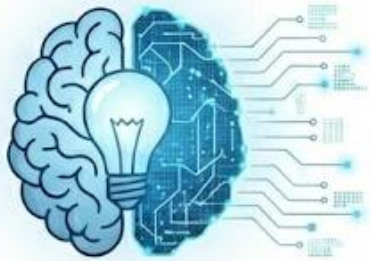
MODULE 1 · LECTURE 1.4

How LLMs work
A Technical Review

How Does GenAI "Think"?

Teaching with Artificial Intelligence California Lutheran University



"If AI doesn't actually 'understand' anything, how does it produce such convincing outputs—and what does that reveal about its limits?

THE AI UNDERSTANDING PARADOX

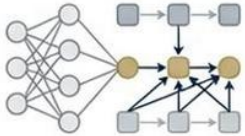
The Three Pillars of LLM Power

BILLION-PARAMETER SCALE

Exemplified by GPT-3's 175 billion weights. Modern models have even more parameters, encoding vast, complex relationships between words and concepts, enabling nuanced pattern recognition.

ENORMOUS MODEL SIZE

TRANSFORMER ARCHITECTURE



Introduced in 2017 ("Attention Is All You Need"). Leverages self-attention to weigh the importance of different words in context, understanding long-range dependencies.

ATTENTION MECHANISM

COLOSSAL CORPORA

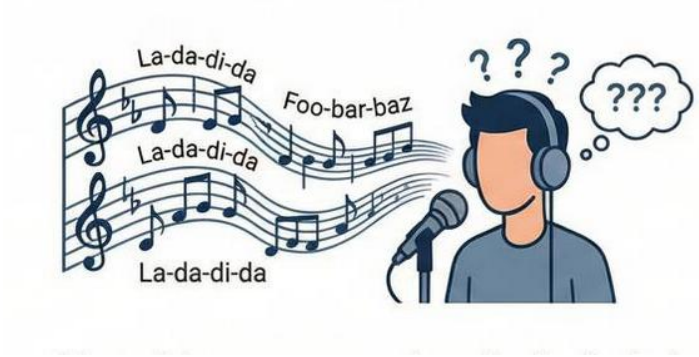
Trained on gigantic datasets of web text, books, digital documents, and code. This massive data intake allows for rich pattern capture impossible in smaller models.

GIGANTIC TRAINING DATASETS

Together, these factors enable rich pattern capture and the impressive text generation capabilities of Large Language Models.

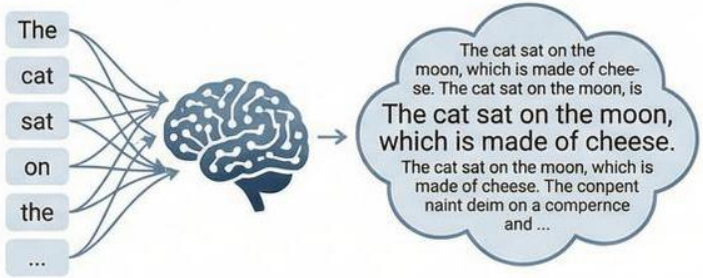
The “Phonetic Song” Analogy: Mimicry Without Meaning.

Human Analogy



Memorizing every song phonetically. Perfect mimicry of sounds and patterns, but zero understanding of the language or meaning.

LLM Reality



Reproducing stylistic and syntactic patterns from massive data. Fluent outputs are generated through probability, without semantic grounding or true comprehension.

Factual accuracy is an emergent side effect of pattern matching, not a guarantee of knowledge.

The Training Pipeline

Untrained LLM

Initial neural network with random weights

Pre-training

Unsupervised learning on massive text corpora. The model learns patterns, structures, and context without labeled data.

