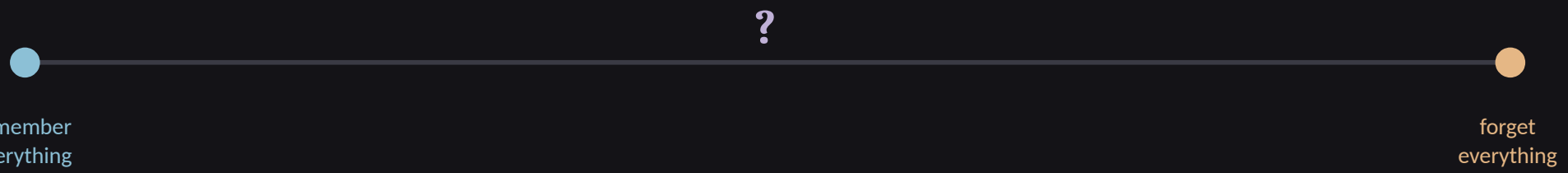
UNIT 9

Overfitting, Memorization, and the Blurry Archive

Borges, “The Library of Babel” and “Funes the Memorious”

Ted Chiang, “ChatGPT Is a Blurry JPEG of the Web”



Germany, 2013. A photocopier. A floor plan with three rooms.

THE ORIGINAL

14.13 m²

21.11 m²

17.42 m²

THE PHOTOCOPY

14.13 m²

14.13 m²

14.13 m²

The copier's compression decided the three numerals were similar enough to store once and reuse.

Observation

A blurry copy announces itself. You can see that it is blurry, so you go and find the original. This copy was crisp, legible, confidently typeset — and wrong. Chiang's word for the danger is not blur. It is that the compression produced "numbers that were readable but incorrect."

KEY QUESTION

Is intelligence what remains after you throw the details away?

Borges answers it twice and gets two different catastrophes. Chiang answers it once and is uneasy about his own answer. The machine-learning literature answers it every day without noticing it is a question.

HOW THIS UNIT IS BUILT

One dial, four positions on it

Parts One · the mechanism

Overfitting and generalization from zero, and the arithmetic that says whether a model could memorise its corpus even if it wanted to. The answer has changed since 2020 and almost nobody has noticed.

Part Two · Chiang

The blurry-JPEG argument, taken seriously and then tested against what is actually known about memorization. The metaphor is excellent and partly wrong, and the partly-wrong is the interesting half.

Parts Three and Four · Borges

The Library: total information, zero knowledge. Funes: total memory, no thought. Two ends of the dial, written in the 1940s, both of them failures of exactly the thing this unit is about.

The claim to test

That forgetting is not a defect of learning but its mechanism — that there is no version of knowing which keeps everything, and that the design question is never whether to discard but how much and which. If that is right, then every archive, every model and every mind is a particular answer to one question, and they can be compared.

PART ONE

The generalization gap

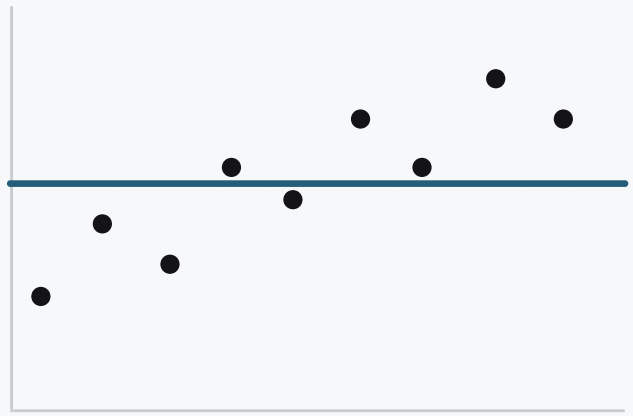
Overfitting and memorization, from zero



One picture, one curve, and one multiplication that settles an argument people are still having.

