DAY 46–55 · Part 5 · Looking Inside

Part 5 Review — Looking Inside

From one dot to the phone in your pocket · one layer deeper · DAY 46-55

No test — a review to read through

One dot can only multiply and add

1. A neural network is dots passing numbers around, stacked in layers.

2. One dot **multiplies each incoming number by a weight and adds them up**, then just checks whether the total clears a threshold.

3. Clear it and the signal goes to the next layer; miss it and it stops there.

Why stacking layers changes things

1. The first layer only catches small pieces, like lines and dots.

2. The next layer builds eyes and ears from those pieces, and the last layer says cat — that's what **deep** means.

3. Without the threshold, stacking any number of layers collapses back into one multiply-and-add.

Attention is a share you split up

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

2. The raw scores aren't used as they are; they become **shares adding up to 1** and get split around.

3. So staring hard at one word automatically weakens all the rest.

The staring decides the meaning

1. Who "it" refers to gets settled when the staring piles onto the word that fits.

2. The big shift was making a sentence **look at itself all at once** instead of reading front to back in one line.

3. That 2017 method became the base of almost every model in use today.

Meaning gets turned into a place

1. A computer can't read letters, so each word is stored as a bundle of several numbers.

2. Nobody hands it those numbers — it works them out **by itself while reading an enormous amount of text**.

3. There is one rule: words that show up in similar company get similar numbers.

Once it's a place, it's math

1. Once meaning is a place, near and far can be measured as numbers.

2. By letters, apple and orange have nothing in common; by place they sit **side by side**.

3. That's the base for finding similar things, which is the next day.

It measures meaning, not letters

1. Sentences, like words, can become number bundles. Then a distance between two of them falls out.

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

3. And nearly identical letters with the opposite meaning move apart.

Why counting letters doesn't work

1. "That was fun" and "that was no fun" share almost every letter but mean the opposite.

2. Measure by letter overlap alone and you would score those two as **very close**, which is wrong.

3. That's why finding text moved from letter matching to places of meaning.

Look it up and read it, instead of reciting

1. Answering from memory alone makes inventing a plausible answer easy.

2. So the question becomes a number bundle and **the text pieces sitting in close places** get pulled out first.

3. Then it reads those beside the question and answers.

What's good about it, what's weak

1. You don't have to retrain the model — **swap out only the text you lay open** and the answers follow.

2. It can show where the words came from, which makes checking easy.

3. But if the text it found is wrong, the answer is wrong with it. Finding is half the job.

Nudge a grown model instead of raising a new one

1. Teaching one more job on top of a base model that already did the big round of learning is called fine-tuning.

2. The numbers inside stay almost untouched; only **a tiny slice** changes.

3. A method called LoRA cut the numbers it touches down to one ten-thousandth.

Which one should you use

1. Teaching it extra is strong at **tone and format**.

2. If you want new facts in there, answering with the book open is the right fit.

3. The big round takes months; the tuning step is far shorter and cheaper by comparison.

Not the right answer, just which one is better

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

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

3. Scoring every answer by hand would never end, so the picks are used to build a scoring model first.

How big was the effect

1. From then on that scoring model hands out the points, and the AI gets tuned toward them.

2. In a study released in 2022, people preferred this tuned small model's answers over those of a model **more than 130 times its size**.

3. It's a rare case where tuning mattered more than size.

The shape of the math was the same all along

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

2. Painting millions of dots on a screen at the same time needs **thousands of simple workers, not a few clever ones**.

3. AI training is the same: not one hard problem, but trillions of easy multiplications at once.

By the numbers

1. One card of the kind used for training holds about **80 billion** tiny switches.

2. That single card does roughly 1,000 trillion calculations a second.

3. Even that wasn't enough: Llama 3 ran 16,000 cards as one block for 54 days.

One is tiny, the pile is not

1. In a measurement released in 2025, one text question came in at a median of **0.24 watt-hours**.

2. That's about a TV left on for 9 seconds, so one question alone is nothing to worry about.

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

Looking at the whole building

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

2. The International Energy Agency expects 945 terawatt-hours by 2030, more than double.

3. 1.5% looks small, but things that double in six years are not common.

It splits the work into pieces

1. Today's phones carry a small model of about **3 billion numbers**.

2. Even with no internet it can summarize, polish text and tidy up notifications.

3. The model on the phone and the one on the server differ in size by tens to hundreds of times.

Why split it at all

1. The phone tries first and sends out only the pieces that are too heavy for it.

2. Finishing inside means **faster, less electricity, and your words never leaving**.

3. Long text and hard questions still have to travel to that building from before.

Inside wasn't magic, it was math.
Simple work, piled up enormously.



Part 6 looks at what comes next.