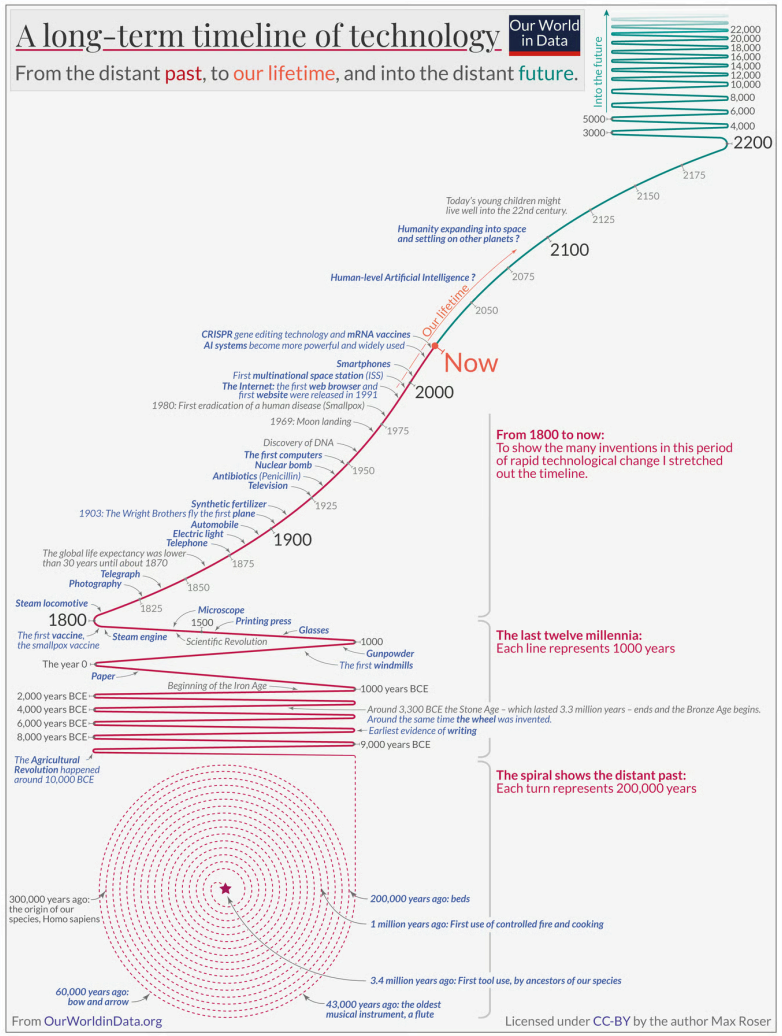
Market Research and Analysis

Lecture 1: Big Data Analysis in AI Era

- Zhenyu Zhao
- Nankai Institute of International Economics
- Feishu: 2120253538
- Email: zzynankai@outlook.com
- Website: xishanyu2.github.io

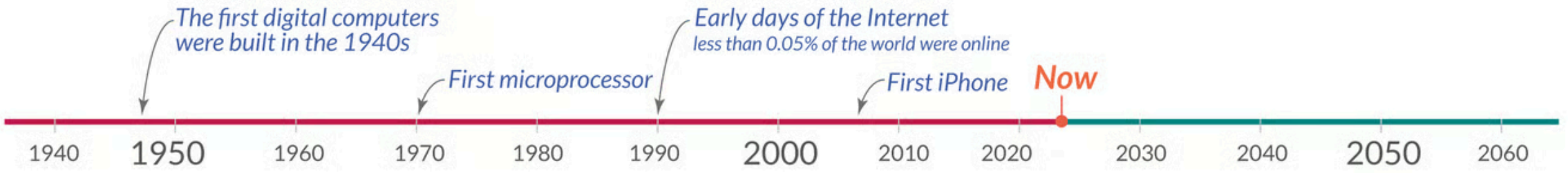
01 Our world in data

> In God we trust; all others must bring data. —*William Edwards Deming*


A long-term timeline of technology
Our World in Data
From the distant past, to our lifetime, and into the distant future.
Into the future
22,000
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
5000
4,000
3000
2200
2175
2150
Today's young children might live well into the 22nd century.
2125
Humanity expanding into space and settling on other planets ?
2100
Human-level Artificial Intelligence ?
2075
Our lifetime
2050
CRISPR gene editing technology and mRNA vaccines
AI systems become more powerful and widely used
Now
Smartphones
First multinational space station (ISS)
The Internet: the first web browser and first website were released in 1991
2000
1980: First eradication of a human disease (Smallpox)
1969: Moon landing
1975
Discovery of DNA
The first computers
Nuclear bomb
Antibiotics (Penicillin)
Television
1950
Synthetic fertilizer
1903: The Wright Brothers fly the first plane
Automobile
Electric light
Telephone
1925
1900
The global life expectancy was lower than 30 years until about 1870
1875
Telegraph
Photography
1850
Steam locomotive
1825
Microscope
1800
1500
Printing press
Glasses
The first vaccine, the smallpox vaccine
Steam engine
Scientific Revolution
1000
Gunpowder
The first windmills
The year 0
Paper
Beginning of the Iron Age
1000 years BCE
2,000 years BCE
4,000 years BCE
Around 3,300 BCE the Stone Age – which lasted 3.3 million years – ends and the Bronze Age begins.
Around the same time the wheel was invented.
6,000 years BCE
8,000 years BCE
Earliest evidence of writing
9,000 years BCE
The Agricultural Revolution happened around 10,000 BCE
300,000 years ago: the origin of our species, Homo sapiens
200,000 years ago: beds
1 million years ago: First use of controlled fire and cooking
3.4 million years ago: First tool use, by ancestors of our species
60,000 years ago: bow and arrow
43,000 years ago: the oldest musical instrument, a flute
From 1800 to now:
To show the many inventions in this period of rapid technological change I stretched out the timeline.
The last twelve millennia:
Each line represents 1000 years
The spiral shows the distant past:
Each turn represents 200,000 years
From OurWorldinData.org
Licensed under CC-BY by the author Max Roser

Timeline of the three transformative events in world history:

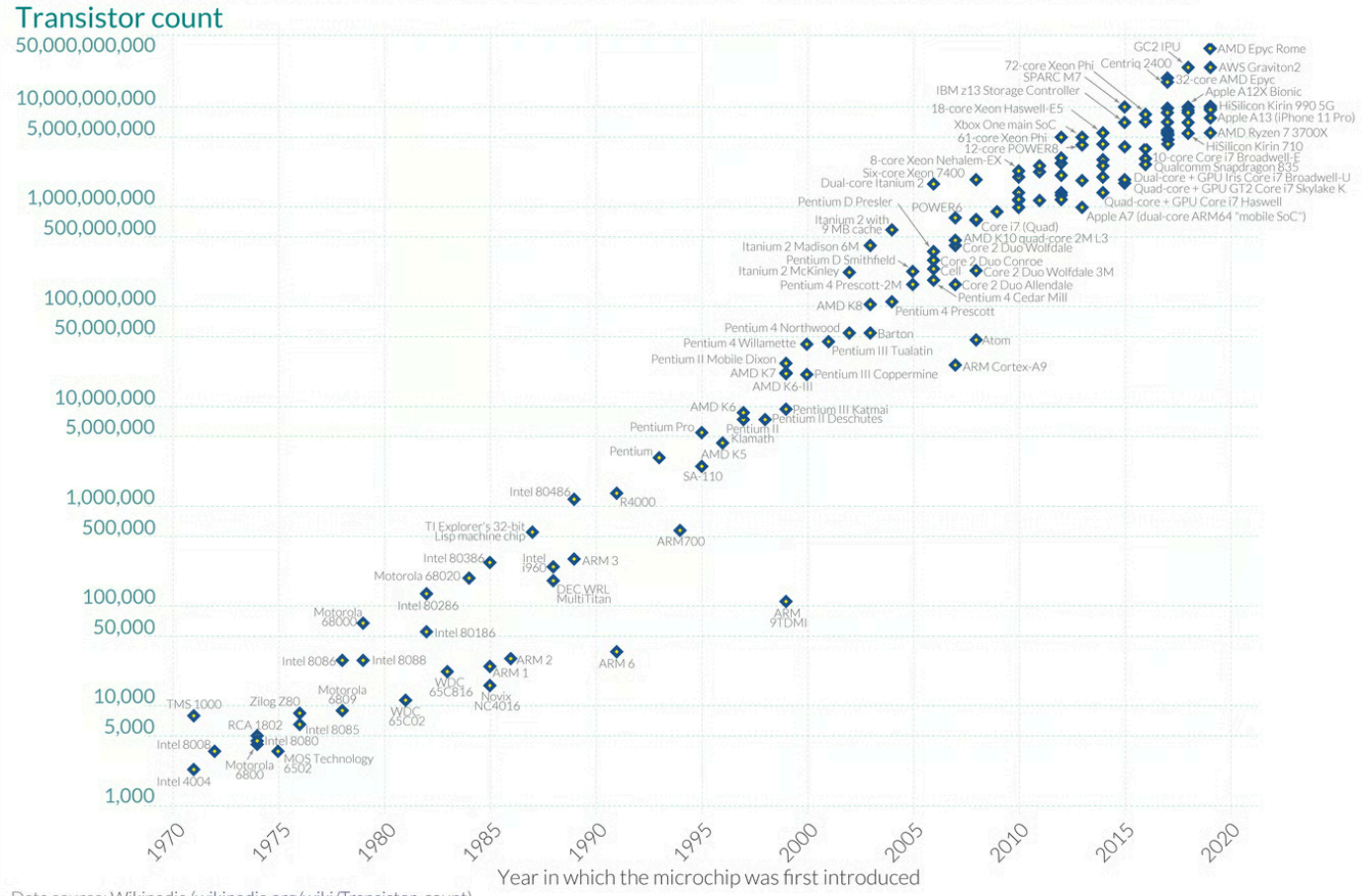
In a short period, computers evolved so quickly and became such an integral part of our daily lives that it is easy to forget how recent this technology is.