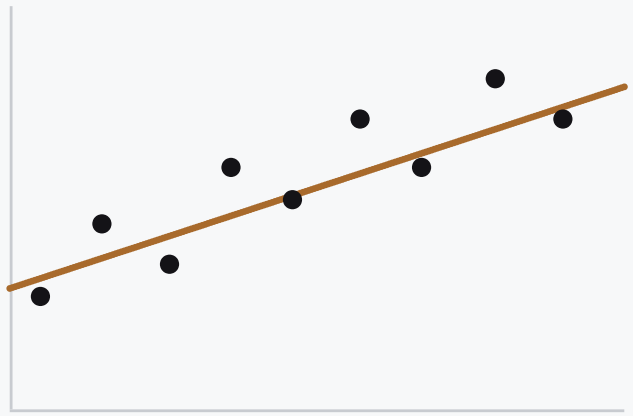
THE WHOLE PROBLEM IN ONE PICTURE

Nine data points, three ways to draw a line through them



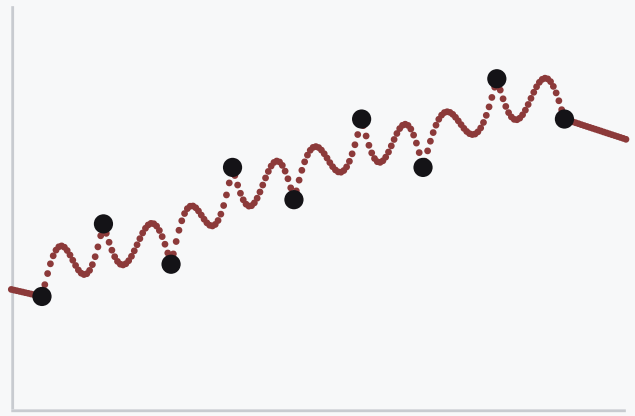
Underfitting

A model too simple to capture the pattern. Wrong on the data it was given, and wrong on everything else. Easy to spot, easy to fix.



Fitting

Captures the trend and ignores the noise. It is wrong about every single point and right about the thing the points are evidence of.



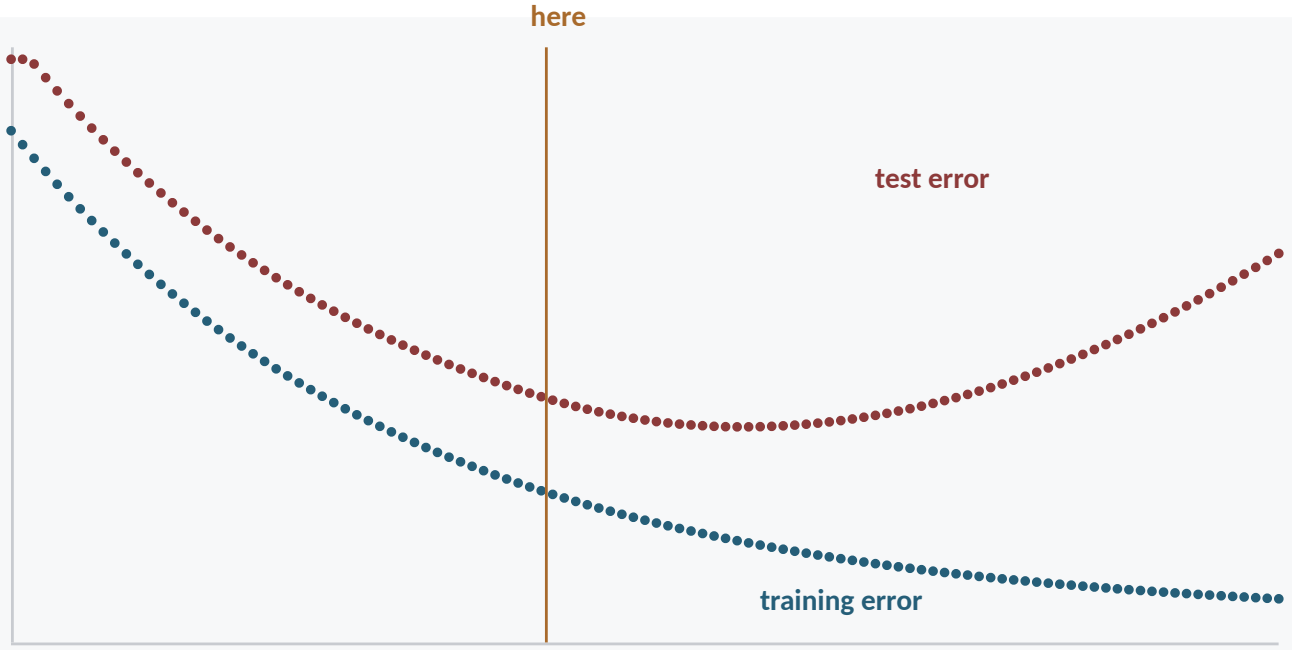
Overfitting

Passes through every point exactly. Zero error on the data it was shown. Useless anywhere else — and from the inside, indistinguishable from brilliance.