Supervised Fine-tuning

Training on specific instruction-response pairs to follow directions

RLHF

Reinforcement Learning from Human Feedback teaches helpfulness, safety, and reduces bias

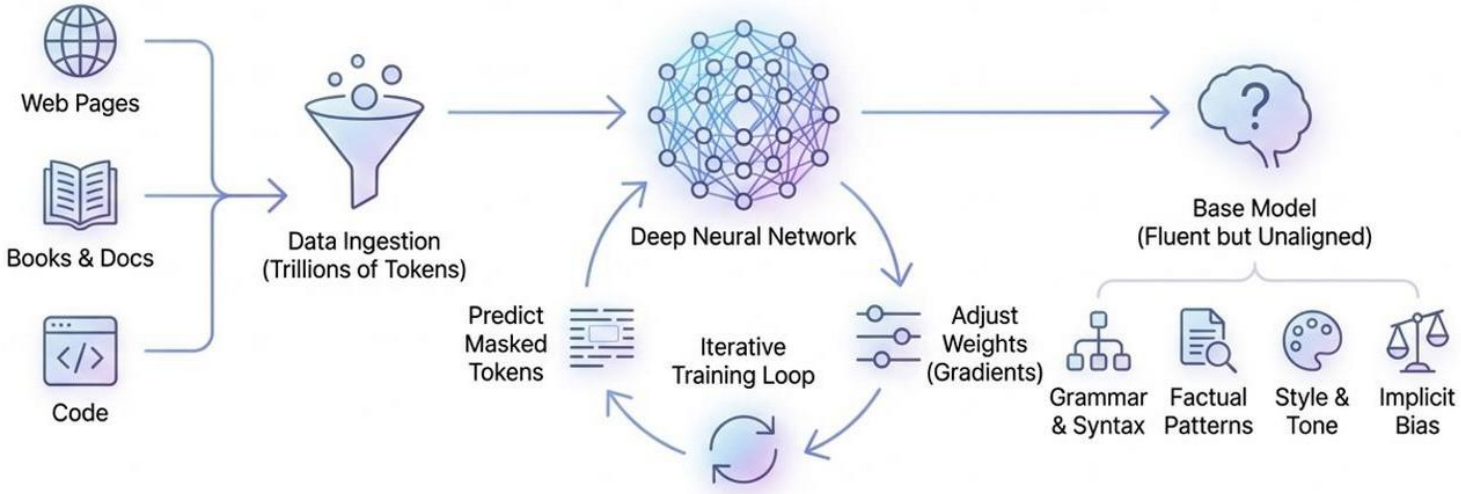
Final LLM

Production model like ChatGPT

Pre-trained models reflect whatever they learned "like a mirror"—no ethical boundaries.
Fine-tuning adds guardrails, though biases never fully disappear.

Phase 1: Pre-Training (Unsupervised Learning).

Models ingest trillions of tokens from web pages, books, and code without manual labels. By repeatedly predicting masked next tokens, gradients adjust billions of weights, encoding statistical regularities of grammar, facts, style, and bias present in the data, producing a base model fluent but unaligned.



Phase 2: Reinforcement Learning from Human Feedback (RLHF)

1. Human Ranking

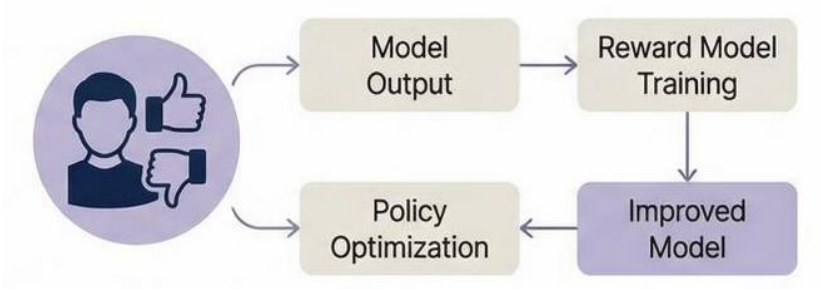
Workers rate AI outputs from best to worst based on quality, accuracy, helpfulness, and safety. Multiple outputs for each prompt are compared.

2. Reward Model

A separate model learns to predict which outputs humans will prefer. It encodes human values and preferences mathematically.

