
disappears.

Generate

Prompts, reasoning
traces, preference pairs.

Train

The one line. Minutes.

Evaluate

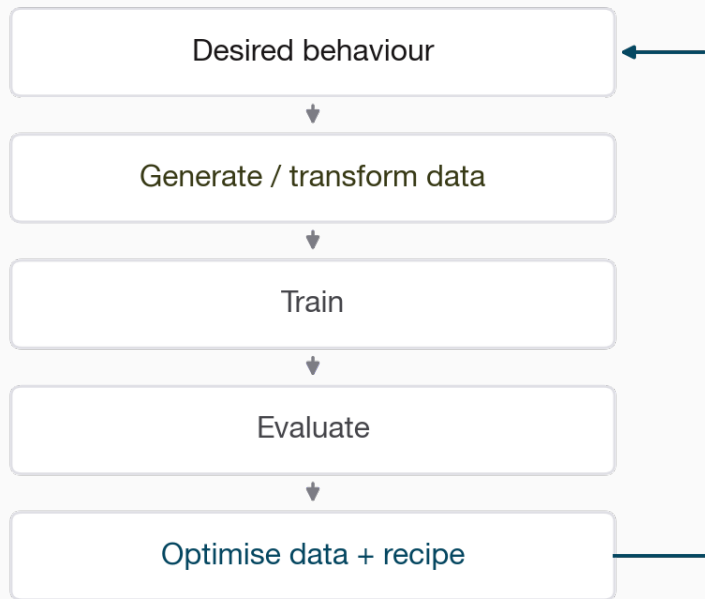
Then go back and
change the data. Again.

The dataset stops being an input.

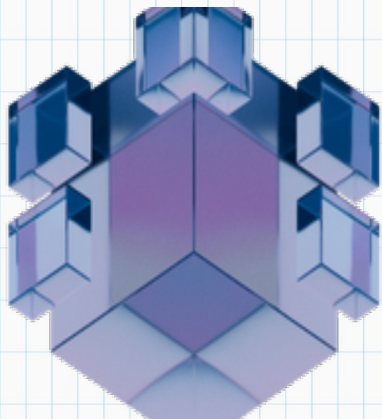
“Traditional ML treats the dataset as an input to the research process. Adaption asks whether the data itself can sit inside the optimisation loop.”

The thesis of this talk.

The adaptation loop

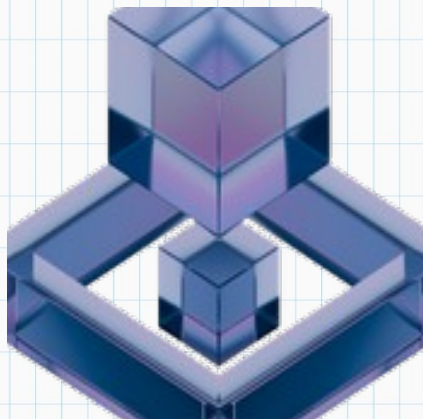


Three components, one question each



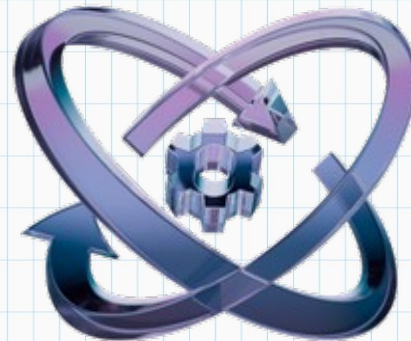
Adaptive Data

You have data. Transform, augment and re-specify the rows you already hold into training data.



Invent a dataset

You have none. Start from a description of the behaviour you want and generate the corpus.