Only the middle one has learned anything. The right-hand one has the best possible score on the only data it has ever seen, and that score is precisely what tells you it has failed.

THE CURVE EVERY PRACTITIONER WATCHES

Training error falls forever. Test error does not.



The gap is the whole subject

Test error minus training error. It is the amount by which the model's confidence about itself exceeds its competence. The field has a name for everything on this slide except, oddly, that sentence.

The vertical line is a choice

Someone decides when to stop. Too early and you have underfit; too late and you have memorised. Nothing in the data tells you where the line goes — it is set by a human, with a holdout set and a judgement.

And the right-hand side is Funes

Perfect recall, rising uselessness. Keep the shape of this curve in your head for the last block of the day; we will put a short story on it.

Everything to the right of the line is a model getting better at the past and worse at the future — and the only way to see it happening is to have deliberately hidden some data from yourself in advance.

Schematic. The shape is standard; the exact curves depend on model, data and regularisation.

WHAT MEMORIZATION IS, PRECISELY

Not a metaphor — a measurable property with a definition

The informal version

A model memorises when it stores specific training examples rather than the regularities across them. Give it an example it has seen and it is perfect. Give it a near-neighbour it has not seen and it has nothing to say.

The operational version

A string is extractable if there is a prompt that makes the model emit it verbatim. This is testable: you take known training data, feed a prefix, and check whether the continuation comes back exactly.

Why this matters today

Because it converts a philosophical argument into a measurement. Whether a model is a blurry copy or a sharp one stops being a matter of taste the moment somebody runs the experiment — and somebody has.

The uncomfortable middle

A real model is not at one end of this. It is blurry in most places and photographically sharp in a few, and which few depends on what was repeated, how big the model is, and how much of a passage you feed it before asking for the rest. Hold that sentence — it is going to be the main objection to Chiang.

THE ARITHMETIC THAT SETTLES IT

Could a model store its training data, even in principle?

$$\text{bits available per token} = (16 \text{ bits} \times \text{parameters}) \div \text{training tokens}$$

Sixteen bits per parameter is the usual training precision. Divide the model's whole storage by the number of tokens it must account for.

MODEL		PARAMS	TOKENS	BITS PER TOKEN
Gopher	2021	280B	300B	**14.93**
GPT-3	2020	175B	300B	**9.33**
Chinchilla	2022	70B	1.4T	**0.80**
Llama 2 70B	2023	70B	2.0T	**0.56**
Llama 3.1 405B	2024	405B	15.6T	**0.42**

The line it crosses

English runs roughly 2 to 5 bits per token once compressed as well as anyone knows how.

Above that line, storing the corpus is physically possible. Below it, it is not — by a factor of ten.

Somewhere between 2020 and 2022 the field crossed a line and stopped noticing. Today's models do not have room to remember their training data. Generalization is not a virtue they were taught — it is the only option left once the arithmetic is done.

Parameter and token counts as publicly reported. Entropy range per token is approximate and depends on tokenizer; the order-of-magnitude conclusion does not.

NEXT

Chiang

The most famous metaphor in the argument, taken seriously and then tested

PART TWO

A blurry JPEG of the web

Chiang, taken seriously and then tested



The argument first, in its strongest form; then what the measurements do to it.

CHIANG'S ARGUMENT

Four steps, and each one is doing work

1

Compression is the frame

A language model is a lossy compression of its training corpus. Chiang asks us to imagine the web's text squeezed at a hundred to one by an algorithm that keeps statistical regularities rather than strings.

2

So output is interpolation

Ask it something and it reconstructs an answer from the neighbourhood of what it kept. That is what interpolation means, and it is the same operation whether the result is a useful paraphrase or an invention.

3

So hallucinations are artifacts

Not a separate bug to be patched. The same mechanism that produces fluent summary produces fluent fabrication, and Chiang doubts the two can be separated at all.

4

And lossy looks smarter

His sharpest line: a lossless model would just quote web pages and look like a search engine. Because it rephrases, it resembles a student putting things in her own words — which is what understanding looks like from outside.

The conclusion, in his own words

"When we're dealing with sequences of words, lossy compression looks smarter than lossless compression." Note what that sentence concedes and what it withholds: the appearance is real, and it is still an appearance.

Ted Chiang, "ChatGPT Is a Blurry JPEG of the Web," The New Yorker, 9 February 2023.

