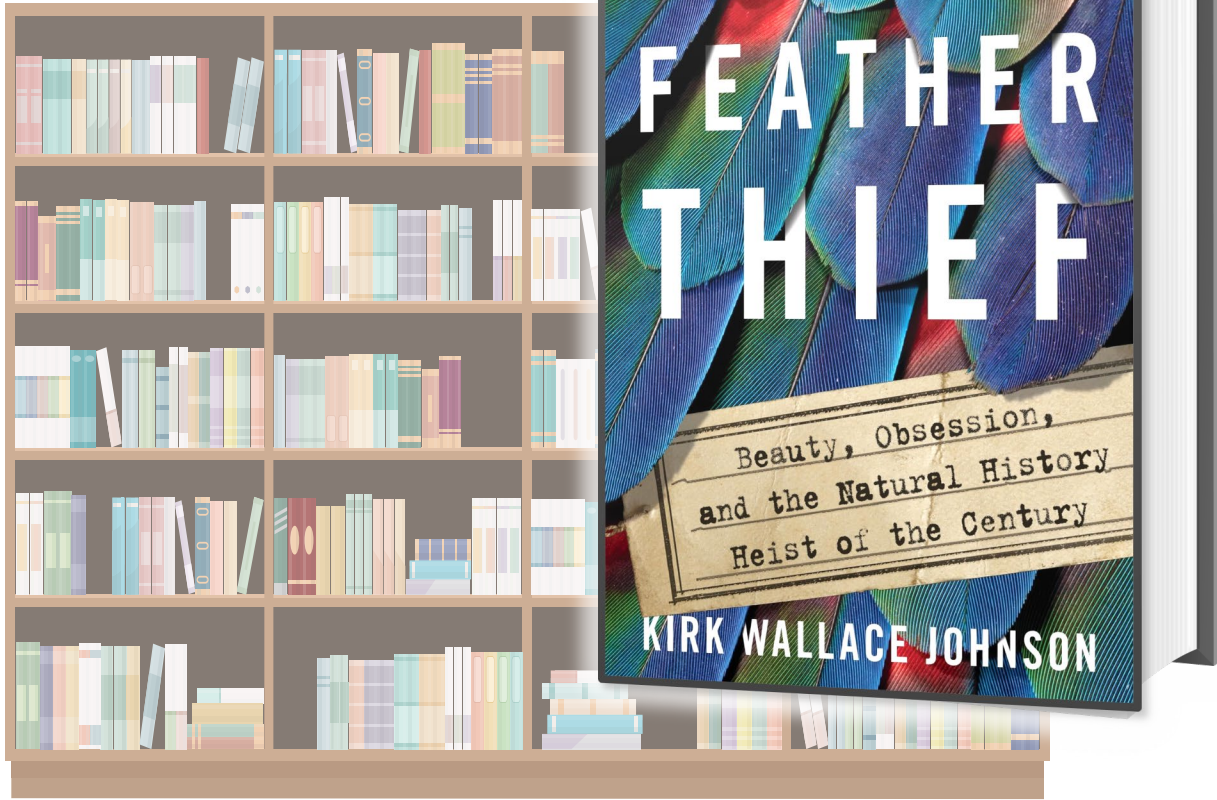


TECHNOLOGY TUTORIAL



ANTHROPIC



Bartz et al v. Anthropic
January 30, 2025

Plaintiffs' Presentation
Justin Nelson, Susman Godfrey LLP

Complaint Alleges Anthropic Used Copyrighted Books Without Permission

THE PILE

THE FEATHER THIEF
Beauty, Obsession,
and the Natural History
Heist of the Century
KIRK WALLACE JOHNSON