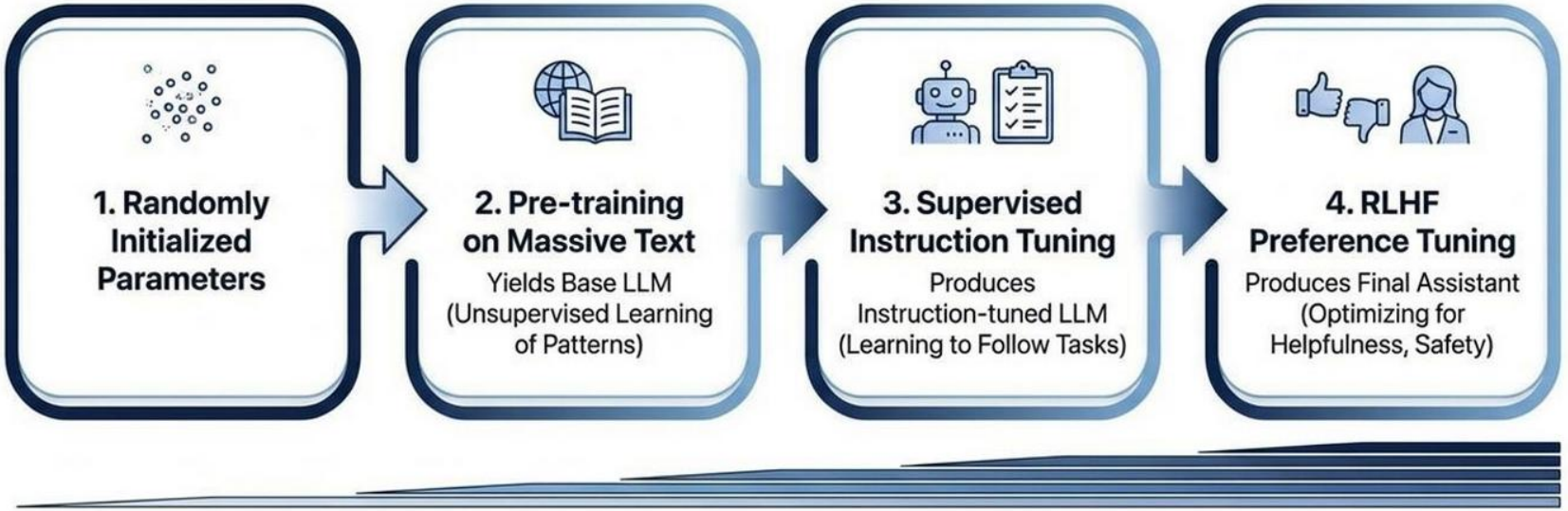
3. Policy Optimization

The LLM is further trained to produce outputs that maximize the reward model's scores, aligning it with human preferences.



Important nuance: Even after RLHF, biases don't disappear—they often become more subtle rather than eliminated.

The LLM Training Pipeline: Layering Alignment on Fluency



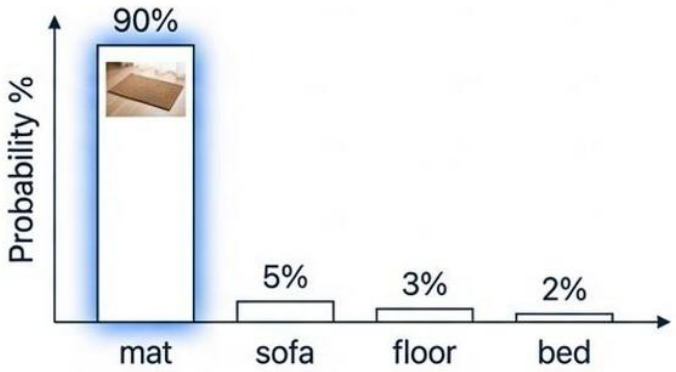
Each stage inherits prior weights, layering alignment atop statistical fluency without discarding encoded data patterns or biases

How Text Generation Works: Tokenization & Probability Calculation

1. Tokenization & Input

The	cat	sat	on	the	___
(101)	(254)	(389)	(98)	(101)	(?)