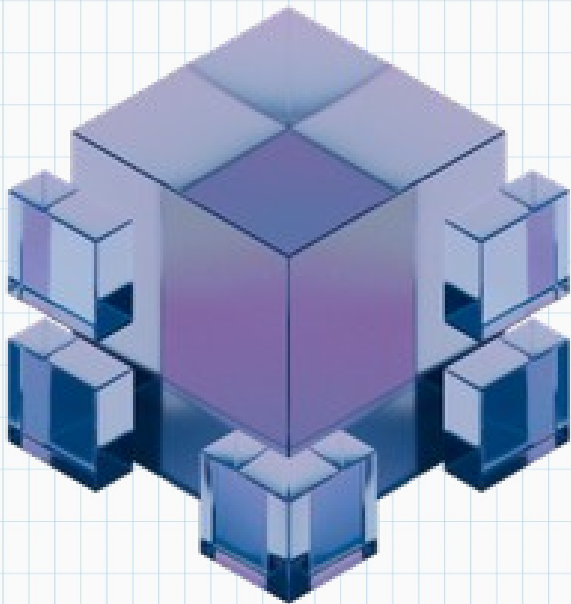
AutoScientist

You have both. Train, evaluate, revise the recipe, train again — a search, not a run.

2

What Adaption
automates

Adaptive Data

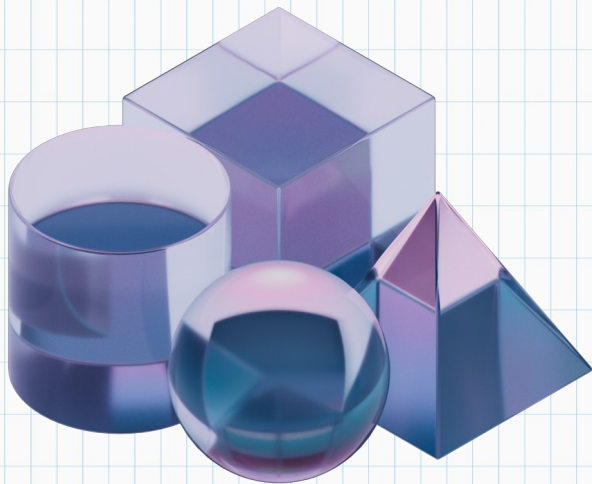


Your dataset is not your training dataset

Column mapping, deduplication, rephrasing, reasoning traces, style and safety controls, web grounding, language expansion, preference pairs.

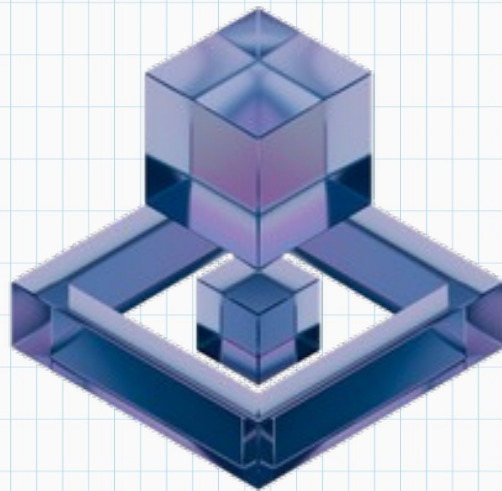
Each of those is normally a hand-made, undocumented decision. Here it is a parameter — which means you can ablate it.

Invent a dataset — from collection to specification



What you do now

Find papers, scrape, annotate, hire people, clean, format. Roughly a term.



What Invent proposes

Specify the behaviour you want, generate the corpus, evaluate what came back.

“Synthetic data is not automatically good data. What it really does is trade one research question — where do I get data — for a better one: how do I measure the quality of data I generated myself?”

Two Invent runs, 1,000 rows each, against a QA gate I wrote first.



Invent, measured

588

of 1,000 rows delivered – and billed at 1,000

Their own answer, published 1 Oct – next slide

Their answer, published yesterday — Invent vs the frontier APIs

19–55%

more diverse than every API tested —
Anthropic, Google, OpenAI, DeepSeek, Zai
— at 200 to 20,000 samples

+17%

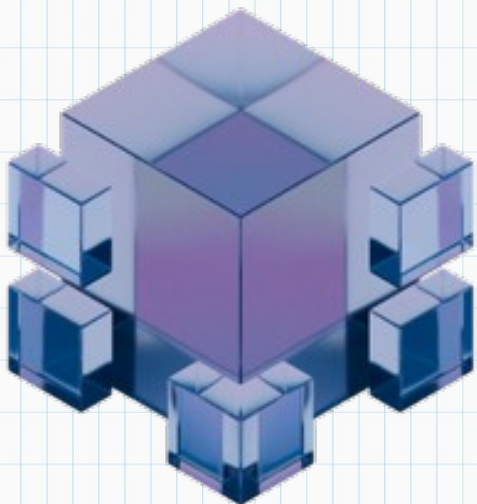
higher quality than the best baselines,
Claude Opus 5 and GLM-5.3, with no
diversity trade-off