Download Now

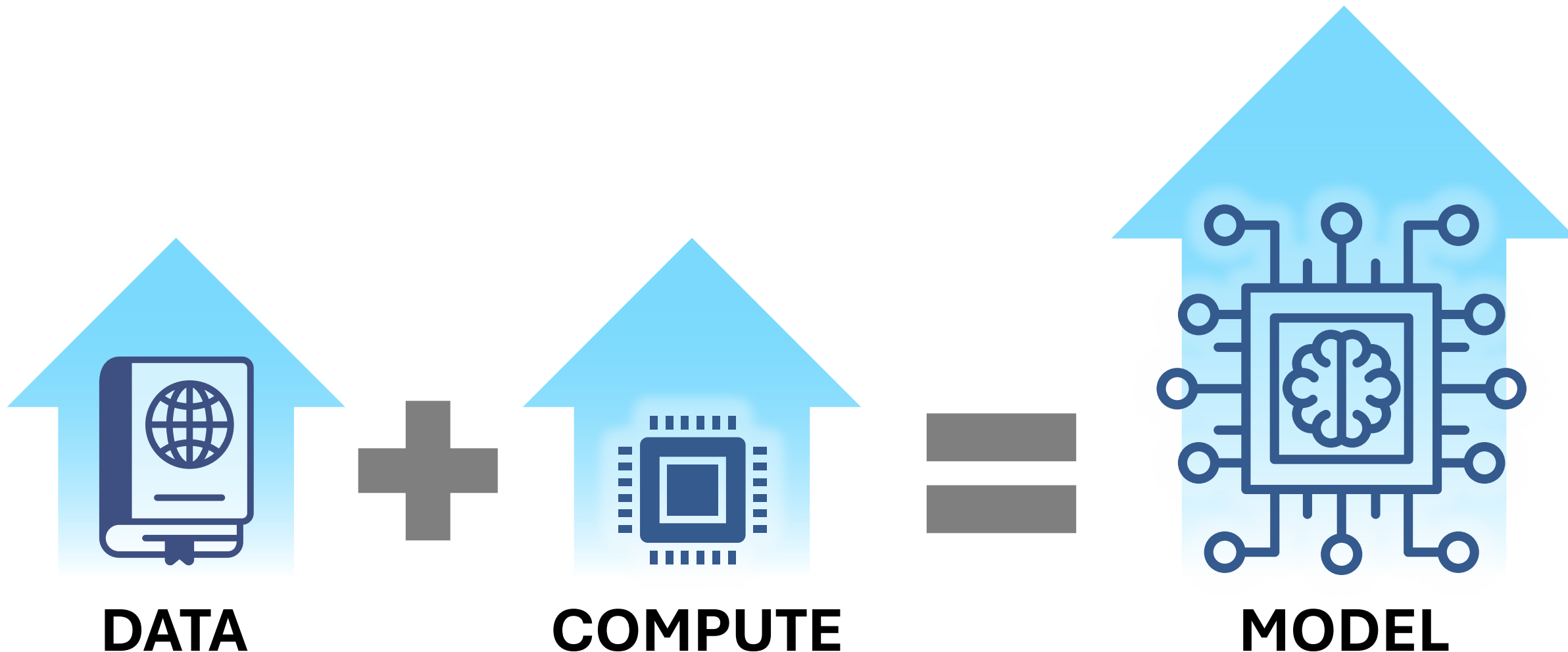
ANTHROPIC



A.I. is really H.I. — “Human Imitation”



The Basic Formula



Better data, Better compute, Better model

LLM, or “Large Language Model”

