THE PRIMER · PART 1

It predicts the next word.

That is the whole machine — and almost everything a lawyer needs to know about these systems follows from it.

A new weekly strand. It explains how these systems actually work, from zero, for lawyers who will have to advise on them — and lands each part on a duty under a real data-protection law.

Rushikesh R. Mahajan, Advocate
Data protection, technology and AI governance · LL.M. in Law and Technology, Queen's University Belfast
Practises before the Bombay High Court and courts across India in civil and criminal matters

What this is / is not. This is general, illustrative and **non-exhaustive** information about how large language models work and how India's Digital Personal Data Protection Act, 2023 engages with them. It is **not legal advice**, **not a compliance determination**, and does **not** create an advocate–client relationship. Readers outside India should map every term to their own law. Current as at 7 September 2026.

Prepared by Rushikesh Mahajan, Advocate.

What a language model actually is

Most of what a lawyer needs to know about artificial intelligence follows from one sentence. Before the sentence, though, the words in it have to mean something — so this issue defines every term it uses, on the page, as it uses it.

Start with the name. A *model*, here, is simply a piece of software that has been tuned until it is good at one narrow task. A *language* model is one tuned to do this:

> **LANGUAGE MODEL**
>
> A program that takes a stretch of text and produces, for every possible continuation, a number saying how likely that continuation is to come next. It then picks one and adds it on.
>
> That is the entire specification. Nothing about truth, meaning, sources or intention is built into it.

Why would anyone build a next-word guesser?

This is the question that never gets answered, and everything else depends on the answer.

Nobody set out to make a machine that guesses words. They set out to make a machine that could *use* language. The obstacle was never ambition — it was teaching. Language cannot be written down as a set of rules. Decades were spent trying: grammar rules, dictionaries, hand-coded exceptions. Those systems broke on the first sentence a real person actually said.

Next-word prediction solved a **teaching** problem, not a language problem. Take any text ever written. Hide the last word. Ask the machine to guess it. You instantly know whether it was right, because the true answer was sitting there in the text all along.

No human has to mark the paper. That single property is the whole reason this approach won. The machine could be set an exam, and corrected, billions of times a day, against writing that already existed — with nobody supervising.

And then came the part that surprised the people doing it. To guess the next word *well*, you cannot avoid absorbing almost everything else:

> **WHY GUESSING FORCES LEARNING**
>
> *"The capital of France is ___"* — unguessable without the fact.
>
> *"The accused was granted ___"* — unguessable without the grammar and the legal register.
>
> *"She dropped the glass and it ___"* — unguessable without knowing something about how the world behaves.

Facts, grammar, tone, register, the ordinary physics of objects — all of it had to be picked up along the way, because all of it is what determines the next word.

Prediction was never the goal. It was the only exercise anyone found that was cheap enough to repeat a trillion times — and repeating it dragged everything else along behind it. What you talk to is the side effect.

What the machine is actually handling

Two more words, because the rest of this issue leans on them.

> TOKEN
>
> The machine does not work in words. It works in **tokens** — small chunks of text, roughly three-quarters of a word on average. *"Petition"* may be one token; an unusual name may be three or four.
>
> Where this issue says "word", it is choosing readability. The machine is handling tokens.