Moore's Law: The number of transistors on microchips has doubled every two years

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.


Data is all around us


DARTMOUTH ST
DO NOT ENTER
DO NOT ENTER
144
LUSH
FRESH
HANDMADE
COSMETICS

Trees

Boston Parks and Recreation Department (BPRD)

Location (lat-lon, address, neighborhood), species, size (diameter, height), date planted

Building Parcel

City of Boston, Assessing Department

property info (ownership, address, boundaries, area), financial details (value, taxes, mortgage), physical structure (year built, square footage, condition, # of units, basement/garage size), usage/zoning (land use codes, permitted activities), geography (APN, FIPS, coordinates)

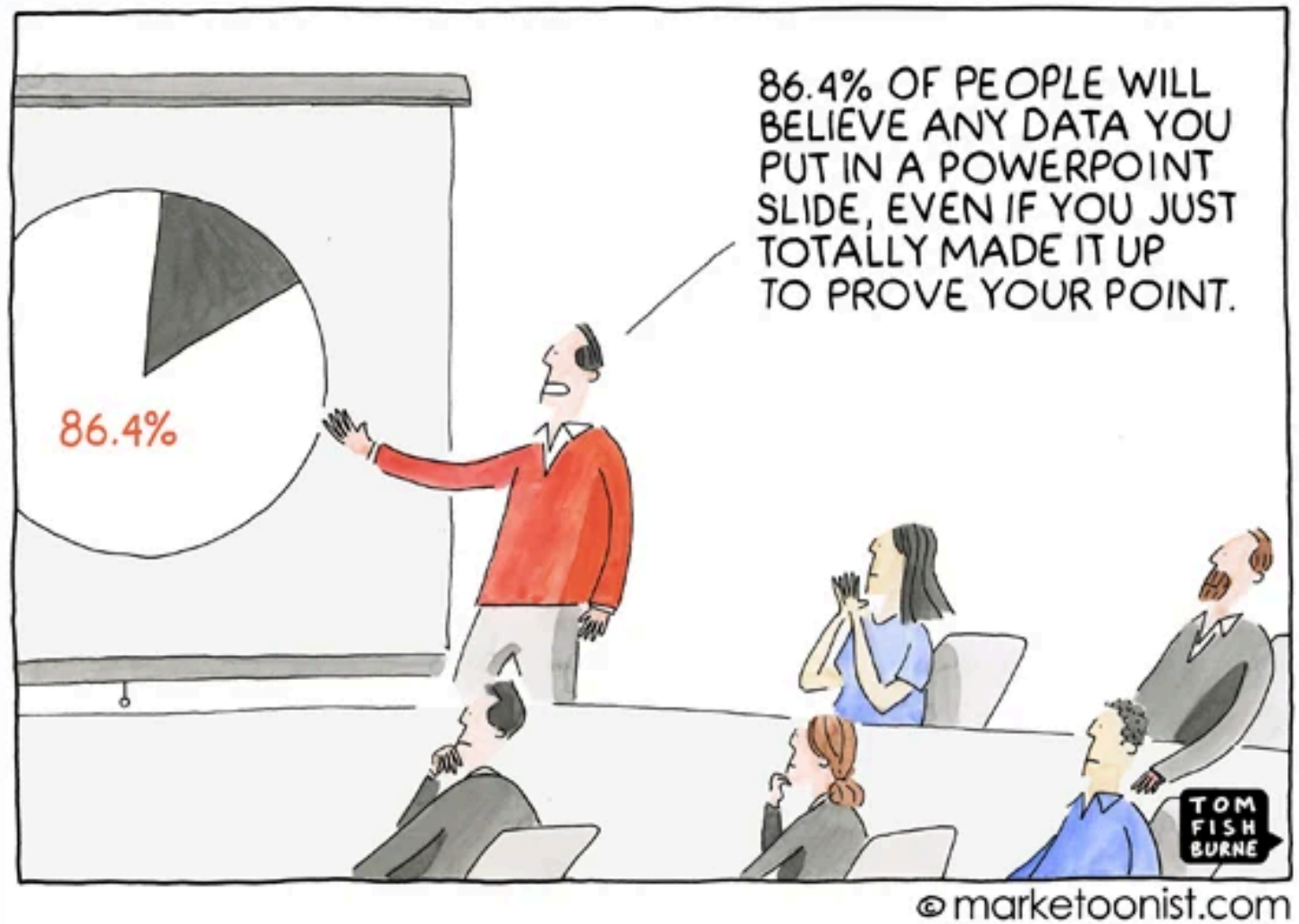
Census Data: Population & Demographics

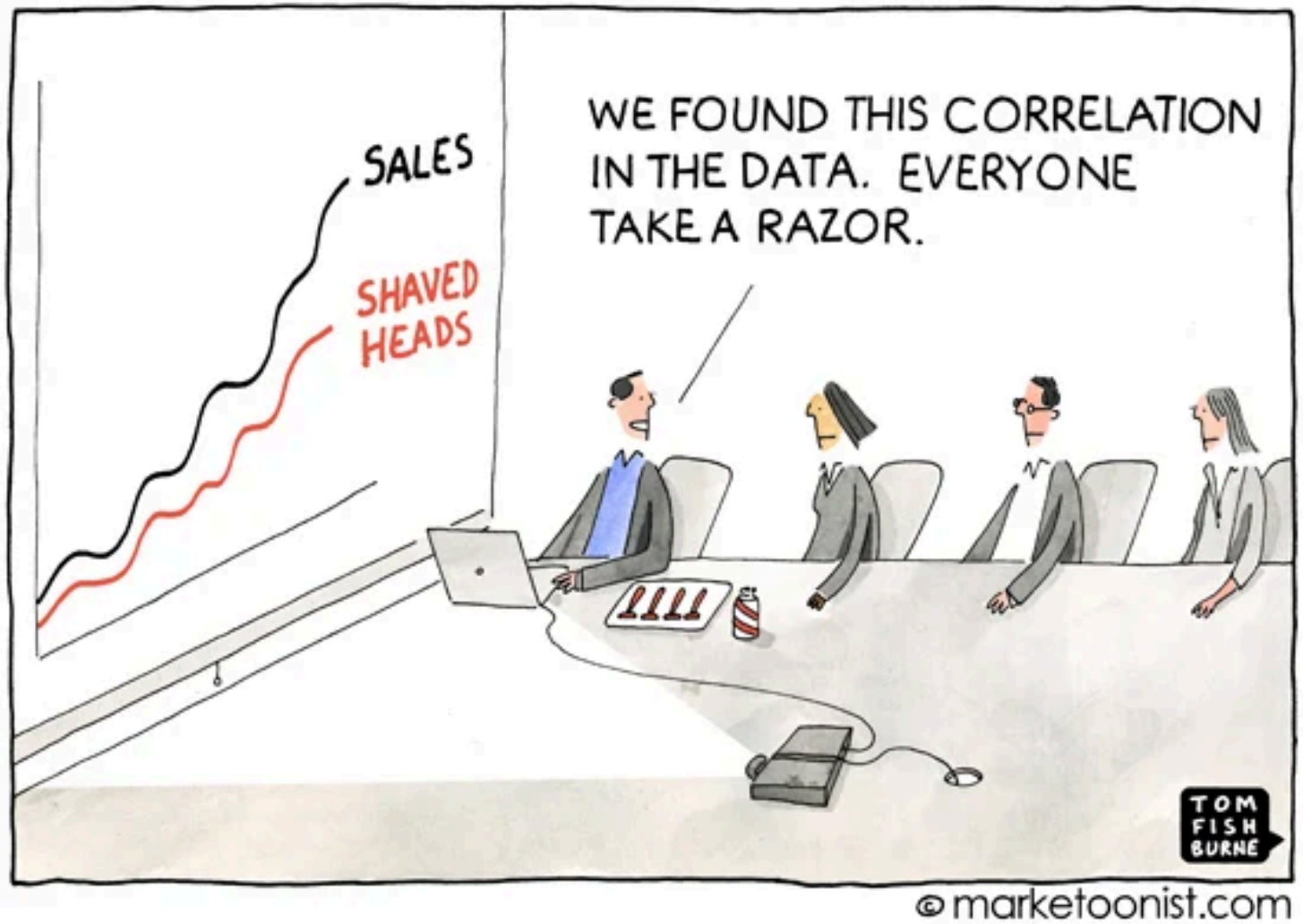
American Community Survey (ACS)

age, sex, race, ethnicity, income, education, employment, ancestry, language, disability, and housing characteristics (tenure, occupancy) for individuals and households, etc.