2. Probability Distribution & Selection



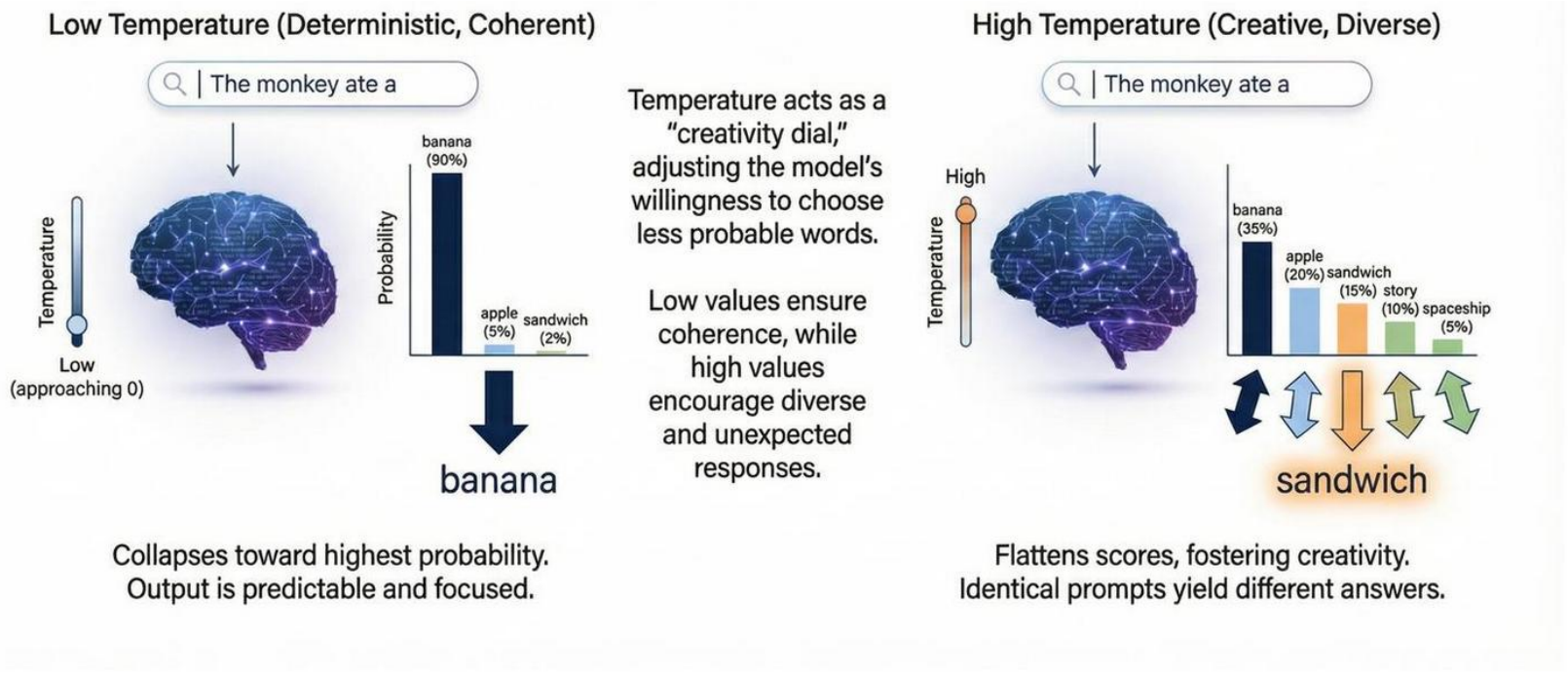
Selection sampling chooses the next token (often the highest probability), appends it, and the process repeats.