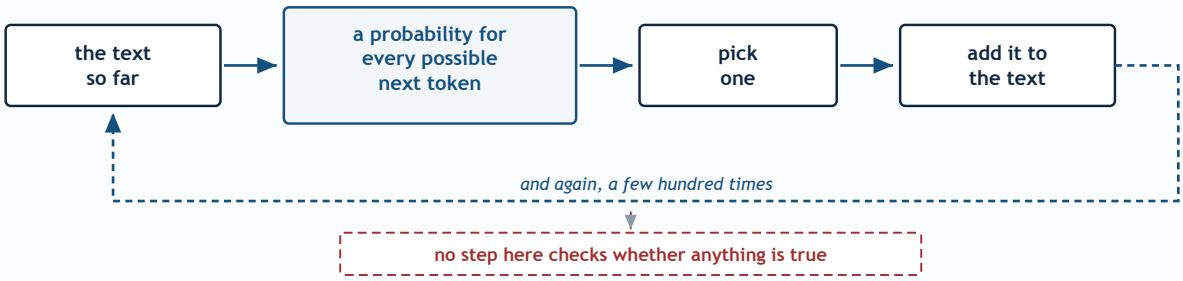
> STRING
>
> A **string** is a piece of text treated purely as a sequence of characters — letters, digits, spaces, punctuation — with no meaning attached.
>
> To the machine, *"AIR 1973 SC 1461"* is a string. So is *"AIR 1999 SC 8402"*. It has no way to ask whether either one corresponds to a judgment, because a string is all it ever had.

With those in hand, the sentence at the top of this issue can be stated properly, and it is now a description rather than a slogan.

A language model predicts the next token. Give it some text. It produces a probability for every possible next token. It picks one. It appends it. Then it does the same thing again with the slightly longer text, and again, until it stops. What you read as an answer is that loop running a few hundred times.

There is no step in that loop where the system decides what is true. There is no step where it consults anything. There is only the next token, and then the next one.

FIGURE 1 — THE LOOP



The entire mechanism. Every answer you have ever read from one of these systems is this loop, run to completion. Nothing in it consults a source, and nothing in it evaluates truth.

Where the numbers came from

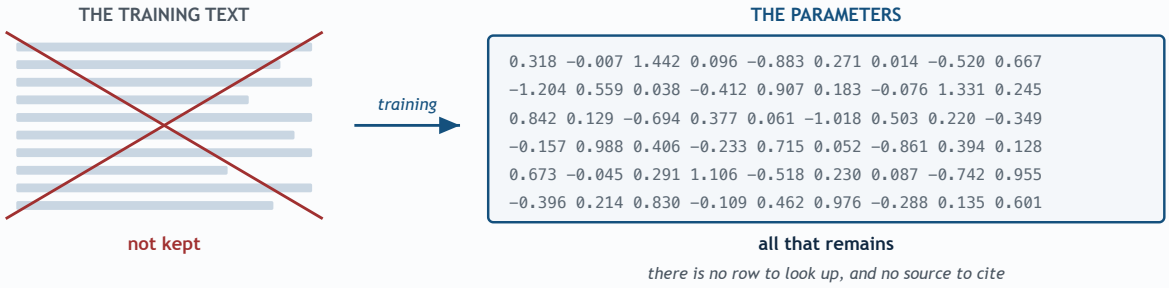
The system was shown an enormous quantity of text, and billions of internal numbers were adjusted, over and over, until its guesses about what came next got good.

> **PARAMETERS (ALSO: WEIGHTS)**
>
> The adjustable numbers inside the model — billions of them. Each one is a dial. Together they decide, for any input, which token comes out.
>
> **Training** is the process of nudging those dials: show it text, let it guess the next token, measure how wrong it was, adjust every dial a fraction in the direction that would have made it less wrong. Repeat, at enormous scale.

When the training finished, the text was not kept. **The parameters are all that remains.** The model is those numbers and nothing else.

This is why the question *"where did it get that from?"* usually has no answer. The honest reply is not that the source is hidden or hard to find. It is that the system never held a source in the first place. It holds a very large number of adjusted weights that make plausible text come out.