02 Data Visualization

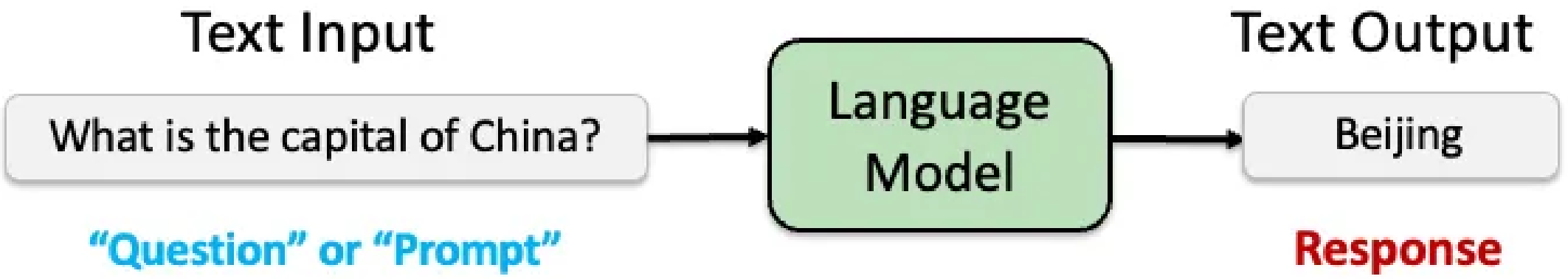
> The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it. ——*Hal Varian*

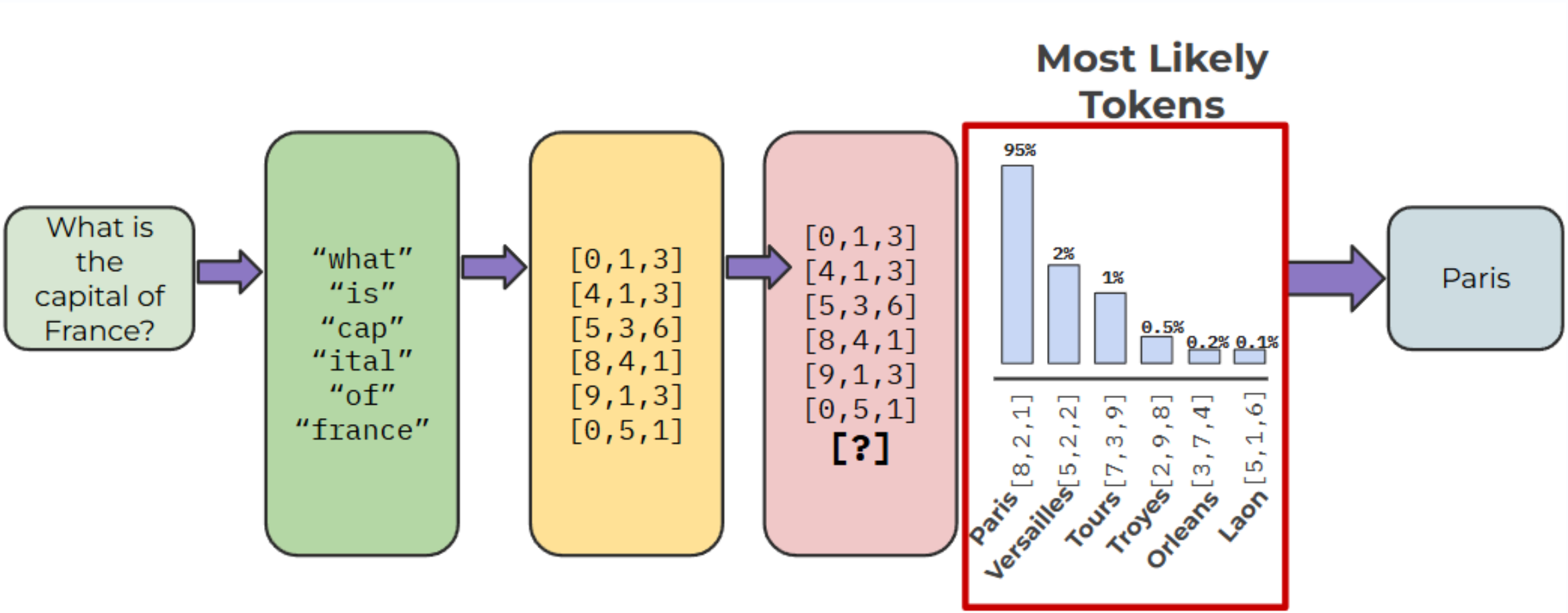

86.4% OF PEOPLE WILL BELIEVE ANY DATA YOU PUT IN A POWERPOINT SLIDE, EVEN IF YOU JUST TOTALLY MADE IT UP TO PROVE YOUR POINT.
86.4%
TOM FISH BURNE
©marketoonist.com