THE TEST HE PROPOSES

Arithmetic, because arithmetic cannot be approximated

WHAT HE OBSERVED

GPT-3 handled two-digit addition well. At five digits, accuracy fell to around ten per cent — and the errors looked like a system that was not carrying the one.

“A superficial approximation of the real thing.”

WHY IT IS A GOOD TEST

Addition has no gist. You cannot be approximately right about 48,273 + 19,856 in a way that counts. Every other domain lets a blur pass as a paraphrase; arithmetic refuses, which makes it a probe for whether a rule was learned or a distribution was.

And the honest caveat, worth noting in the margin

The essay is from February 2023 and describes the models of that moment. Arithmetic in particular has moved, partly because systems now hand the sum to a calculator — which does not refute Chiang so much as sidestep him. When you cite a three-year-old essay about a moving target, say which part you are citing: the mechanism, which has held up, or the measurements, which have not.

NOW TEST THE METAPHOR

Models do memorise. It has been measured.

Size

A tenfold increase in model size raised memorization by about nineteen percentage points. Bigger models are sharper copies, not blurrier ones.

Repetition

Sequences repeated often in training are far more extractable. Deduplicating the corpus cut memorization from 3.6% to 1.2% — roughly threefold.

Context

Feed more of a passage and more of the rest comes back verbatim: 33% extractable at 50 tokens of context, 65% at 450.

Scale

For a 6-billion-parameter model the authors estimate at least 1% of its training corpus is extractable. One per cent of a corpus that size is an enormous number of exact strings.

What this does to the JPEG

A JPEG is uniformly blurry — that is what makes it a good metaphor and what makes it wrong. A model is blurry across most of its range and exact in specific places, and the exact places are not random: they are whatever the corpus repeated. So it is less a blurry photograph than a photograph with some regions at full resolution, and no legend telling you which. The compression is selective, and what it preserves is frequency.

Carlini et al., "Quantifying Memorization Across Neural Language Models" (2023), on GPT-Neo/GPT-J models trained on The Pile.

AUDITING THE METAPHOR

What the blurry JPEG earns, and what it charges

What it gets right

That output is reconstruction rather than retrieval. That fluent summary and fluent fabrication come out of one mechanism. And that a system which rephrases will read as comprehending, because rephrasing is the evidence we have always accepted for comprehension.

What it gets wrong

A JPEG degrades evenly; a model does not. A JPEG has no objective and nothing it was rewarded for keeping. And a JPEG cannot reproduce any region of the original exactly — whereas a sufficiently repeated passage comes back whole.

What the metaphor costs

“Blurry copy” implies uniform loss, which makes the copyright question look like a question about degree. “Selectively exact archive” implies something else entirely: that the model is sharpest precisely where the corpus repeated itself, and so sharpest on whatever the web said most often.

The discipline, not the verdict

Chiang’s essay is very good, which is exactly why it is worth doing this to. A metaphor that is eighty per cent right is more dangerous than one that is obviously wrong, because the twenty per cent travels unexamined inside the eighty.

NEXT

The Library of Babel

Back eighty years, to find all of it already written

PART THREE

Everything, and therefore nothing

“The Library of Babel”



One of the two ends of the dial, specified precisely enough to calculate.

BORGES GIVES A SPECIFICATION

He is unusually precise, and the precision is the joke

410	40	80	25
pages per book	lines per page	characters per line	orthographic symbols

410 × 40 × 80 = 1,312,000 character positions in every book

And the Library contains every possible arrangement of those positions. Not a selection. Every one.

So it contains every true book

The catalogue of the Library. Your biography. The proof of every theorem. The correct account of your own death, in every language, including ones not yet invented.

And every false one

Each true book accompanied by millions of near-identical copies differing in one character, plus every lie, every plausible forgery, and every text that is one typo away from the truth.

Which makes it useless

Not hard to search. Useless — in a precise sense we are about to compute. A library of everything has exactly as much information in it as a library of nothing.

Specifications as given in "The Library of Babel" (1941).

THE SIZE OF THE LIBRARY

$$25^{1312000} \text{ books}$$

1,834,098

digits long, if you wrote the number out

$\approx 10^{80}$

atoms in the observable universe, for comparison

0.76 MB

of information in a single one of the books

The number of books exceeds the number of atoms by a factor of about ten to the one-million-eight-hundred-thousandth. The Library cannot be built, searched, or sampled from. It can only be specified — which Borges does, in one sentence, and that is the point.