FIGURE 2 — WHAT SURVIVES TRAINING



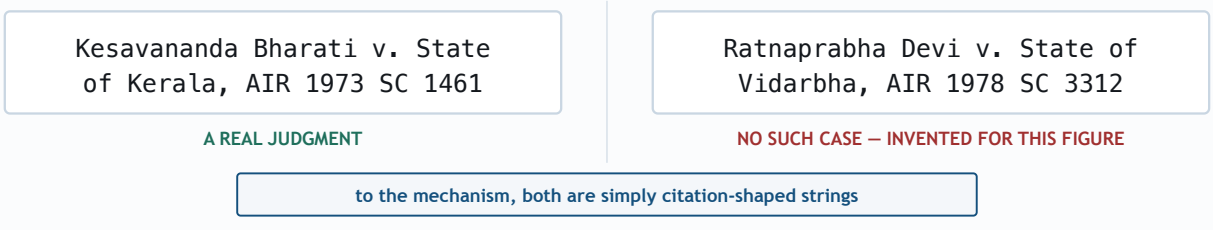
The model is the numbers. Training adjusts billions of weights until the next-word guesses improve; the corpus itself is discarded. This is why a model cannot tell you where an answer came from — **it never held a source to begin with.**

Why it invents things — and why that is not a bug report

Every lawyer who has used one of these has watched it produce a case that does not exist, with a citation formatted perfectly.

Here is what is happening. A real citation and an invented one look **identical to the machine** — both are simply strings of a familiar shape. Nothing in the mechanism distinguishes a string that corresponds to a judgment from one that merely follows the pattern of citations. It is producing citation-*shaped* text, and it does that equally well whether or not the case is real.

FIGURE 3 — INDISTINGUISHABLE TO THE MECHANISM



There is no tell. The right-hand citation is fabricated for this illustration. Nothing in the generating process distinguishes it from the left — which is why verification cannot be delegated to the tool that produced it.

In September 2025, OpenAI published a paper by Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala and Edwin Zhang titled **"Why Language Models Hallucinate."** Its argument is that these errors are neither mysterious nor incidental.

> *"Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty… We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty… Hallucinations need not be mysterious — they originate simply as errors in binary classification."*
>
> Kalai, Nachum, Vempala & Zhang, *Why Language Models Hallucinate*, arXiv:2509.04664 (4 September 2025)

Two things in that follow through to practice.

First, the error is structural, not accidental. The paper's position is that *"if incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures."* Distinguishing a real citation from a citation-shaped string is exactly such a case. The error is a property of the arrangement, not a defect in a particular product.

Second, the way these systems are scored rewards guessing. Standard evaluations give no credit for *"I do not know."* A model that guesses scores better than one that abstains. In the authors' phrase, *"language models are optimized to be good test-takers, and guessing when uncertain improves test performance"* — precisely as a student sitting a paper with no negative marking will outscore one who leaves blanks.

That is the part worth being blunt about: **the confidence of the answer tells you nothing about its accuracy.** These systems were optimised to be good test-takers. They are.

The three things that follow, for practice

1. "The AI said so" is not a source. It cannot be, because the system has no concept of a source. Anything it produces that matters must be verified against the actual instrument — the bare Act, the reported judgment, the notified rule.

2. The verification burden has not moved. It sits with the professional who signs, exactly as it did when the first draft came from a junior. Nothing about the machine changes who is answerable for the output.

3. Fluency is not competence. The single most dangerous property of these systems is that wrong output arrives in the same confident register as right output. There is no tell. There is no tremor in the voice. The paper above explains precisely why you should not expect one.

And the legal turn — where this series is going

Here is the thing that makes this a matter for a data-protection lawyer rather than a technologist.

The Digital Personal Data Protection Act, 2023 does not use the words "artificial intelligence." Neither do the Rules made under it. India's AI Governance Guidelines, released on 5 November 2025, are expressly a reference framework rather than a compliance requirement.