Large

Language

Model

Trained on a
large dataset

“Data is the Fossil Fuel of AI”—Because it is a Finite Resource

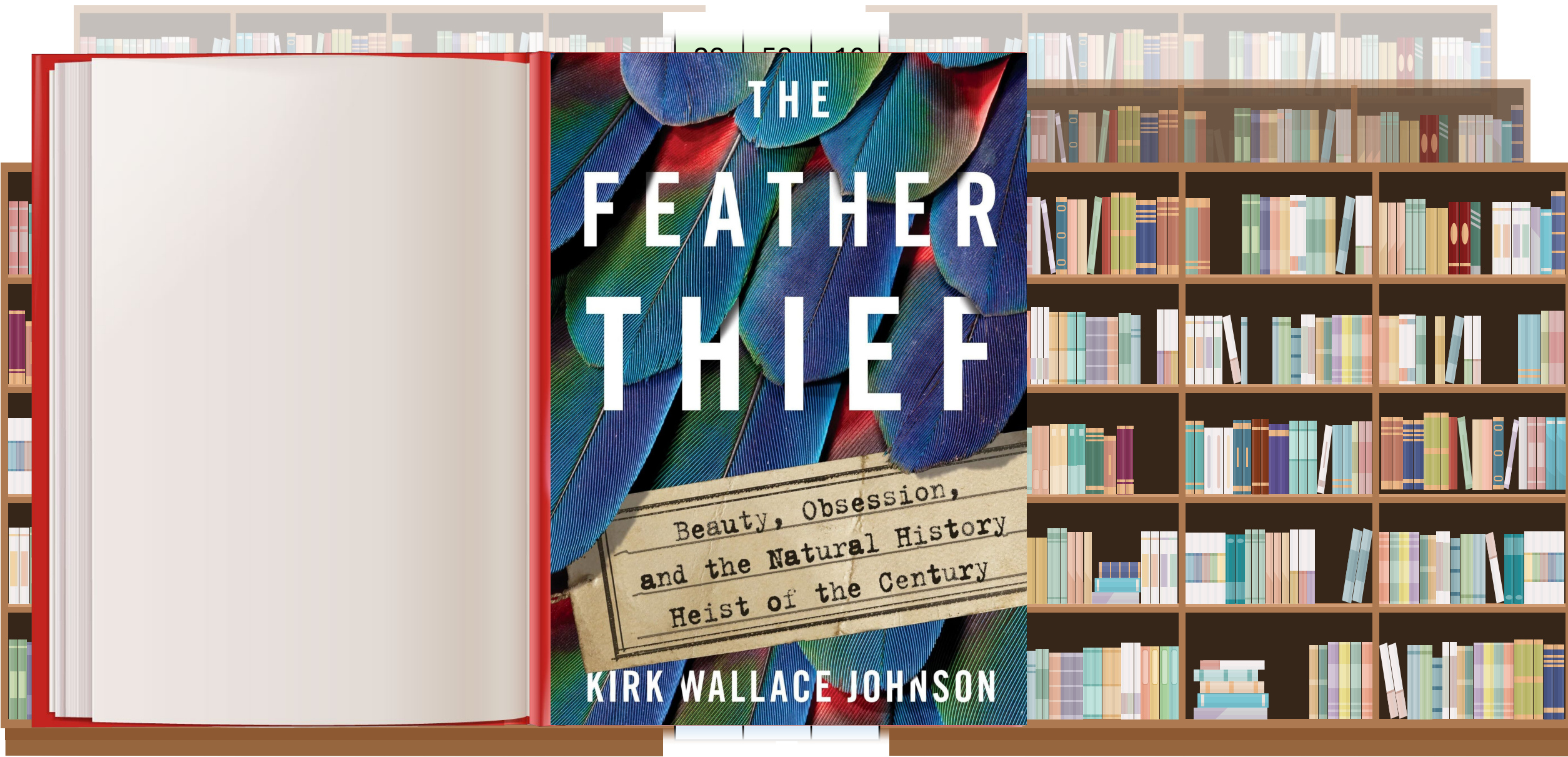


Ilya Sutskever
Former Chief Scientist of OpenAI

“Data is the fossil fuel of AI. It was like, created somehow. And now we use it.”



What is an LLM?



High Sierra: A Mystery Novel

JC Write me the opening two pages of a mystery novel, set on the John Muir trail in the 1970s.



← High Sierra Mystery - Opening Pages ×

The first sign something was wrong was the ravens. Three of them, wheeling in tight circles above the treeline, their harsh calls echoing off the granite walls of Evolution Valley. Rangers learn to read the birds, especially the ravens. They're nature's undertakers, and where they gather, death usually follows.

I shouldered my pack higher and picked up the pace. The September air was thin at 10,000 feet, and my lungs burned with each step. Twenty-three years of patrolling these trails, and

JC Write me the opening two pages of a mystery novel, set on the John Muir trail in the 1970s.

