The three numbers that shape every AI interaction

Parameters, tokens, and context window — the dials behind every model.

Three questions, one shape of answer.

Each points at a different dial.

Why did it forget what I told it 10 messages ago?

DIAL 03 — CONTEXT WINDOW

Why is the bigger model slower, and not obviously smarter?

DIAL 01 — PARAMETERS

Why can't it count the r's in "strawberry"?

DIAL 02 — TOKENS

Three dials.

Every model is a machine with three settings. Learn them and most of what looks mysterious stops being mysterious.

DIAL 01

PARAMETERS

The size of the brain

DIAL 02

TOKENS

The unit of thought

DIAL 03

CONTEXT

The working memory

DIAL 01

Parameters — the size of the brain.

~1T

parameters at the frontier

Frontier models have hundreds of billions of them — the numerical weights the model learned during training.

WHAT TO KNOW

- Bigger isn't monotonically better.
- Diminishing returns past a point.
- Better recipes beat more parameters.
- Affects cost & speed > smartness.

A 2026 70B model matches a 2024 500B model.

Parameters got smaller. The recipe improved.

2023	500B+	*frontier flagships raced on parameters*
2024	400B	*post-training techniques maturing*
2025	200B	*distillation and data quality gains*
2026	70B	*recipe beats raw scale at many tasks*

Pick by task fit, not by parameter count.

DIAL 02

Tokens — the unit of thought.

~4

characters per token, typically

Tokens are the chunks the model actually sees — statistically common slices of text. Not letters. Not words.

WHAT TO KNOW

- Model sees tokens as atomic.
- Can't introspect characters inside.
- Tokenization varies by language.
- You pay per input + output token.

The strawberry problem.

Why trivial-for-humans is sometimes hard-for-models.

strawberry

straw **berry**

Two tokens. The model never sees the letters inside.

Character-level tasks? Add tools. Don't push prompts harder.

DIAL 03

Context window — the working memory.

200k+

tokens per inference call

Shared between input and output. Stateless between calls. Everything that feels like "remembering" is the app re-sending history.

WHAT TO KNOW

- Input + output share the budget.
- Stateless between calls.
- "Lost in the middle" is real.
- Bigger window isn't better memory.

What fits in 200,000 tokens?

Concrete equivalents. Make the abstract budget tangible.

150k

English words

a full non-fiction book

~300

novel pages

a medium-length novel

1

medium codebase

a mid-sized repo, start to finish

1

month of Slack

a busy channel's month of history

Sounds infinite. It isn't. And attention quality across that window is a separate problem.

The three dials, side by side.

DIAL 01

PARAMETERS

Size of the brain

~1T

at the frontier

Cost and speed
more than capability.

DIAL 02

TOKENS

Unit of thought

~4

chars per token

Why it can't
count letters.

DIAL 03

CONTEXT

Working memory

200k+

tokens per call

Why it "forgot"
what you said.

Three dials → three operational questions.

Most AI decisions collapse into one of these.

PARAMETERS

Which model?

Pick by task fit, not by size. Mid-tier frontier is usually best value.

TOKENS

How do I prompt it?

Don't fight tokens. If the task is character-level, add tools.

CONTEXT

How to structure context?

Window sets ceiling. Attention sets floor. Retrieval often beats stuffing.

THE CATCH

Benchmarks mislead.

A benchmark is a weighted sum of all three dials plus a fourth — inference efficiency.

Different labs weight the dials differently. Which means two models with near-identical scores can feel radically different in production — because the dial that dominated the benchmark isn't the dial that dominates your workload.

Build a private benchmark. Five or six tasks that matter to you.

Three dials, plus a quiet fourth.

Inference efficiency — how fast and how cheaply a model runs.

DIAL 01

PARAMETERS

Brain size

Some labs race here

DIAL 02

TOKENS

Unit of thought

Mostly stable

DIAL 03

CONTEXT

Working memory

Some labs race here

DIAL 04

EFFICIENCY

Cost & speed

Open-weight races here

"Model X beats model Y" always means: on these dials, at this weighting, for this task.

Takeaways

01

Three dials explain most of AI's quirks.

Parameters (size of brain). Tokens (unit of thought). Context window (working memory).

02

Each dial maps to one operational question.

Which model? How to prompt? How to structure context?

03

Benchmarks mislead. Build a private one.

Five or six tasks that matter to you, run across every model you're considering.