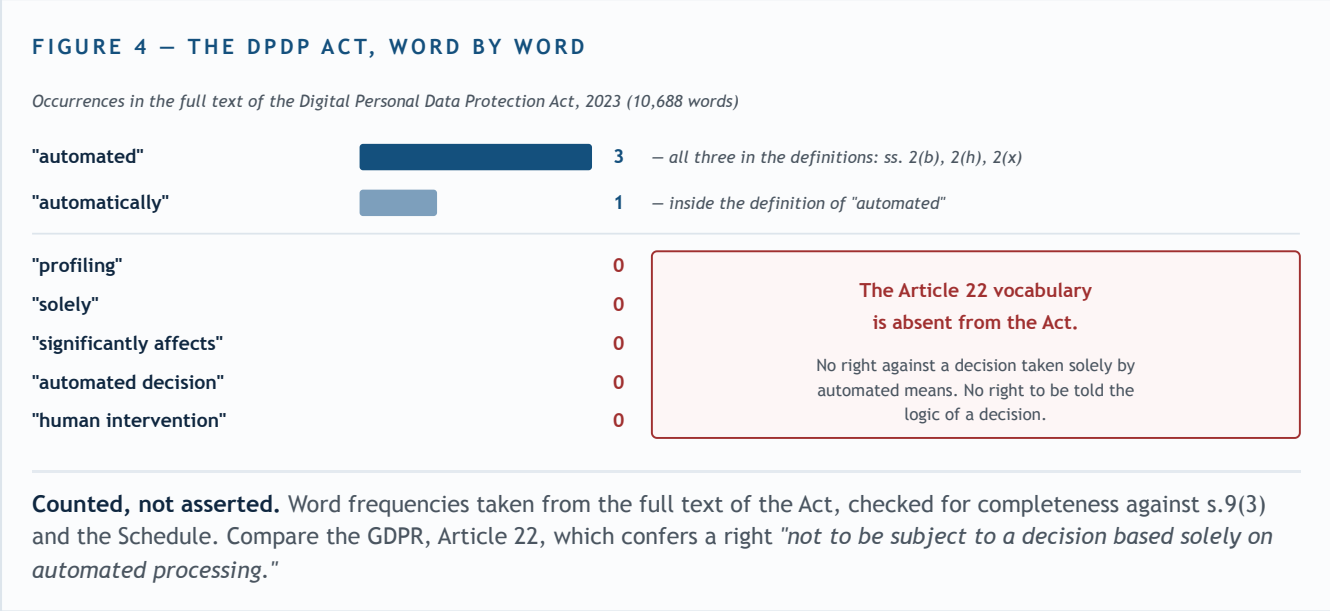
The conclusion most people draw from that is that nothing binds them yet.

The conclusion the drafting supports is different. **The Act does not regulate AI as a technology. It regulates processing.** And every stage of what you have just read is processing: the text the system was trained on, the text you type into it, the text it keeps in order to be useful to you next time.

Which is why this series starts at the machine rather than at the statute. You cannot apply a processing statute to a system until you know, concretely, what the system does with the material it is given.

India Desk

There is no automated-decision provision, and the absence is precise. Read the Act itself. The word *"automated"* appears three times and all three are in the definitions — of *"automated"* (s.2(b)), of *"data"* (s.2(h)) and of *"processing"* (s.2(x)). The words *"profiling"*, *"solely"*, *"significantly affects"*, *"automated decision"* and *"human intervention"* do not appear in the Act at all.



There is, in short, no Indian equivalent of the GDPR's Article 22 right, and no right to be told the logic of a decision. If a client's system refuses a loan, declines a claim or filters a candidate, the individual's handles under the Indian statute are the ordinary ones — notice, consent, purpose limitation, correction, erasure, grievance. **The output being machine-made adds nothing to their position.**

The binding AI rules are not in the data statute. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 were notified on 10 February 2026 and came into force on 20 February 2026. They introduce a regime for "synthetically generated information": labelling obligations for AI-generated visual and audio content, and a takedown window cut from thirty-six hours to three. They bind intermediaries — and reach platforms that enable the creation of synthetic content. **These are in force now**, while the substantive DPDP obligations are not.