TOTAL INFORMATION, ZERO INFORMATION

Why a complete archive tells you nothing

THE INFORMATION-THEORY VERSION

Information is a reduction of uncertainty. Finding a book in the Library reduces your uncertainty about the world by nothing at all, because the Library's contents are fixed in advance and independent of how the world happens to be. Every claim is on a shelf somewhere, so a shelf is not evidence.

A source that emits everything emits nothing.

WHAT THE LIBRARIANS DO

— Some search for the catalogue of catalogues.

— Some hunt the Man of the Book — the librarian who has read the one volume that explains the rest.

— The Purifiers destroy books they judge worthless, which is futile: every book has near-copies.

— And some throw themselves down the ventilation shafts.

Read the librarians as a literature review

Each is a retrieval strategy for an un-indexed archive, and each fails in a way we would now recognise. The catalogue-seekers want an index. The Man of the Book is a dream of a model that compresses the whole library into one readable artefact. The Purifiers are doing data curation, badly, with no criterion. Borges has written the failure modes of information retrieval before the field named one of them.

WHERE THE LIBRARY SITS ON THE DIAL

It is the archive with nothing thrown away at all.

Zero compression, zero loss, zero generalization — and therefore zero use. The Library is what you get if you refuse, completely and on principle, to forget.

The Library

Contains every possible string. Contains every truth and is therefore not a source of any.

An overfit model

Contains its training set exactly. Right about everything it has seen and silent about everything else.

A well-fit model

Contains a smoothed account of its training set. Wrong about every particular and right about the shape.

Only the third one can be asked a question it has not already been given the answer to.

NEXT

Funes the Memorious

The other end of the dial, which is worse

PART FOUR

A man who cannot forget

“Funes the Memorious”



The Library is a thought experiment about an archive. Funes is one about a mind, and it costs more.

THE CONDITION

Thrown from a horse, paralysed, and perfected

Before

An ordinary young man in a small town, notable only for always knowing the exact time without a watch — a small detail Borges plants and never explains.

The fall

He is thrown from a horse and left paralysed. He wakes with, in the story's terms, perception and memory that are infallible. The body stops; the recording does not.

After

"I alone have more memories than all mankind has probably had since the world has been the world." He can reconstruct an entire day — and reconstructing it takes an entire day.

The detail that should bother you

A day of memory costs a day to replay. His recall is perfect and therefore has a one-to-one ratio with lived time, which means he can never catch up with his own life. Perfect memory is not a faster mind. It is a mind with no compression, running in real time, falling permanently behind.

Borges, "Funes the Memorious" (1942).

HIS NUMBER SYSTEM

What a mind looks like with no abstraction in it

Funes found the decimal system wasteful and replaced it with names. In place of seven thousand and thirteen, he would say Máximo Pérez. In place of five hundred, he would say nine.

It is not a system

A numeral system is a rule that generates names for numbers you have never met. Funes's scheme has no rule — only a table, one entry per number, extended by memorising more entries. It cannot reach a number he has not already named.

It is a lookup table

Which is exactly the data structure of a memorised model. Perfect on what it contains, undefined one step outside it. The narrator's objection is that this is "the exact opposite of a system of enumeration," and he is right on technical grounds.

And Funes cannot see the objection

He thinks he has improved on arithmetic. From the inside, a lookup table that covers everything you have encountered is indistinguishable from understanding — which is the same sentence that described the overfit curve earlier.

Borges has written the failure of abstraction as a character trait, and he has made it sound, from the inside, like a refinement.

THE SENTENCE THE WHOLE UNIT TURNS ON

> *“It bothered him that the dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front).”*

1

He is not wrong

They are different. Different photons, different angle, different second. Every argument for calling them the same dog is an argument for throwing information away, and Funes can see exactly what is being thrown.

2

A category is a decision to ignore

The word “dog” is a declaration that certain differences do not matter. Every concept you own is a deliberate blindness that someone found useful. Funes has no blindnesses and therefore no concepts.

3

Which is why he cannot think

Not because he is slow. Because thought operates on classes, and classes are made by discarding. With nothing discarded there is nothing to reason over — only an endless, perfectly accurate inventory.

“To think is to forget differences, generalize, make abstractions.” — and Borges’s narrator concludes that Funes was “almost incapable of ideas of a general, Platonic sort.”

The analogy, and then the place it breaks

WHERE IT HOLDS

— Perfect performance on everything seen before.

