
NINJA LABZ • OUR OWN CYBER MODEL

Ninja Cortex, plainly.

A security brain that reads from a library instead of remembering things. Five short stories about how it works.

12M

facts in the library

184 / 200

hand-written questions right, live

0

queries written by the model, ever

THE WHOLE IDEA IN ONE LINE

The facts stay in the library. The model only reads.

Most chatbots answer from memory, and memory is where the making-things-up happens. Ours is not allowed to remember. Every answer is read, that moment, from a graph we refresh every night.

STORY ONE • THE GRAPH

The library of twelve million index cards

card

This gang hit that
hospital on this date.

a ransomware victim, from the leak sites

card

This bug is in that product
and is being exploited now.

a CVE, its vendor, the KEV entry

card

This malware uses that
trick to stay hidden.

a family and its ATT&CK technique

Every card is linked to the others, so a question about one gang can walk to its victims, their sectors, the bugs it used and the tools it shares with other crews. The library is rebuilt from the real feeds every night.

The librarian never answers from memory

1 · you ask

“Tell me
about LockBit.”

2 · it walks
to the shelf

Pulls every card
about LockBit,
nothing else.

3 · the dossier

Lays them on the
desk as one page
of plain facts.

4 · it answers

Only from
what is on the
desk. Quoted,
not paraphrased.

If the card is not there

it says “the dossier does not say”. That is
a correct answer and we score it as one.

Why it does not make things up

the facts never lived in its head. You cannot
misremember a thing you only ever read off a page.

Every question follows the same recipe, one way, no loops

1

Who is this about?

Find the names in the question. Carry the subject from the last turn if you said “they”.

2

Fetch the dossier

The one page of cards for that entity, from the live graph.

3

Pick from our shelf of queries

Counting or comparing across the whole library? Choose one we wrote and tested. Never write a new one.

4

Write the answer

From the dossier and the query results only, naming the lines it used.

5

Second look

Run the names through our maths: tempo, signals, persistence. Say what stands out.

The model never writes a database query. It chooses from ours.

A directed acyclic graph of steps: each one feeds the next, nothing runs backwards, nothing improvises. That is what keeps a chatbot from wandering into fiction.

“Which sectors does LockBit target?”

- 1 **Resolved:** LockBit (a ransomware group).
“target” is a verb here, not the retailer.
- 2 **Dossier:** LockBit’s page pulled from the live graph.
- 3 **Query chosen from our shelf:** sectors_ranked,
crew = LockBit. Counts every victim card by sector.
- 4 **Answer, under a second:** Manufacturing, Professional Services, Technology, Other, Healthcare, Retail, Financial Services, Education, Government, Transportation.
- 5 **Second look:** the leak-site tempo for LockBit against its own expected rate, flagged if unusual.

what the query returned

sector	victims
Manufacturing	547
Professional Services	516
Technology	349
Other	201

Real rows from the live graph, 27 October 2026.
The numbers are on the screen because the model was given them, not because it remembered them.

STORY FOUR • THE ADAPTER AND THE WEIGHTS

A brilliant graduate and a pad of sticky notes

the weights • the brain

14,000,000,000

tiny dials that together know English and a great deal of general knowledge. The base model, Qwen 2.5 at 14 billion parameters.

We do not touch a single one.

the adapter •
the sticky notes

80,311

worked examples: here is a dossier, here is a question, here is the right answer. A thin extra layer of dials learns the habit. **Read the file. Quote it. Say when it does not say.**

the training • one shift

6.5 h

one pass over the examples on a rented 96 GB graphics card. About twenty dollars. We can rerun it tonight with a new family of examples and mark it in the morning.

The notes are tiny next to the brain, but they change how it behaves. That is LoRA, a low-rank adapter: all the general ability, none of the fine-tuning bill, and the base can be swapped for a bigger one when it is worth it.

STORY FOUR, CONTINUED • SERVING IT

Then we shrink it to fit on one rented card

as trained

28 GB

brain plus sticky notes, merged
into one model at 16-bit precision

as served

12.1 GB

every dial rounded to 6 bits.
Built on our own CPU box
in 13 minutes. Same score.

where it runs

1 card

a rented 32 GB graphics
card, 72 cents an hour,
pulled or re-rented at will

Under a second a question

0.93 seconds on average through
the live path, two questions at once.

If the card goes away

a smaller 3B sibling, trained
the same way, is always on,
on our own box, for free.

A switch, not a migration

3B, 14B or a 120B showcase: one toggle
on the console moves the whole stack.

STORY FIVE • THE MIXTURE OF EXPERTS

It is not one model. It is a newsroom with four desks.

desk 1 • the 3B

The junior, always in

On our own box, free,
trained the same way.
168 of 200. The fallback
that never leaves.

desk 2 • the 14B

The main reporter

On the rented card.
184 of 200 live. Reads
the dossier, picks the
queries, writes the answer.

desk 3 • the 120B

The star columnist

Switched on for a
showcase, off the rest
of the time. 195 of 200,
untrained, on a 96 GB card.

desk 4 • Claude

The prose desk

Writes the morning briefing
from our facts, compares
two gangs, gives the
analyst's take, does the
ScottG voice. Never hands
over a number of its own.

the desk
editor

**A set of rules,
not a model**

Facts go to our reporters, who read our library. Composition and explanation go to Claude, with those facts as the only material. Every line on screen says which desk wrote it, and the reader switch picks which of our own desks is on shift.

HOW YOU KNOW IT IS HONEST

Every answer comes with its receipt

on every card in the console

The exact lines it read, highlighted
in the dossier as you hover

The queries it ran, with their rows, as the source

Which desk wrote it and how long it took

A second look from our own
maths, kept apart from the answer

A thumbs up or down that lands in
the graph and becomes the next test

We mark it like a junior analyst

Two hundred hand-written questions whose answers we know, plus a hundred for the vulnerability role and eighty-eight for attack techniques. Scored by a script, not by us.

The fails are published too

If a training run makes it worse, or makes it refuse too often, it does not go live, and the result goes on the public status page anyway.

Cast iron or it does not ship.

THE SCORECARD • THE SAME 200 QUESTIONS, EVERY MODEL

Size buys fifteen points. Reading buys the rest.

14B, from memory alone

86

the base model answering
with no dossier.
Confident, often wrong.

3B, reading,
our CPU box

168

the free fallback, trained
the same way, always on.

14B, reading,
live today

184

through the production
path, under a second,
zero refusals of
answerable questions.

120B, reading,
showcase

195

untrained, on demand,
ten points away for
the 14B to chase.

What moved the number

Giving the model the dossier: 86 to 168 on the same questions.
Then a bigger base on exactly the same training set: 168 to 184.

What did not

Stacking clever training families. Two that each
passed alone regressed together, and refusals
went to 49 percent. One family at a time now.

THE FIVE STORIES, ONE MORE TIME

A library it reads. A recipe it follows. Notes it was taught. A newsroom that signs its work. A receipt on every answer.

Two people, our own graph, our own model, rented cards by the hour, and a rule we do not bend: cast iron or it does not ship.

ninjalabz.io
status and the public log at skunkworkz.io