DAY 09 · Part 1 · What AI Is

# Is a bigger model always better? — Big models, small models

Five-Minute AI Lessons · Sep 15, 2026

There's no one best size — it depends on the job

# In a nutshell

1. A **big** model is one with lots of parameters, which you met on DAY 05 — it knows a lot and holds up on hard problems it's never seen.

2. But big models are **slow and expensive**, and they won't even fit on your phone. They only run on servers across the internet.

3. So sometimes what a big model learned gets **poured into a small model**. The job decides the right size.

# Same job, but they live in different places

## Small model

Runs on your own device

Fast · cheap · works without internet

Knows a narrower range

Fixing typos · reading text in photos · instant game replies

**Passing on what it learned**

(distillation)

## Big model

Runs on servers (data centers)

Knows a lot · holds up on hard problems

Slow · expensive · needs internet

Reading long text closely · solving new problems step by step

**Not better or worse, just different jobs — the job decides the size**

Try it

# Which one gets this job?

## Small model

### Better for

- ✓ **Reading text in a photo in airplane mode**
- ✓ **Fixing typos before sending a text**
- ✓ **A game NPC answering within 1 second**
- ✗ Solving a new competition problem

Fast, cheap, and runs without internet

## Big model

### Better for

- ✓ **Solving a new problem step by step**
- ✓ **Reading 30 pages and questioning it**
- ✓ **Comparing two countries' systems**
- ✗ Fixing a single typo

Knows a lot, but slow and expensive

## Distillation

### Big model → small model

- ✓ **The big model first makes tons of answers**
- ✓ **The small model learns from those answers**
- ✓ **The small model beats its own size**
- ✗ But it still can't fully replace the big one

Formalized by Hinton's team in 2015

Big isn't what's good — the right size for the job is.

Source: Hinton, Vinyals, Dean, Distilling the Knowledge in a Neural Network (2015) · arxiv.org/abs/1503.02531

Quick check

# Just three questions

**Q1**

## When a model is “big,” what does it mostly have a lot of?

① Letters shown on the screen
② Parameters (numbers tuned during training)
③ People using it
④ Time it takes to answer

**Q2**

## What does a small model clearly do better than a big one?

① Solves hard new problems better
② Knows more facts
③ It's fast, cheap and runs offline on your device
④ Never says anything wrong

**Q3**

**(True / False)** There's a way to pour what a big model learned into a small model.

# Know the why, too

Q1

② Size means the number of parameters — the number you got a feel for on DAY 05. The more there are, the more it can hold.

Q2

③ A small model fits whole on your device, answers instantly, and uses less power and money. It doesn't win on hard-problem skill — it wins on where it lives.

Q3

T It's called distillation. Teach a small model using a big model's answers as its textbook, and it does far better than its size suggests.

# Big isn't what's good — the right size for the job is.

Tomorrow: Open and closed models — AI anyone can download and use, and AI you can't.