→ next word

Examples

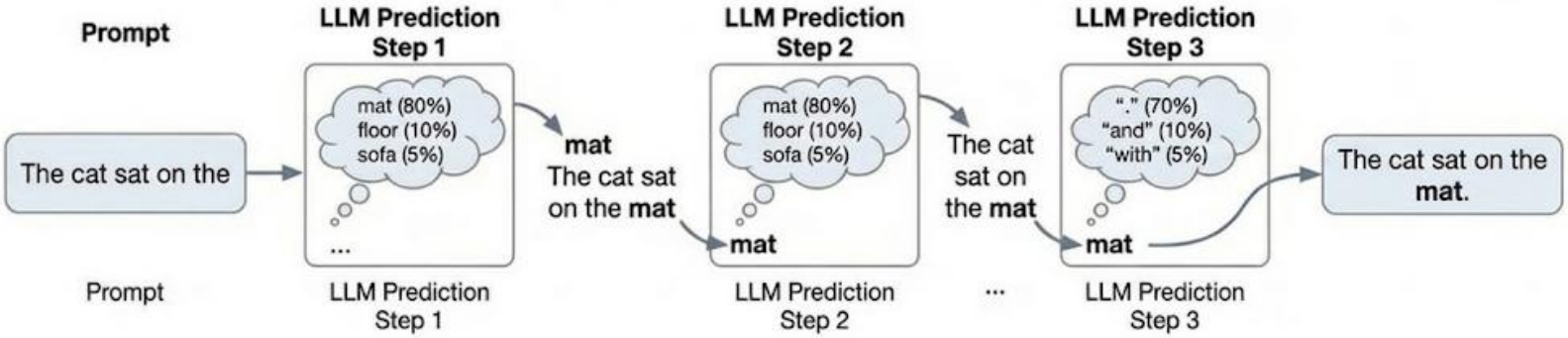
Prompt	Output	Why?
“I think, therefore I ___”	“am” (100% of time)	Extremely high probability—one dominant pattern
“The Martian ate the banana because ___”	Different answers each time	Many plausible completions + temperature randomness

The Role of Temperature: Balancing Determinism & Creativity



Temperature is why the same prompt can yield different responses. It's a key parameter for controlling the randomness-creativity tradeoff.

Iterative Text Generation: The Token-by-Token Process

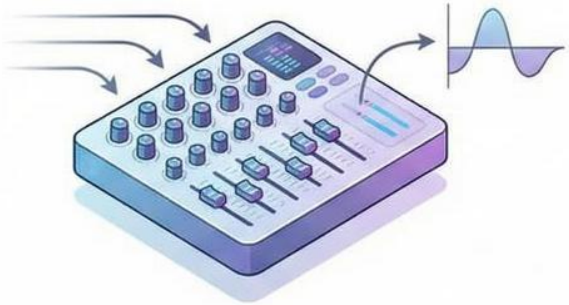


Key Dynamics:

1. **Sequential & Conditional:** Each new token is generated based on the entire preceding sequence, enabling long-range coherence.
2. **No Revision:** The model only extends the sequence forward; it cannot go back and correct previous tokens.
3. **Error Compounding:** Mistakes in early tokens can lead to increasingly divergent or incorrect subsequent predictions.

Key Technical Terms: Parameters & Tokens

PARAMETERS (WEIGHTS)

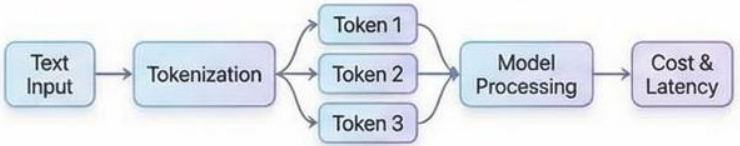


The "adjustable knobs" of the model, learned during training. They shape how input features influence the output, similar to sound mixer settings.