Global — data protection & AI governance

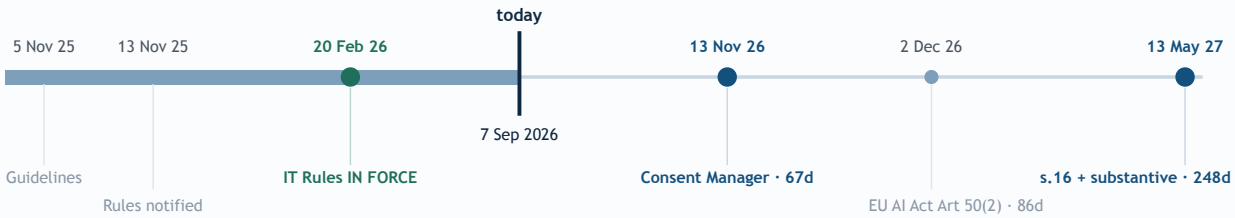
A European regulator has already held that an AI model is not automatically anonymous. In Opinion 28/2024, adopted 17 December 2024, the European Data Protection Board considered when an AI model trained on personal data can be treated as anonymous, and concluded that it cannot be assumed — it must be assessed case by case. The Board accepted that although these models are not designed to produce personal data, such data may nonetheless be **embedded in the model's parameters as mathematical objects**.

Read that against Figure 2. The parameters are all the model keeps. If personal data can survive inside them, then *"we deleted the training data"* is not the end of the analysis.

We come back to this properly in a later part of this series. It is flagged here because it is the direct consequence of how the machine is built.

Running clock — computed 7 September 2026

FIGURE 5 — THE DATES THAT ARE ACTUALLY RUNNING



The rules already in force are not the ones being counted down to. Distances are as at 7 September 2026 and nowhere else in this issue; every date above is stated absolutely so the text reads correctly whenever it is opened.

DATE	WHAT IT IS	DISTANCE
5 Nov 2025	India AI Governance Guidelines released	306 days ago
13 Nov 2025	DPDP Rules, 2025 notified	298 days ago
20 Feb 2026	**IT Amendment Rules, 2026 in force — SGI labelling + 3-hour takedown**	**199 days ago — in force**
13 Nov 2026	**Rule 4 / s.6(9) — Consent-Manager registration commences**	**67 days**
2 Dec 2026	EU AI Act Art 50(2) transitional relief expires	86 days
13 May 2027	**s.16 + Rule 15, and the substantive package**	**248 days**

> **Practitioner's verdict — what to tell a client this week**
>
> If they are relying on one of these systems for anything that reaches a client, a court or a regulator, the question to ask is not *"is it accurate?"* — it is *"who checks it, and what do they check it against?"* There is no configuration, no prompt and no subscription tier that makes the output self-verifying, and the people who build these systems have published the reason why. If the answer is that the tool is trusted because it sounds right, that is not a workflow, it is an exposure. And if they tell you no Indian law applies yet because there is no AI statute — they are right about the statute and wrong about the exposure. **The data law does not need to name the technology to reach it.**

NEXT IN THIS SERIES

Part 2 — training versus using. Your client's data can end up in the weights, or in the prompt. Those are different events with different consequences, and most privacy policies do not distinguish them.

This publication is informational and educational only. It is not legal advice and creates no advocate-client relationship. Authored from India; readers in other jurisdictions should map terms to local law.

Every statutory reference and every date in this issue was checked against at least two independent sources before it went out. Where a claim could not be verified to that standard it does not appear, or it is stated as an absence of record rather than as a fact.

ABOUT THE AUTHOR & WHERE TO FIND THIS

Rushikesh R. Mahajan

Advocate · Data protection, technology and AI governance

LL.M. in Law and Technology, **Queen's University Belfast**

Practises before the **Bombay High Court** and courts across India in **civil and criminal matters**

Enrolment No. MAH/11261/2021 · Bar Council of Maharashtra & Goa

Writes on the Digital Personal Data Protection Act, cross-border data transfer, AI governance and technology contracts — and publishes open reference tools for Indian data and AI law as **wolfgang_rush**.

THIS PUBLICATION

The India Data & AI Governance Desk

Weekly. The Primer runs in every issue.

READ EVERY ISSUE · FREE

buttondown.com/the_desk/archive

WRITING AND OPEN TOOLS

wolfgangrush.github.io

LINKEDIN

linkedin.com/in/rushirmahajan