54%

head-to-head win rate for a model
trained on it — six times the nearest
baseline

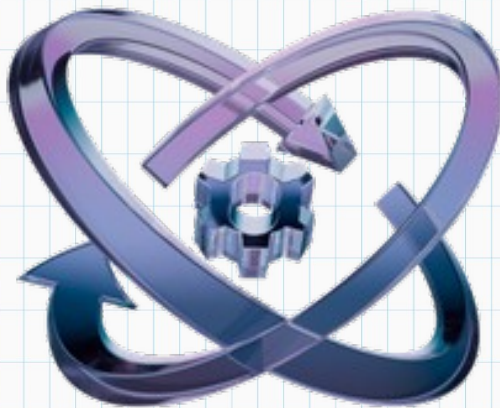
Adaption Labs, 1 Oct 2026 · their product, their benchmark

AutoScientist — what is actually being optimised



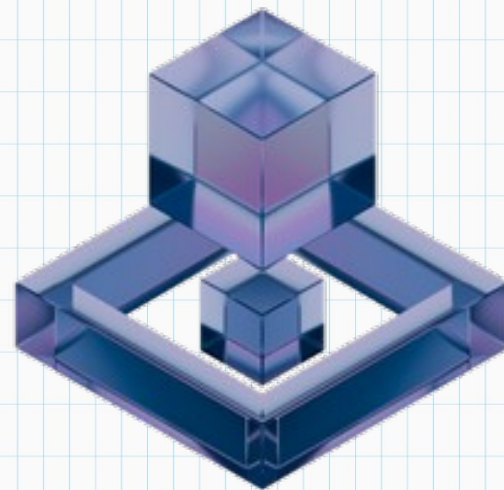
Data

Quantity, domain augmentation, instruction form, preference pairs, transformations.



Training

Learning rate, epochs, batch, LoRA or full fine-tune, base model.



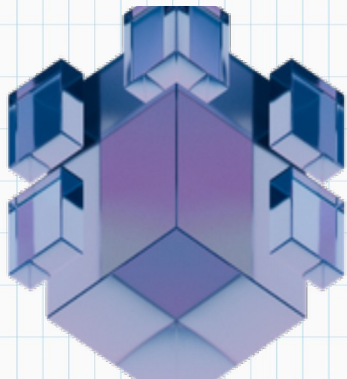
Evaluation

Base vs adapted, judge preference, held-out domain, task vs domain split.

3

What I
measured

Finding 1 — the data knob moves more than the recipe knob



+14

mean gain in held-out domain win rate, over 13 seeds

Paired design

Same seed and recipe.
Only the augmentation
varies.

13/13

seeds positive

0.0002

two-sided sign test

“I ran one matched pair. Expanding 1,882 rows to 21,882 dropped the held-out score from 76 to 72, and I wrote in a public model card that the platform overstated its own augmentation benefit.

Run properly — paired within seed, 13 seeds — it was 13 of 13 positive. I withdrew the claim in the same document that made it.”

176

fine-tuning jobs it took to overturn one run

The same job, scored two ways.

Every base under 10B improved on its own data while degrading on the held-out domain. The dashboard shows the first number; you are judged on the second.

Five-base controlled ablation, eval-loss curves read from every adapter.

Did the model improve?