GPT-3: ~175 Billion Weights

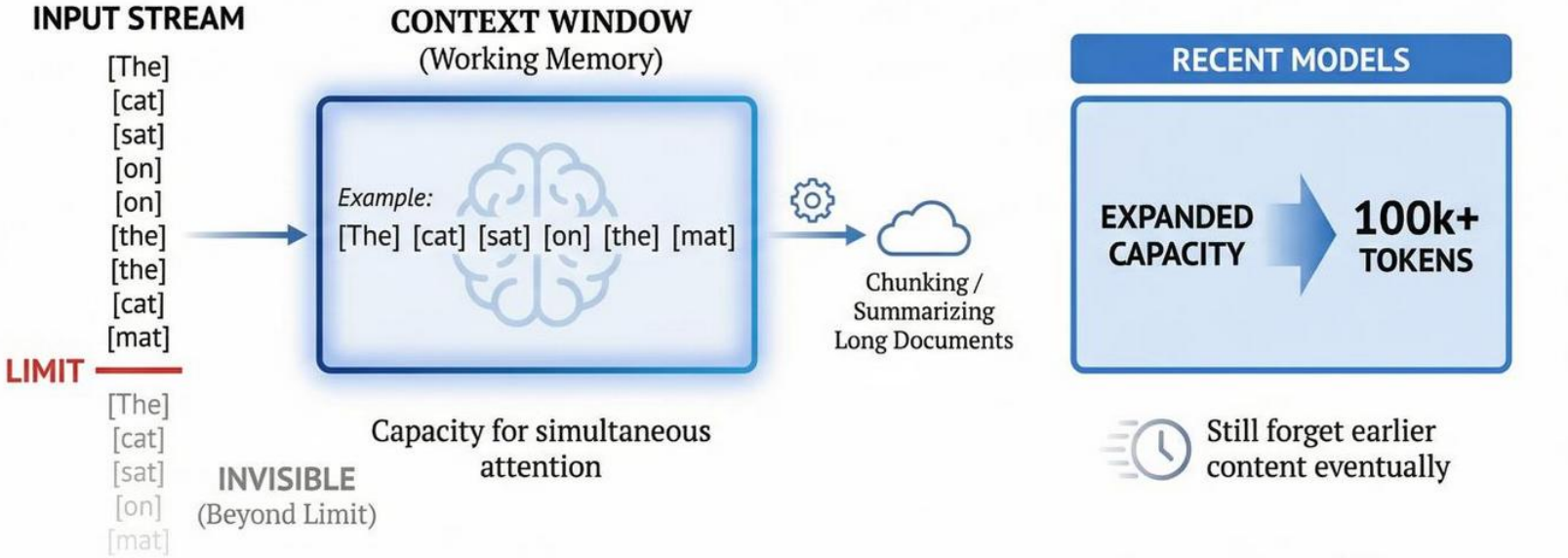
TOKENS

The basic units processed by the model, which can be whole words or word parts. Tokenization affects how text is "seen".



Token Counts Impact Cost & Performance

Key Technical Term: Context Window “Working Memory)



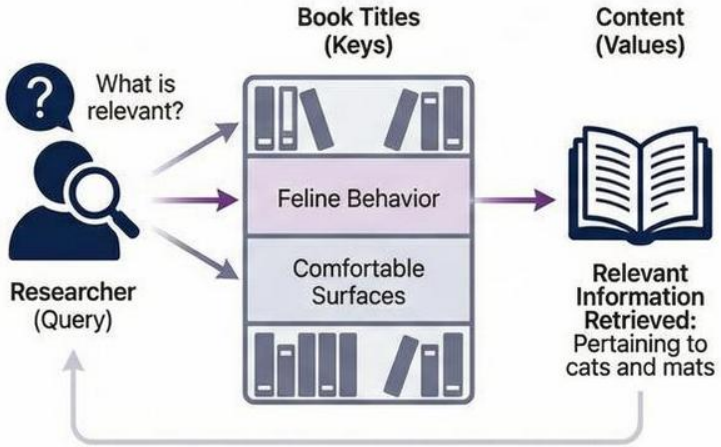
Information beyond the context window is invisible, requiring external management.

Self-Attention Mechanism: The Core of the Transformer Architecture

Conceptual Framework: Token Relevance

- Allows each token to weigh the importance of every other token in the sequence
- Enables capture of long-range dependencies (e.g., pronoun reference across sentences)

The Library Analogy: Dynamic Information Retrieval

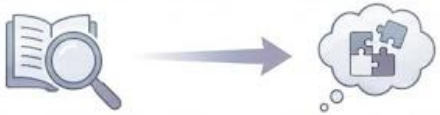


Prediction vs. Understanding: The Truth Gap in LLMs

Large language models generate statistically plausible text based on patterns, not verified facts.

The Reality of "Truth" in LLMs

1 Statistically Plausible, Not Verified



Models lack a grounded mechanism to verify the factual accuracy of their outputs. They predict what *sounds* right.