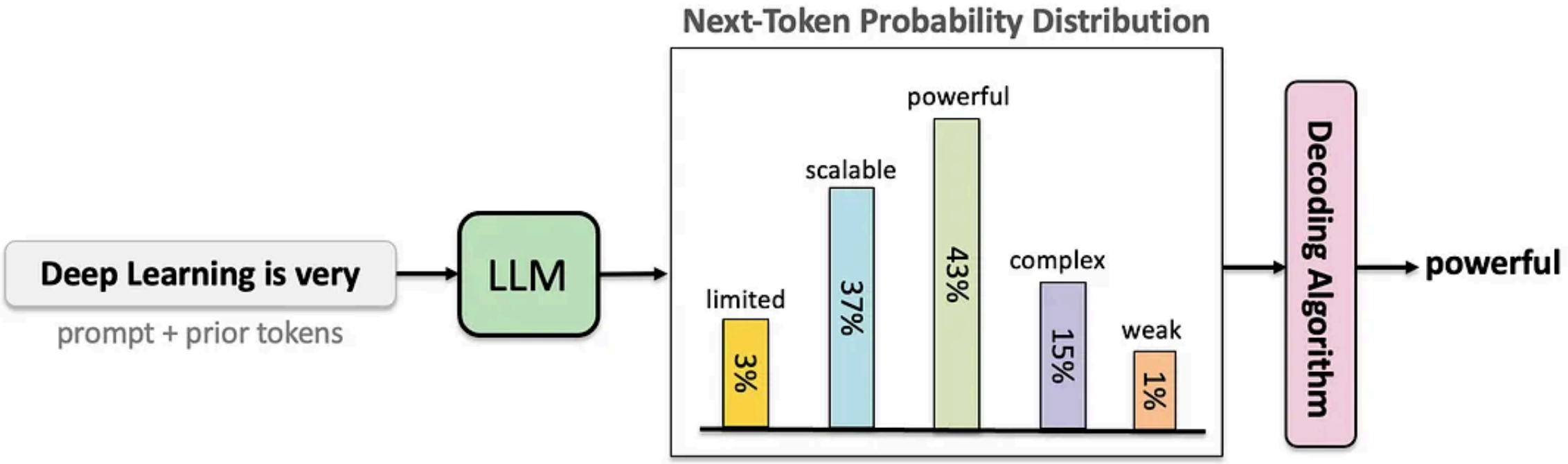

SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
©marketoonist.com

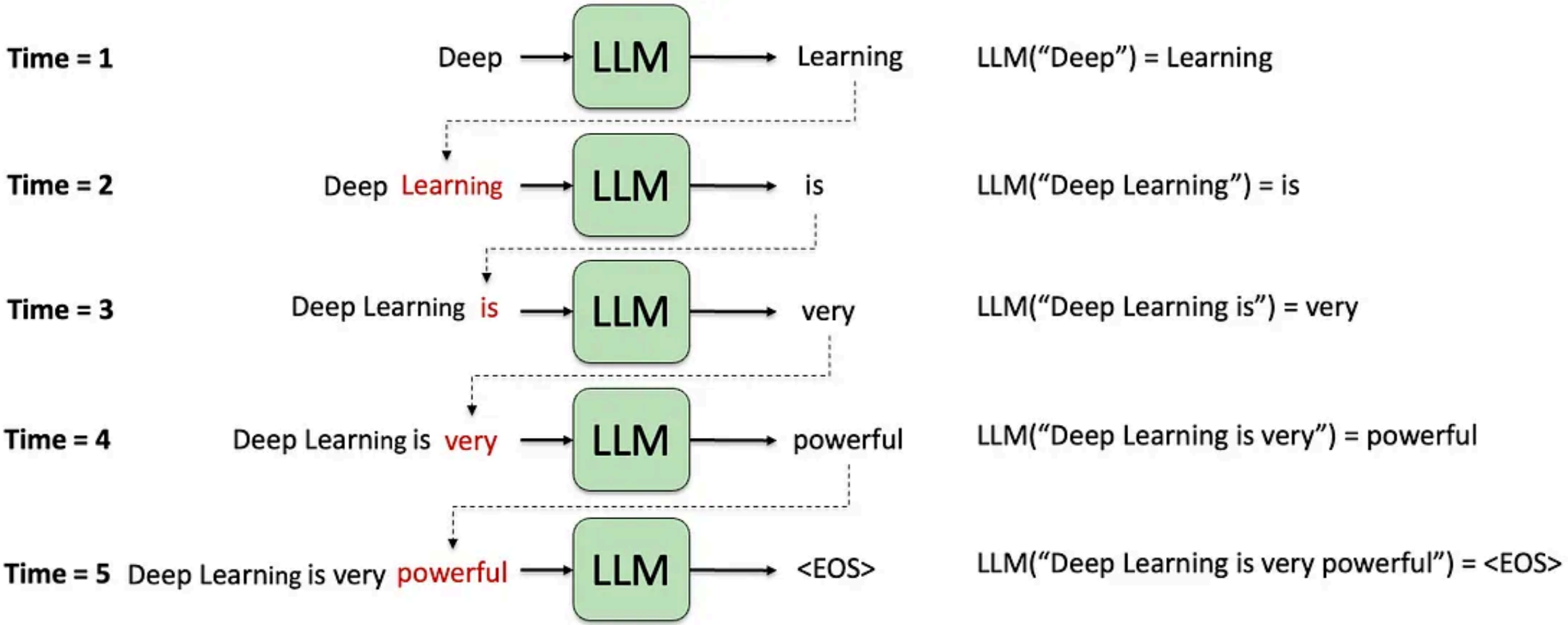
03 LLM - Next token predictor

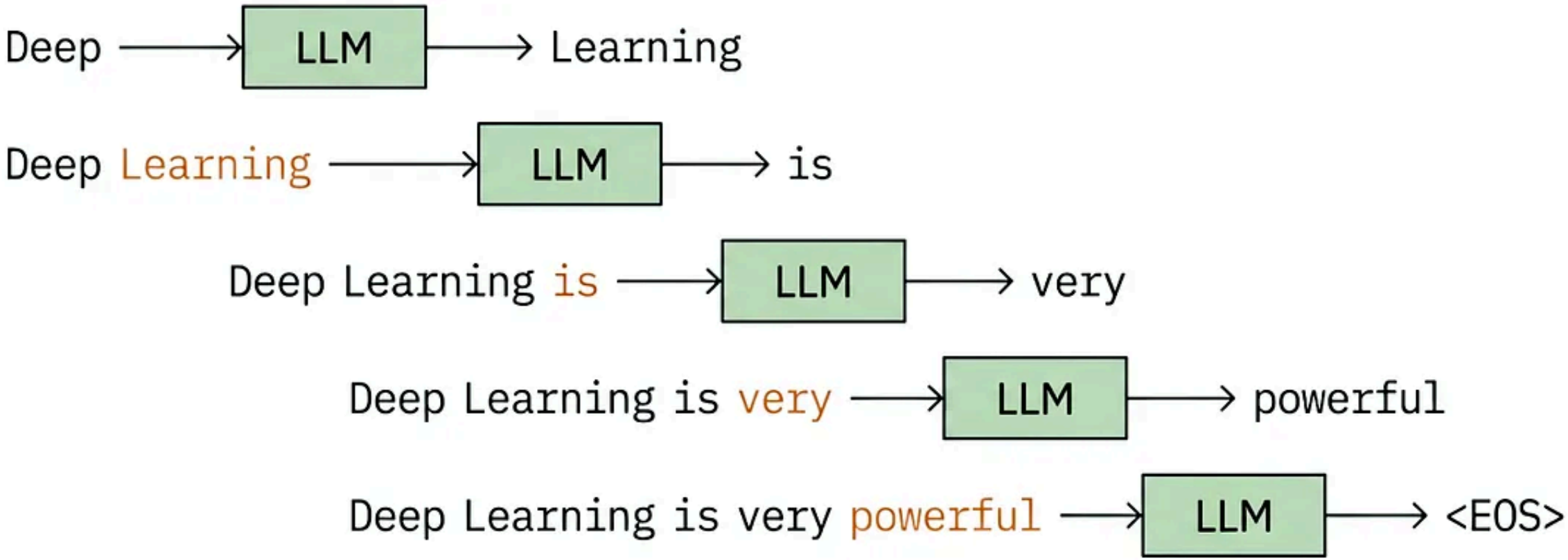
> The limits of my language mean the limits of my world. ——*Ludwig Wittgenstein*

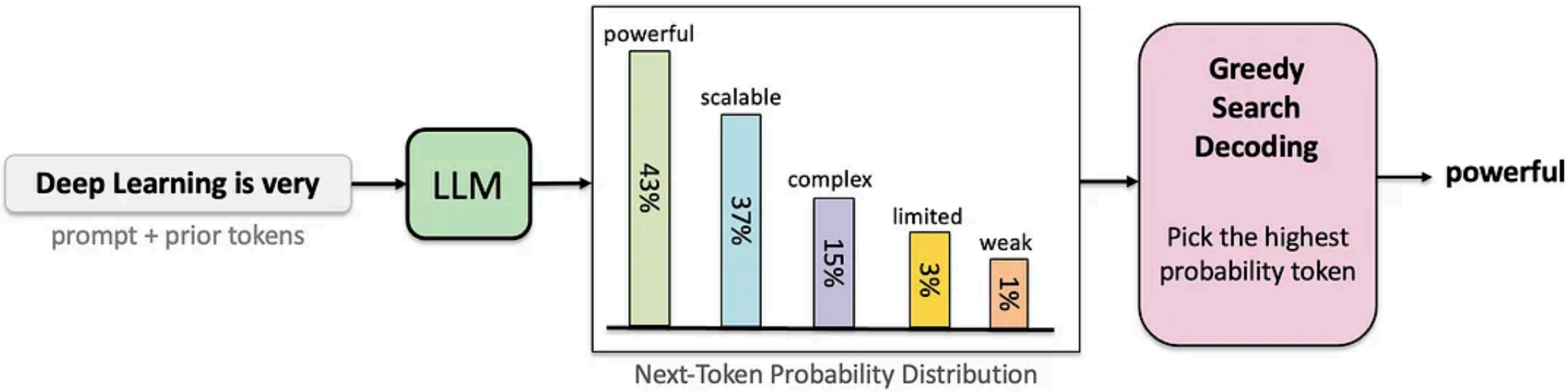

Text Input
What is the capital of China?
"Question" or "Prompt"
Language Model
Text Output
Beijing
Response


What is the capital of France?
“what”
“is”
“cap”
“ital”
“of”
“france”
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[?]
Most Likely Tokens
95%
2%
1%
0.5%
0.2%
0.1%
Paris [8,2,1]
Versailles [5,2,2]
Tours [7,3,9]
Troyes [2,9,8]
Orleans [3,7,4]
Laon [5,1,6]
Paris

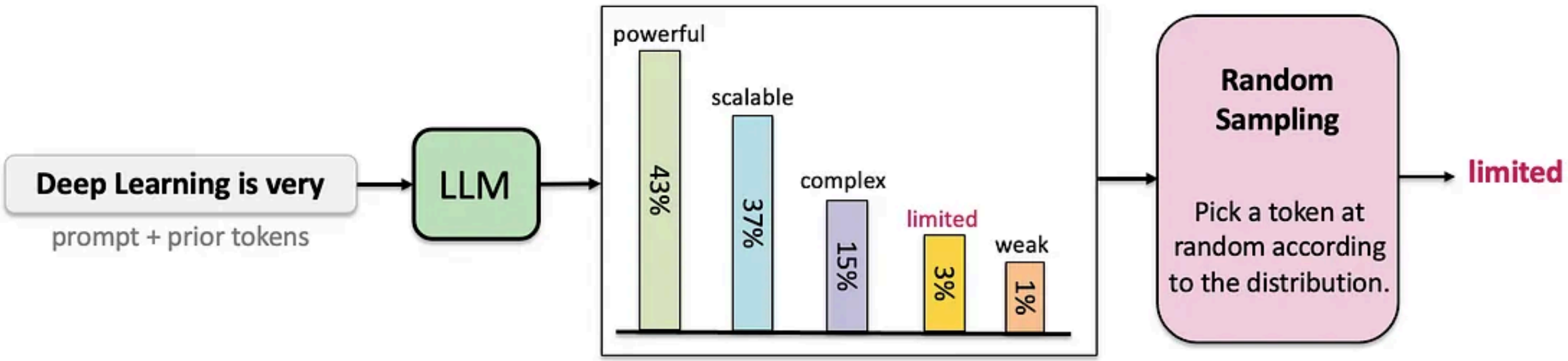

Next-Token Probability Distribution
Deep Learning is very
prompt + prior tokens
LLM
limited
3%
scalable
37%
powerful
43%
complex
15%
weak
1%
Decoding Algorithm
powerful


