DAY 46–55 · Part 5 · Looking Inside

Part 5 Review — Looking Inside

From one dot to the phone in your pocket · short, with analogies · DAY 46-55

No test — a review to read through

One dot can only multiply and add

1. Open up an AI and you find dots passing numbers around, stacked in layers. That stack is called a neural network.

2. All one dot does is **multiply, add, and check a threshold** — that's the whole job.

3. The smartness isn't in any single dot. It builds up as numbers pass through layer after layer.

Attention is a share you split up

1. Attention is the math that decides how hard one word in a sentence stares at the other words.

2. The staring doesn't stretch forever — the words **split up a fixed share** of it.

3. Stare hard at one word and the rest get weaker, and where the staring piles up becomes the meaning.

Meaning gets turned into a place

1. A computer can't read letters, only numbers. So each word gets stored as a bundle of numbers.

2. Words used in similar company learn similar numbers, so **words with close meanings land in close places**.

3. Once meaning is a place, near and far become something you can calculate.

It measures meaning, not letters

1. Whole sentences can become number bundles too. Then you can measure the distance between two of them.

2. Even with no letters in common, **same meaning means a close place**.

3. And when the letters are nearly identical but the meaning flips, they move apart - like "that was fun" and "that was no fun".

Look it up and read it, instead of reciting

1. Answering from memory alone makes it easy to invent something plausible when you don't know.

2. So before answering, it **goes and finds the relevant text first** and lays it open beside the question.

3. That turns a closed-book test into an open-book one. And it can say where the words came from.

Nudge a grown model instead of raising a new one

1. You can take a base model that already did the big round of learning and teach it one more job. That's called fine-tuning.

2. It isn't learning from scratch — only **a tiny slice of the numbers inside** gets changed.

3. It's strong at changing tone and format; for new facts you use answering with the book open instead.

Not the right answer, just which one is better

1. You get two answers to the same question and have a person pick the better one, tens of thousands of times over.

2. What you collect isn't right-and-wrong, it's **what people prefer**.

3. Then you build a small scoring model that copies that preference, and tune the AI toward its scores.

The shape of the math was the same all along

1. A graphics card is a part built to do easy math a huge number of times at once.

2. AI training has exactly that shape, which is how it became the heart — one card holds about **80 billion** tiny switches.

3. Even that wasn't enough: Llama 3 was built on 16,000 cards tied into one block, running for 54 days.

One is tiny. the pile is not

1. One text question came in at a median of **0.24 watt-hours** - about a TV left on for 9 seconds.

2. The catch is that questions like that keep piling up, nonstop.

3. Data centers worldwide used 415 terawatt-hours of electricity in 2024, about 1.5% of all electricity.

It splits the work into pieces

1. Today's phones carry a small model of about 3 billion numbers, so they can summarize and polish even with no internet.

2. Long text or hard questions get sent out to **that building from before** and the answer comes back.

3. Whatever it can finish inside, it finishes inside - faster, less electricity, and your words never leave.

Open it up and there's no magic. Simple math, piled up enormously.



Part 6 looks at what comes next.