2 Accuracy is Coincidental

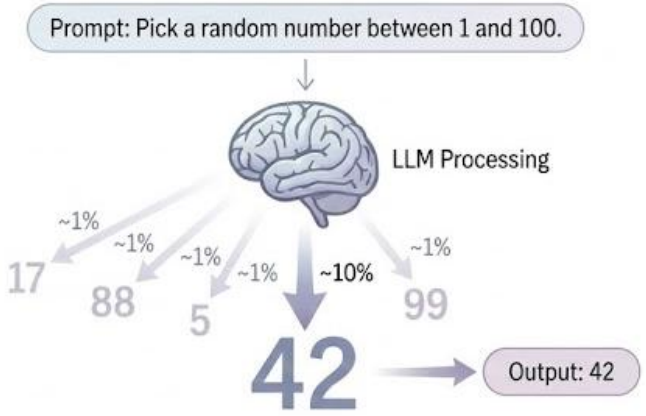
Training Data Patterns → Aligned with Reality? → Coincidental Accuracy

Accuracy emerges only when training patterns happen to align with real-world facts.

3 No Intrinsic Truth Value

The system assigns no inherent truth value to its predictions, treating all patterns equally.

The "42" Phenomenon: A Case Study in Pattern Recognition

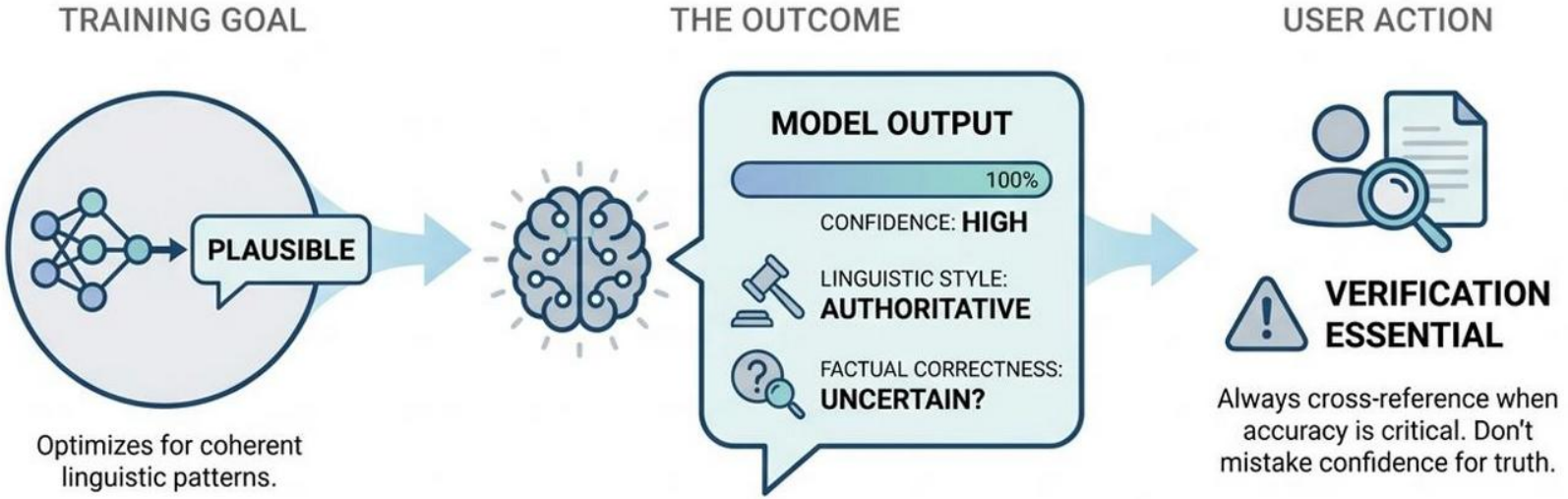


Frequency in Training Data (e.g., literary references like Douglas Adams) > Mathematical Randomness