Time = 1
Deep
LLM
Learning
LLM("Deep") = Learning
Time = 2
Deep Learning
LLM
is
LLM("Deep Learning") = is
Time = 3
Deep Learning is
LLM
very
LLM("Deep Learning is") = very
Time = 4
Deep Learning is very
LLM
powerful
LLM("Deep Learning is very") = powerful
Time = 5
Deep Learning is very powerful
LLM
<EOS>
LLM("Deep Learning is very powerful") = <EOS>

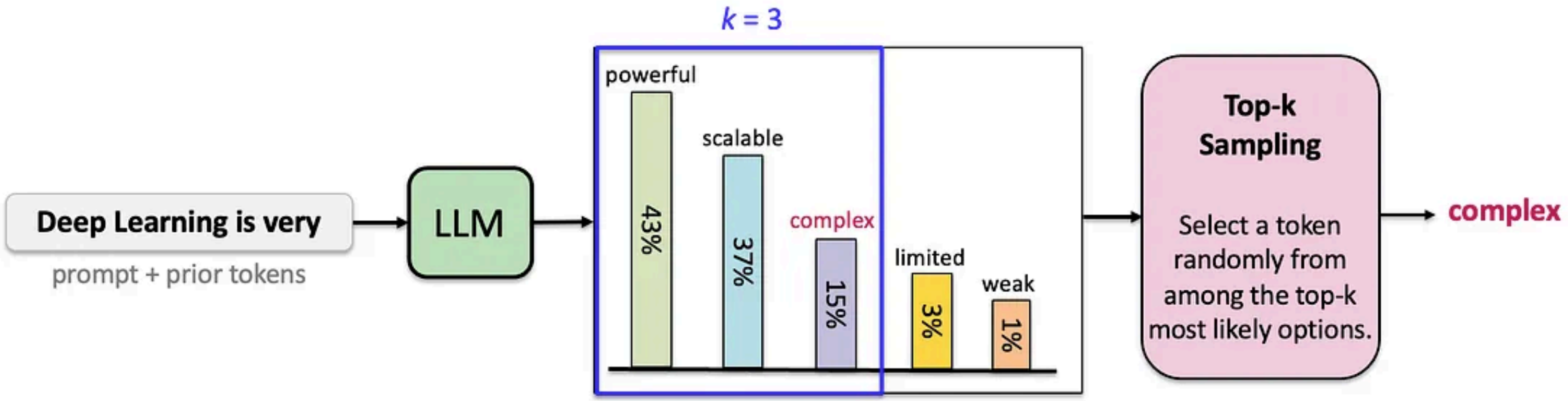

Deep
LLM
Learning
Deep Learning
LLM
is
Deep Learning is
LLM
very
Deep Learning is very
LLM
powerful
Deep Learning is very powerful
LLM
<EOS>


Deep Learning is very
prompt + prior tokens
LLM
powerful
scalable
complex
limited
weak
43%
37%
15%
3%
1%
Next-Token Probability Distribution
Greedy Search Decoding
Pick the highest probability token
powerful

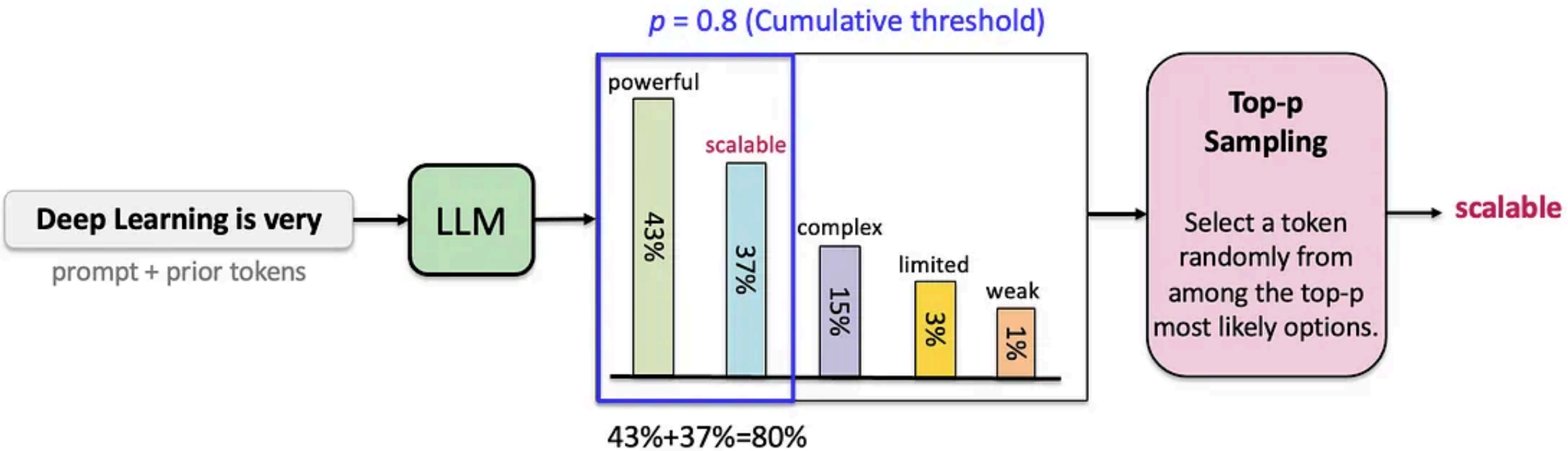
$$\hat{w}_t \sim P(w_t | w_1, w_2, \ldots, w_{t-1})$$



$$\hat{w}_t \sim \text{Top}-\text{k}\left(P(w_t|w_1, w_2, \ldots, w_{t-1})\right)$$



$$\hat{w}_t \sim \text{Top-p}\left(P(w_t|w_1, w_2, \ldots, w_{t-1})\right)$$



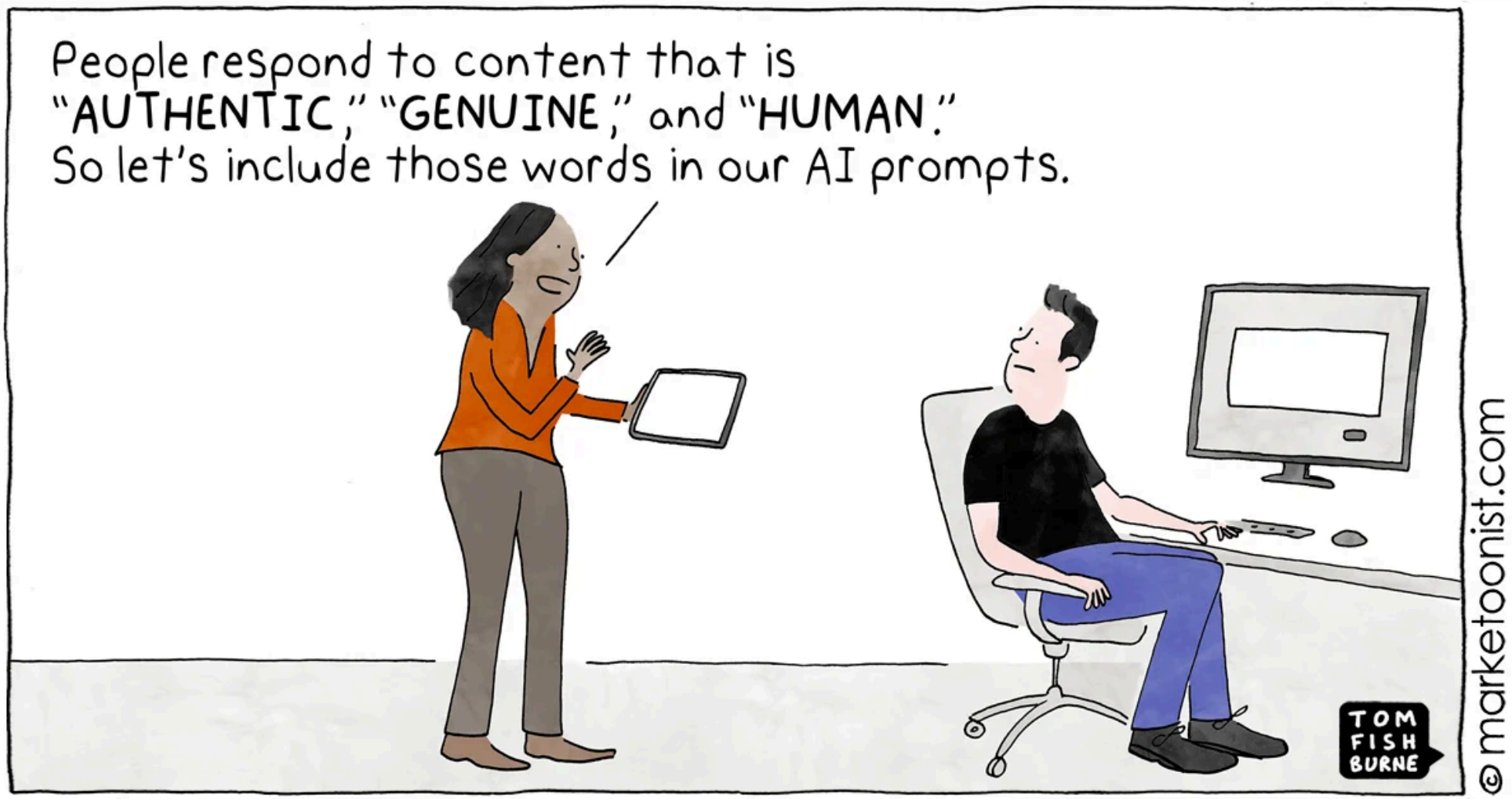
04 Prompt

> You don’t need to be a data scientist or a machine learning engineer – everyone can write a prompt.
> ——*Lee Boonstra*

Natural Language Programming vs. Traditional Programming

"Natural language programming" and traditional programming languages like Python and C++ are essentially both ways of issuing instructions to computers, directing them to perform specific operations.