— Nothing to say about anything one step outside it.

— A lookup table where a rule should be.

— And, from the inside, no way to tell the difference between this and competence.

WHERE IT BREAKS

— Overfitting is a failed attempt at generalizing. Funes never attempts it.

— A model has a training set and a test set. Funes has a life, and no one is scoring him.

— Overfitting is a tuning error someone can correct. Funes's condition is an injury.

— And the model has no interior. Funes is in distress, and the story is about the distress.

The rule this course keeps applying

Take the structure, leave the explanation. Funes shows you what a mind without compression would be unable to do, and that demonstration survives the fact that he is not a neural network and was never trained on anything. The resemblance is real; the identity is not. Say what the analogy illuminates and say where it stops — the same discipline you applied to mode collapse, to the genie, and to Chiang an hour ago.

ALL OF IT ON ONE LINE

Five archives, and the place the line gives out

The ordering

The Library keeps every possible string; Funes keeps every actual perception; a search engine keeps documents and ranks them; a 2026 model keeps a smoothed account weighted by frequency; a human expert keeps almost nothing and can still answer.

Where the line gives out

Two of the five do not belong on it. The Library never observed anything — it is a possibility space, not a record. And a search engine discards nothing at all; it defers the discarding to whoever reads page one. Neither is a point on an axis of forgetting.

So the second axis

Not how much is thrown away, but on what grounds. A search engine ranks by proxies for relevance. A model discards by frequency. A human expert discards by relevance to a purpose they could state if asked — and only the last of those has a reason behind it.

Which is why the dial is the wrong picture

A single axis from remembering to forgetting makes the design question look like a tuning problem with a right answer somewhere in the middle. The moment you add the second axis it stops being tuning and becomes a choice about criteria — and criteria belong to somebody.

PART FIVE · SYNTHESIS

Three answers to one question

Borges, twice

Keep everything and you get the Library: every truth present, none of them findable. Keep everything and you get Funes: every detail retained, nothing thinkable. Both are the same failure — a refusal to discard — and he wrote them a year apart.

Chiang, once

We have built the opposite and should be careful what we call it. The blur is the understanding, and also the fabrication, and he does not think the two come apart. The danger is not the blurry copy but the copy that is crisp and wrong.

And the arithmetic

Which says the choice was not really ours. Below about one bit per token nothing can be stored, so generalization is forced. We did not decide to build blurry archives. We decided to build big ones, and this is what big ones are.

So the question is not whether to forget. It is the one Borges asked twice and the field has never once put in writing: how much, and which, and chosen by whom.

TO TAKE AWAY

The question, and the version of it that is harder

Is intelligence what remains after you throw the details away? And: name one thing you know that you could not state the details of — was losing those details the cost of knowing it, or just a loss?

Three live positions

Yes; no; and the question is badly formed. The third has the strongest case and the least support, which is usually a sign it deserves another look.

What counts as an answer

A claim someone could disagree with, resting on one specific thing — a number, a line, a scene. “Borges shows us” is not evidence. “A day of Funes’s memory takes a day to replay” is.

Next unit

We keep the archive and change the question: from what is lost to who is missing from it.

SOURCES

Everything cited today, and where to check it

Borges, “The Library of Babel” (1941) and “Funes the Memorious” (1942)

Specifications — 410 pages, 40 lines, 80 characters, 25 symbols — are the story’s own. 25^1,312,000 and its 1,834,098 digits follow by multiplication; check them. Funes quotations from the standard English translations, which differ: compare yours against a neighbour’s.

Ted Chiang, “ChatGPT Is a Blurry JPEG of the Web”

The New Yorker, 9 February 2023. Xerox case, room areas and the arithmetic figures are his. Remember the date when you cite the measurements — the mechanism has aged well and the numbers have not.

Carlini et al., “Quantifying Memorization Across Neural Language Models” (2023)

Source of the three findings: 10× model size → +19 points of memorization; deduplication 3.6% → 1.2%; 33% extractable at 50 tokens of context, 65% at 450. Measured on GPT-Neo and GPT-J models trained on The Pile — not on current frontier models, and worth saying so if you cite it.

The capacity table

Parameter and token counts as publicly reported (GPT-3 175B/300B, Gopher 280B/300B, Chinchilla 70B/1.4T, Llama 2 70B/2T, Llama 3.1 405B/15.6T). Bits per token is arithmetic on those. The entropy comparison is approximate; the order of magnitude is not.