On your own data



On the held-out domain

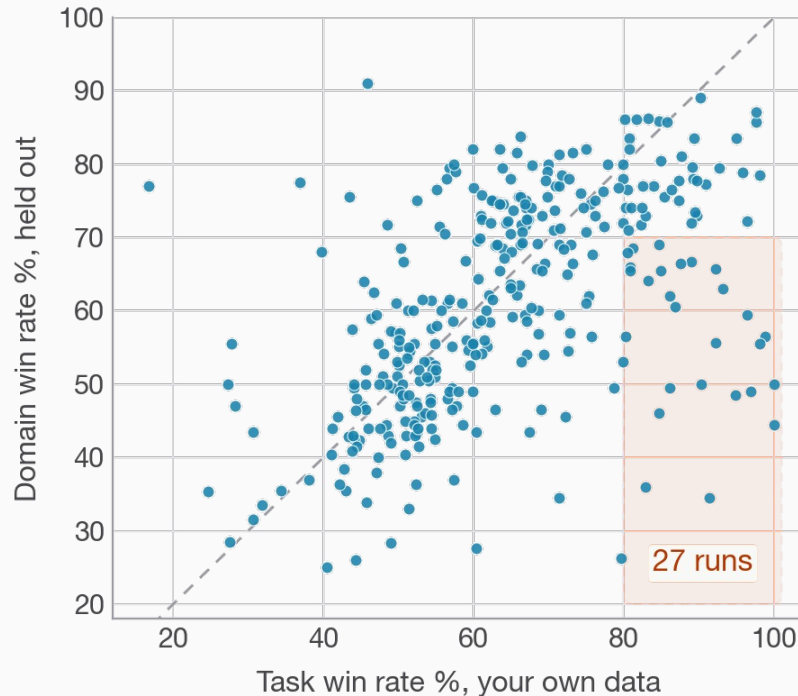


Not four models — all 333 runs.

$r = 0.53$ between the two metrics across every public submission. One run in twelve scores 80 or better on its own task and under 70 held out.

Public AutoScientist leaderboard, the full population on 16 Sep 2026.

Task win vs held-out win



What that costs you, in points

2.77

replicate SD over identical configs. My first estimate, 6.12, was withdrawn.

≈ 8

points a single-run A/B has to clear before you can call it.

+2.4

all that best-of-3 buys you. The recipe mean is the lever, not more draws.

4

Demo
15 minutes

Three screens, no waiting



Nothing trains live

Everything here is pre-run, with a recorded fallback on the machine.

A recording is fine. Four minutes of a progress bar is not.

01 The cost

`estimate=True` returns the price before a single row exists.

02 The rows

Generated output, unedited. Then I ask whether you would publish it.

03 A finished run

Iterations, both evals, the adapter — including the iteration that did not help.

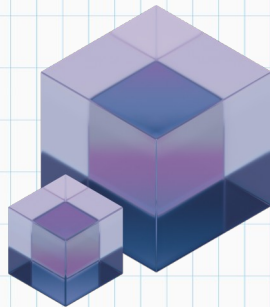
5

Your turn
15 minutes

Write the spec for your own problem

Ten minutes, on paper

Take the behaviour you actually need from a model and write it as a specification rather than a dataset hunt. Three of you read yours out; I will tell you what Invent would do with it, and where it would fail.



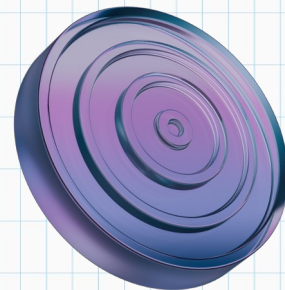
1 · The behaviour

One sentence. What should the model do that it does not do now?



2 · The judge

How would you know it worked? Name the held-out set before you build.



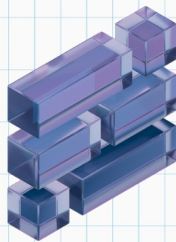
3 · The failure

What would a plausible-but-wrong answer look like? That is your gate.

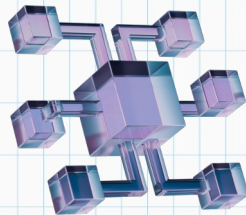
Four projects you could actually run

And credits to run them

Each of these is one pre-registration and a handful of arms away from a result. I have platform credits from Adaption — come and take some.



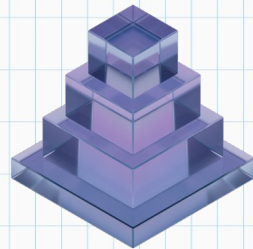
Replication
does 1 Oct hold up?



Recipe ablations
which step matters?



Judge reliability
does it track humans?



Domain transfer
what breaks
elsewhere?



Thank you

carsonrodrigues.com · adaptionlabs.ai