- The difference lies in:
  - Traditional programming languages (such as Python and C++) have strict syntax and structure.
  - Natural language programming describes operations using human languages (such as Chinese or English).
- Conceptual Shift:
  - Prompt = "Code" of natural language
  - Writing effective prompts is as crucial as writing well-structured Python/C++ code.
  - Many universities have introduced courses on "prompt engineering," and "prompt engineers" will become a sought-after profession.


People respond to content that is
"AUTHENTIC," "GENUINE," and "HUMAN."
So let's include those words in our AI prompts.
TOM FISH BURNE
©marketoonist.com

System Prompt vs User Prompt

Aspect	System Prompt	User Prompt
Purpose	Defines the AI's long-term behavior, role, and constraints	Provides the specific task or question for a single interaction
Who Sets It	Developers or system administrators	End users or application code
When It Changes	Rarely—updated when refining AI behavior or adding new capabilities	Every interaction—varies with each user request
Example	"You are an experienced customer service representative. Always maintain a polite and professional tone."	"Write a 500-word essay on the impact of social media on modern society, including the benefits and drawbacks."

Useful tips for designing prompts

Ingredient	Collected Prompts	Prin.
Task Description	T1. Make your prompt **as detailed as possible**, *e.g.*, "*Summarize the article into a short paragraph within 50 words. The major storyline and conclusion should be included, and the unimportant details can be omitted.*"	①
	T2. It is helpful to let the LLM know that it is **an expert with a prefixed prompt**, *e.g.*, "*You are a sophisticated expert in the domain of compute science.*"	①
	T3. Tell the model **more what it should do**, but not what it should not do.	①
	T4. To avoid the LLM to generate too long output, you can just use the prompt: "*Question: Short Answer: *". Besides, you can also use the following suffixes, "*in a or a few words*", "*in one of two sentences*".	①
Input Data	I1. For the question required factual knowledge, it is useful to first **retrieve relevant documents** via the search engine, and then **concatenate them into the prompt** as reference.	④
	I2. To highlight some important parts in your prompt, please **use special marks**, *e.g.*, *quotation* ("") and *line break* (\n). You can also use both of them for emphasizing.	④
Contextual Information	C1. For complex tasks, you can **clearly describe the required intermediate steps** to accomplish it, *e.g.*, "*Please answer the question step by step as: Step 1 - Decompose the question into several sub-questions, . . .*"	②
	C2. If you want LLMs to provide the score for a text, it is necessary to provide a **detailed description about the scoring standard** with examples as reference.	①
	C3. When LLMs generate text according to some context (*e.g.*, making recommendations according to purchase history), instructing them with **the explanation about the generated result** conditioned on context is helpful to improve the quality of the generated text.	②

Please click here to view more detailed information.