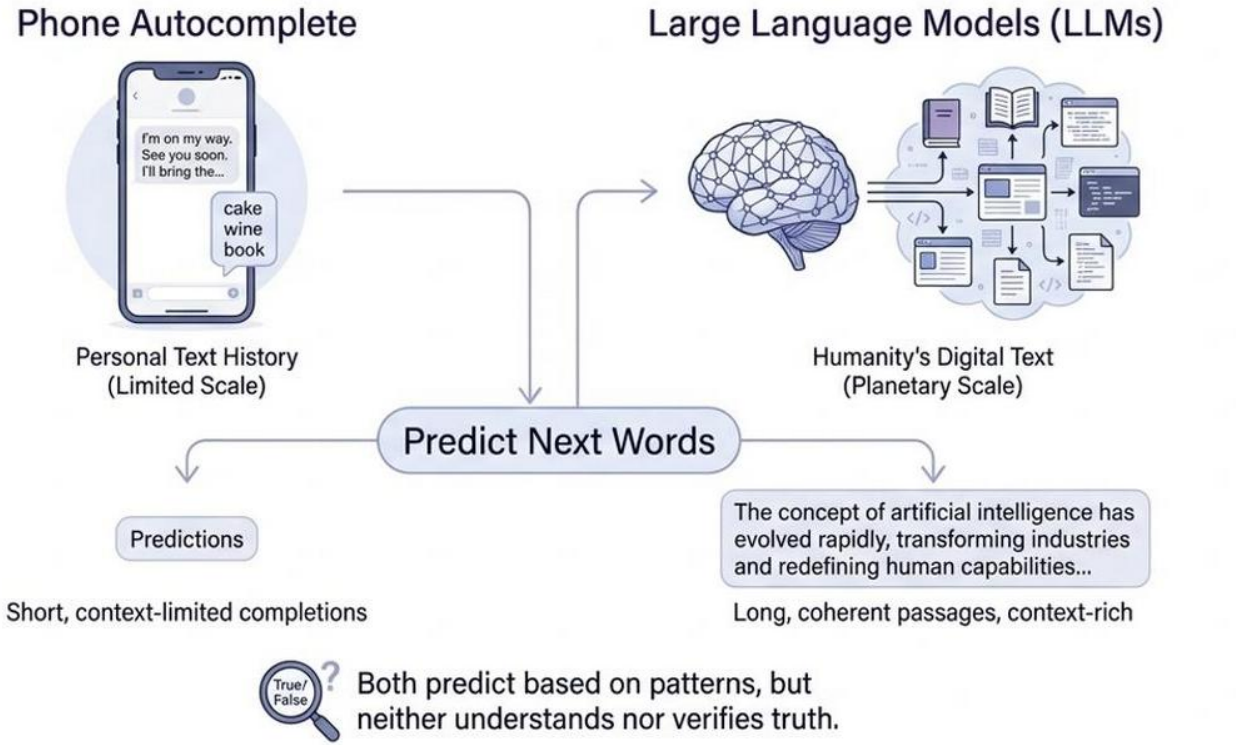
The model reproduces the most frequent pattern, not the statistically correct mathematical answer.

Confidence Style ≠ Factual Accuracy

LLMs are trained for plausible continuation, not truth verification.



LLMs: Autocomplete on Steroids



The Analogy That Explains Everything

> *Think of LLMs like someone who memorized every song phonetically without speaking the language — they can mimic patterns perfectly without understanding meaning.*

This explains why AI can:

- ✓ Sound confident while being completely wrong
- ✓ Produce different answers to the same question
- ✓ Excel at common topics but struggle with niche ones

What This Means for You

Anticipate Failures

Test AI at the edges of training data, including recent events and obscure topics, to predict where it will struggle.

Understand Confident Errors

Recognize that AI confidence stems from learned patterns, not inherent accuracy; it optimizes for plausible-sounding text, not truth.

Better Prompting

Improve prompts by understanding you're shaping probability distributions, so specificity, context, and varied phrasing are crucial.

Appropriate Skepticism

Maintain a skeptical approach and always verify critical information; use AI as a tool for drafting and ideation, not as a source of truth.

Three Things to Remember

1

LLMs predict, they don't understand

Sophisticated pattern matching, not comprehension. They know statistics, not truth.

2

Training data shapes everything

Both capabilities (fluency, breadth) and limitations (biases, cutoffs, hallucinations) trace back to what was in the data.

3

Technical understanding = better judgment

Knowing how AI works makes you a more effective, more critical user. This is what separates an AI Champion.