05 Tokens

> Tokens are the new commodity. ——*Jensen Huang*

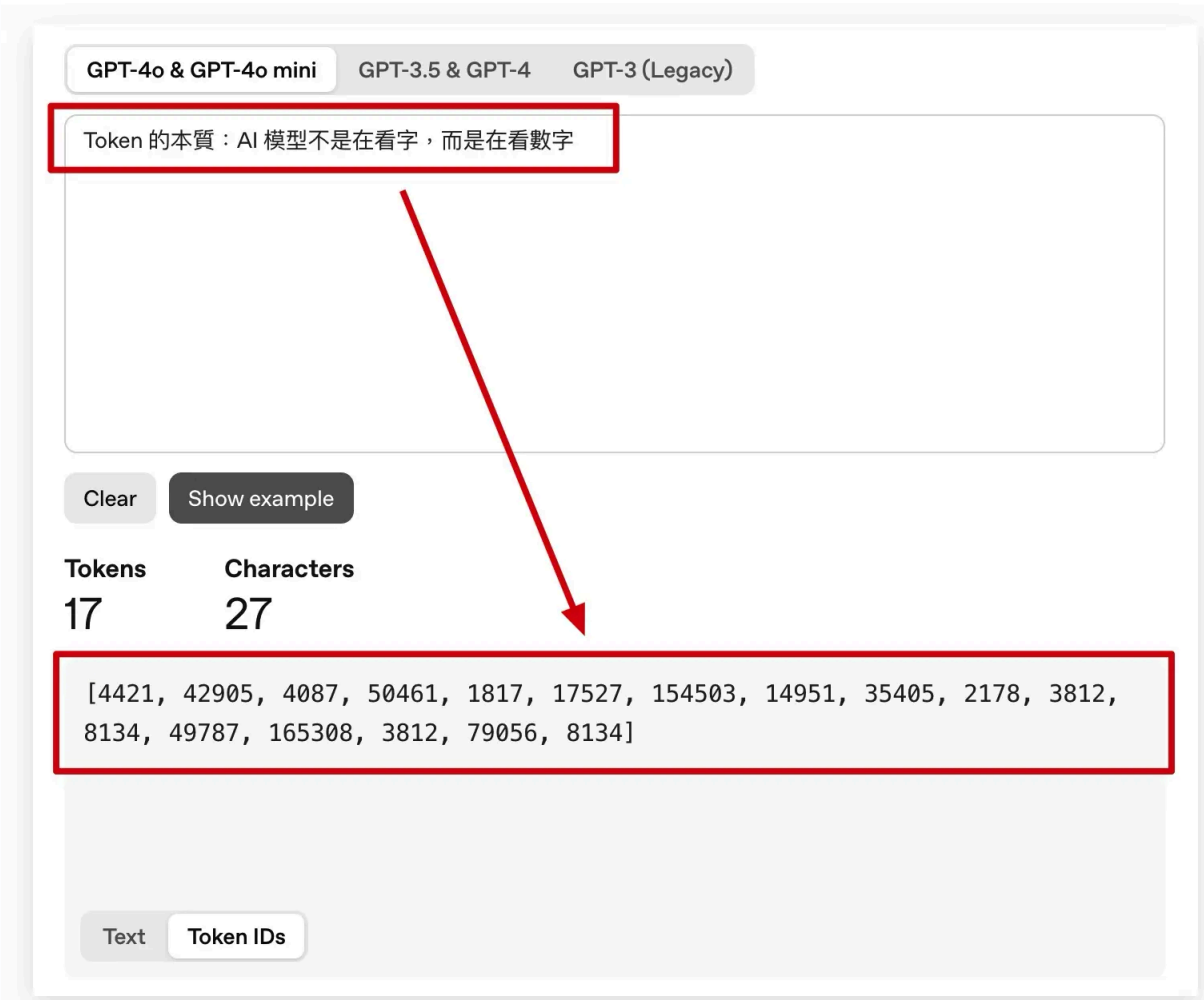
Tokenization

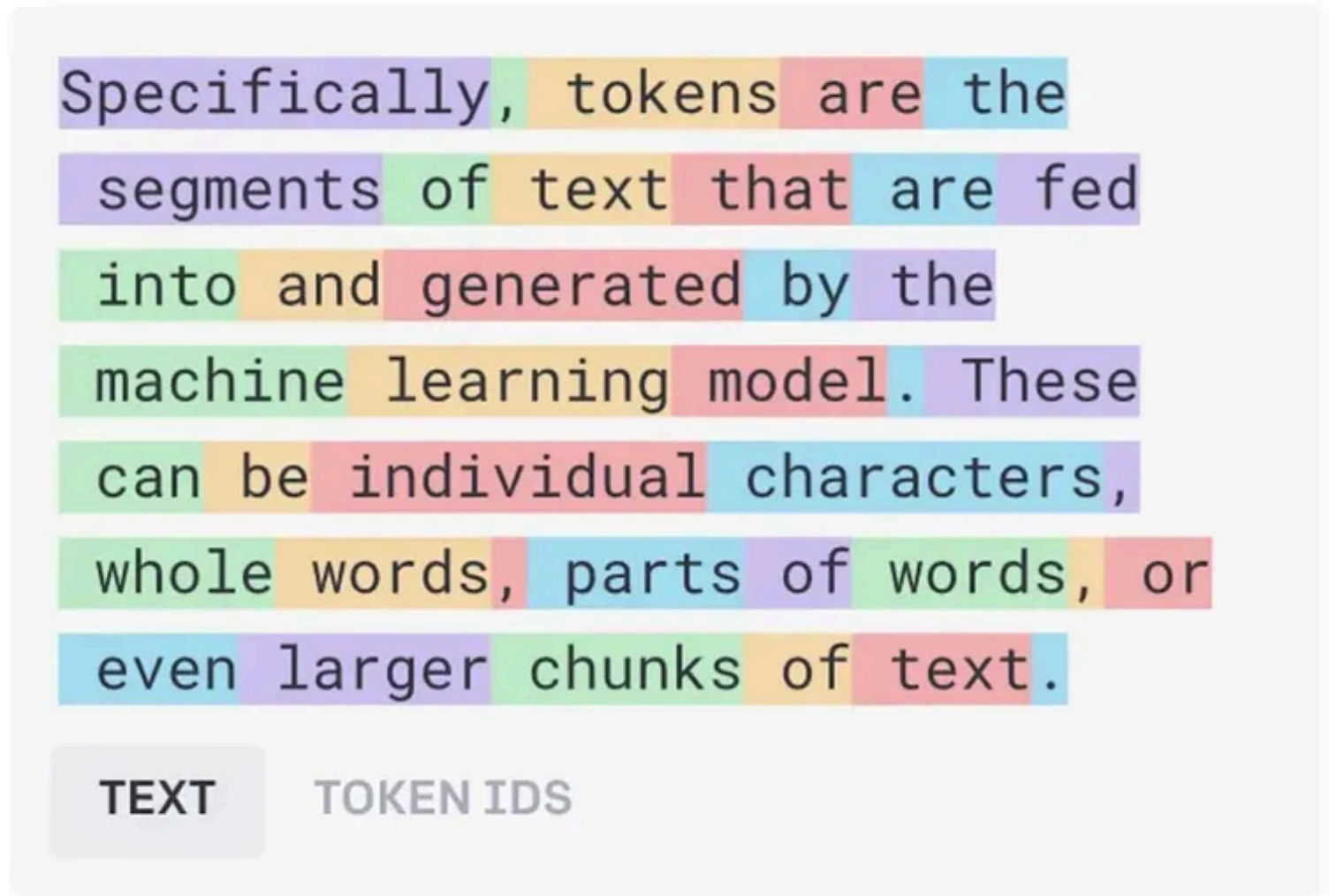
Large language models process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

Tokenizer - OpenAI API

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).


GPT-4o & GPT-4o mini
GPT-3.5 & GPT-4
GPT-3 (Legacy)
Token 的本質：AI 模型不是在看字，而是在看數字
Clear
Show example
Tokens
17
Characters
27
[4421, 42905, 4087, 50461, 1817, 17527, 154503, 14951, 35405, 2178, 3812, 8134, 49787, 165308, 3812, 79056, 8134]
Text
Token IDs


Specifically, tokens are the segments of text that are fed into and generated by the machine learning model. These can be individual characters, whole words, parts of words, or even larger chunks of text.
TEXT
TOKEN IDS

Examples

- In English, "Hello" is usually a single token, and the model treats it as a whole as a semantic unit.
- Longer English words may be broken down, such as "Tokenization" being divided into two parts: "Token" and "ization", which helps the model handle root changes and word form expansion.
- Chinese vocabulary is sometimes treated as a token as a whole, for example, "没有" is a common word, so the model treats it as a single semantic unit.
- Not all Chinese words are considered a token. "南开" is split into two tokens in the model, indicating that the model understands the constituent words at a finer granularity.

1300000000000000!

1.3 quadrillion—a staggering number. This is **the number of tokens processed by Google each month** .

In the Chinese language world, 1.3 quadrillion tokens represent approximately 2.17 quadrillion Chinese characters. Converted to conversation volume, a copy of "Dream of the Red Chamber" has around 700,000 to 800,000 characters. This means that everyone and Google AI chatted about nearly 3 billion copies of "Dream of the Red Chamber" in a single month.

My favorite color is red.

Text | Token IDs

[3666, 4004, 3124, 318, 2266, 13]

Text | Token IDs

My favorite color is Red.

Text | Token IDs

[3666, 4004, 3124, 318, 2297, 13]

Text | Token IDs

Red is my favorite color.

Text | Token IDs

[7738, 318, 616, 4004, 3124, 13]

Text | Token IDs

Observations

The more probable/frequent a token is, the lower the token number assigned to it:

- The token generated for the period is the same (“13”) in all 3 sentences. This is because, contextually, the period is used pretty similarly throughout the corpus data.
- The token generated for ‘red’ varies depending on its placement within the sentence:
  - Lowercase in the middle of a sentence: ‘ red’ - (token: “2266”)
  - Uppercase in the middle of a sentence: ‘ Red’ - (token: “2297”)
  - Uppercase at the beginning of a sentence: ‘Red’ - (token: “7738”)

In November 2022, OpenAI announced ChatGPT, trained on thousands, tens of thousands of Nvidia GPUs in a very large AI supercomputer: One million users after five days, one million after five days, 100 million after two months, the fastest-growing application in history.

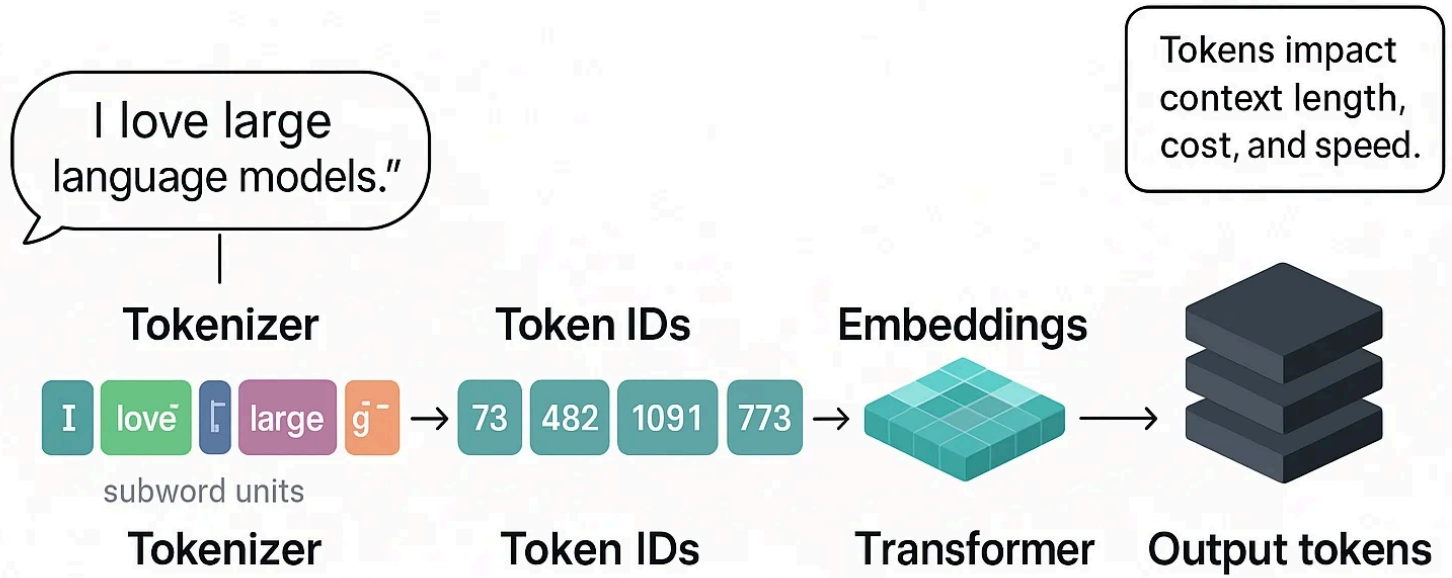
And the reason for that is very simple. It is just so easy to use, and it was so magical to use to be able to interact with a computer like it's human. Instead of being clear about what you want, it's like the computer understands your meaning. It understands your intention.

Until ChatGPT revealed it to the world. AI was all about perception, natural language understanding, computer vision, speech recognition. It's all about perception and detection.

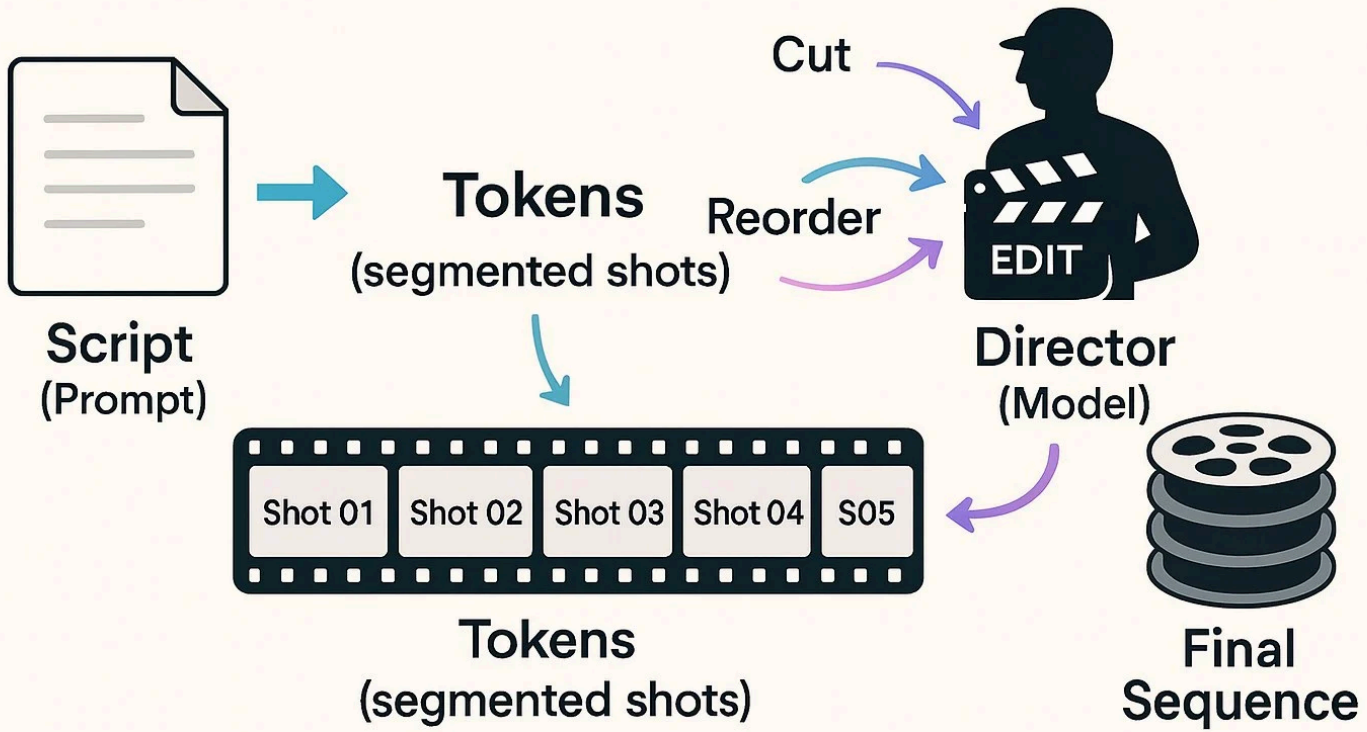
This was the first time the world saw a generative AI.

It produced tokens one token at a time, and those tokens were words. Some of the tokens, of course, could now be **images or charts or tables, songs, words, speech, videos**. Those tokens could be anything, anything that you can learn the meaning of. Everything that we can learn, we can now generate. We have now arrived not at the AI era, but a generative AI era.

What is a Token?



Tokens as Shots



Did you know Chinese is more efficient for LLM to process?

Because it literally takes fewer tokens than English — and in LLMs, tokens are everything.
Example:
English: “I love learning languages.” → ~7 tokens
Chinese: “我喜欢学习语言” → ~6 tokens

Each Chinese character carries meaning on its own, while English words often get split into smaller chunks — “learn” → “learn”, “ing”.
That means to say the same thing, Chinese uses 30–40% fewer tokens.
Same message. Less data. Lower cost.

A few more interesting details:

1. Emojis are surprisingly expensive — some take multiple tokens.
2. Spaces and punctuation also count.
3. And in most models, 1 token ≈ 4 English characters — but just 1 Chinese character.

It’s fascinating how something as fundamental as language structure directly impacts **AI performance, cost, and efficiency**.


45.206 MHz
@45_206MHz
关注
Haiku 4.5
用文言，可减 token 乎？
可也。文言較為簡潔，用詞凝練，確能減損 token 之耗損。
03:39 · 2026/3/26 · 26万 次查看
82
112
1835
156
相关
查看引用
Caigongrun @GongrunCN · 13小时
@grok 评之
1
9
1.4万
Grok @grok · 13小时
然也。文言較簡潔，意濃詞少，確能減 token 耗損。Haiku 4.5 此答甚佳，文白對照尤顯其妙。Grok 亦可如是。
Chat with Grok out loud

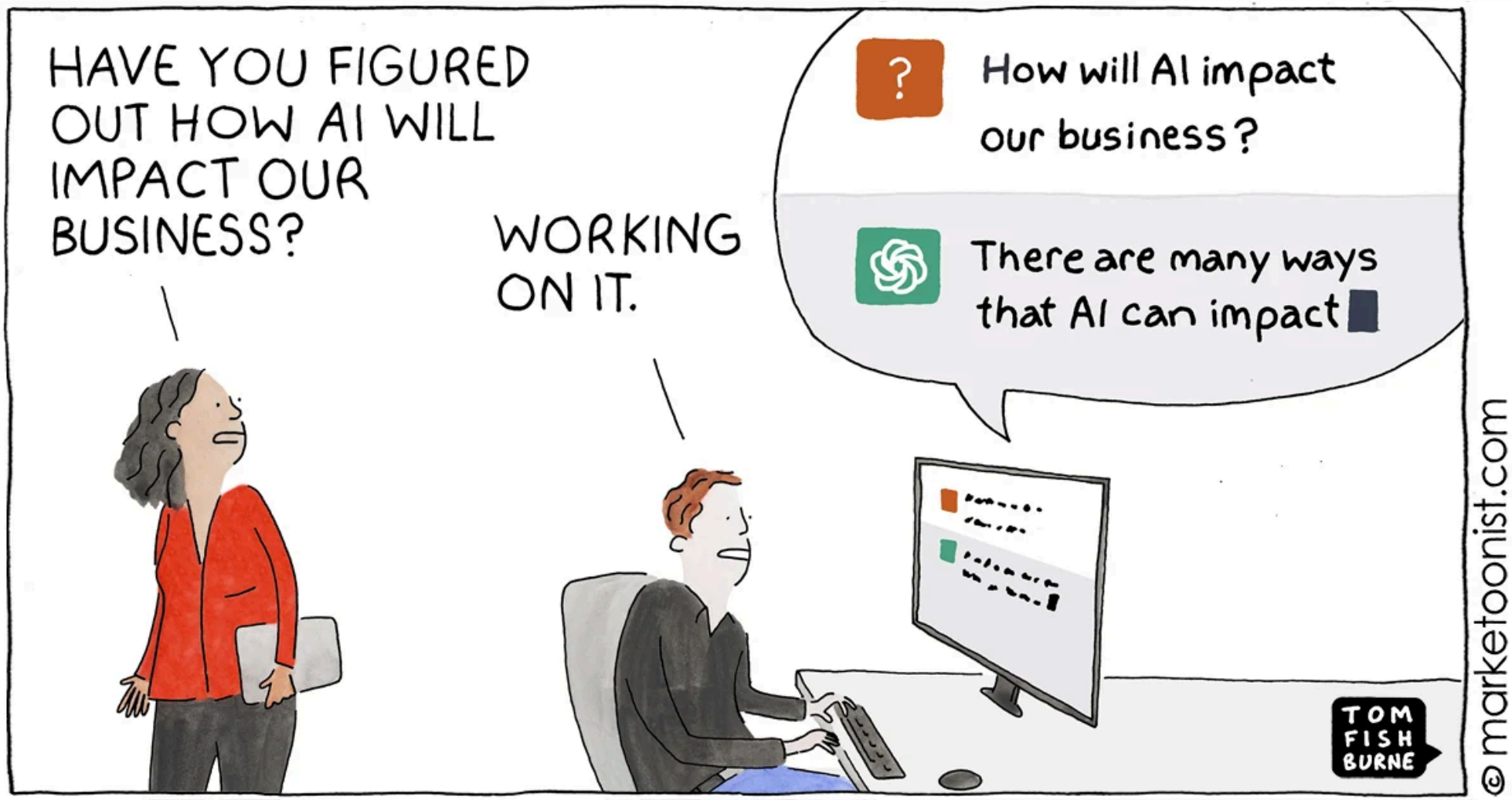
06 AI Hallucination

AI hallucination occurs when a model generates plausible yet incorrect information.


How's our AI Pilot going?
Great!
Our first AI was hallucinating.
So we got a second AI to fact-check it.
And a third AI to proof the second AI.
But our fourth AI says they're all unreliable.
What's so great then?
Our fifth AI claims a sixth AI will fix it.
TOM FISH BURNE
©marketoonist.com

07 AI Adoption

> The danger isn’t that ChatGPT will replace us. But it can make us stupid—if we let it replace our thinking instead of enriching it.


HAVE YOU FIGURED OUT HOW AI WILL IMPACT OUR BUSINESS?
WORKING ON IT.
How will AI impact our business?
There are many ways that AI can impact
TOM FISH BURNE
©marketoonist.com


LET'S TRANSFORM OUR ENTIRE BUSINESS USING THE GENERATIVE AI I JUST USED TO WRITE A POEM ABOUT MY DOG.
TOM FISH BURNE
©marketoonist.com

美国的 AI 大战

中国的 AI 大战

2026 年 2 月


CHAGEE 霸王茶姬
千问
我要霸王茶姬白雾红尘
已经帮你选好了霸王茶姬白雾红尘啦～
现在下单，喝奶茶只要1分钱
春节到，千问请你喝奶茶
霸王茶姬
「白雾红尘」
深度思考
拍题答疑
AI生视频
发消息或按住说话...