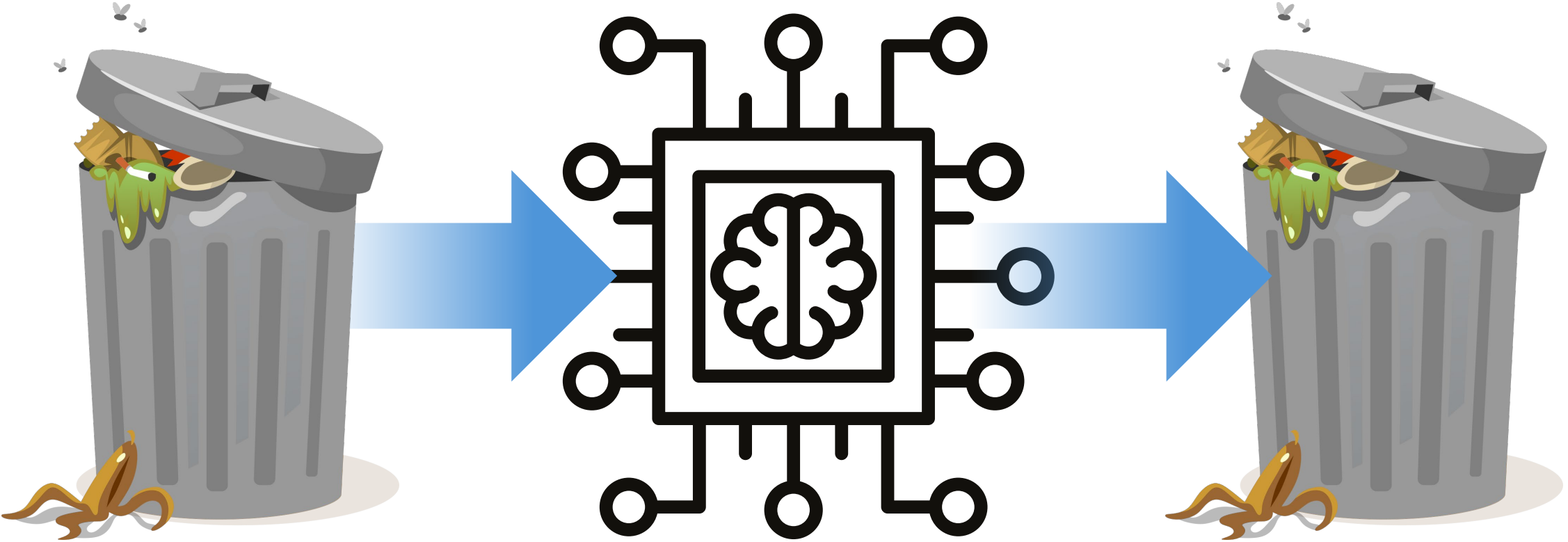
"We've got a situation at McClure Meadow. Hiker reported finding an abandoned campsite. Blood on a sleeping bag. No sign of the occupant. Nearest backup is four hours out. Over."

My free hand instinctively went to the holster at my hip. In '76, most rangers still carried revolvers – simple, reliable Smith & Wessons that could handle the abuse of backcountry patrol. We rarely needed them for anything more threatening than the occasional rattler or injured deer. But there had been changes lately. More drug runners using the remote trails. More desperate people fleeing into the wilderness, carrying their troubles with them.

"Copy that. I'm about thirty minutes out. Tell them not to touch anything. Over."

"Wilco. Be careful up there. Out."

The ravens had disappeared behind the ridgeline. I checked my watch – 2:47 PM. At this altitude, darkness would come early, and I needed to document whatever I found before



Presentation Roadmap

1 Background of LLMs

2 How an LLM is trained and how it works

3 Why books are so important to LLMs

What is a Large Language Model?

Large

Trained on a
large dataset

Language

Works with
human languages

Model

Type of
statistical model

First Neural Networks Proposed in the 1950s

Psychological Review
**THE PERCEPTRON: A PROBABILISTIC MODEL FOR
 INFORMATION STORAGE AND ORGANIZATION
 IN THE BRAIN¹**

F. ROSENBLATT
Cornell Aeronautical Laboratory

If we are eventually to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking, we must first have answers to three fundamental questions:

1. How is information about the physical world sensed, or detected, by the biological system?
2. In what form is information stored, or remembered?
3. How does information contained in storage, or in memory, influence recognition and behavior?

The first of these questions is in the province of sensory physiology, and is the only one for which appreciable understanding has been achieved. This article will be concerned primarily with the second and third questions, which are still subject to a vast amount of speculation, and where the few relevant facts currently supplied by neurophysiology have not yet been integrated into an acceptable theory.

With regard to the second question, two alternative positions have been maintained. The first suggests that storage of sensory information is in the form of coded representations or images, with some sort of one-to-one mapping between the sensory stimulus

¹The development of this theory has been carried out at the Cornell Aeronautical Laboratory, Inc., under the sponsorship of the Office of Naval Research, Contract Nonr-238(00). This article is primarily an adaptation of material reported in Ref. 13, which constitutes the first full report on the program.

***THE PERCEPTRON: A
 PROBABILISTIC MODEL FOR
 INFORMATION STORAGE
 AND ORGANIZATION IN THE
 BRAIN***

By Rosenblatt, 1958

**THE PERCEPTRON: A PROBABILISTIC MODEL FOR
 INFORMATION STORAGE AND ORGANIZATION
 IN THE BRAIN¹**

F. ROSENBLATT

Cornell Aeronautical Laboratory

With regard to the second question, two alternative positions have been maintained. The first suggests that storage of sensory information is in the form of coded representations or images, with some sort of one-to-one mapping between the sensory stimulus and the stored pattern. According to this hypothesis, if one understood the code or "wiring diagram" of the nervous system, one should, in principle, be able to discover exactly what an organism remembers by reconstructing the original sensory patterns from the "memory traces" which they have left, much as we might develop a photographic negative, or translate the pattern of electrical charges in the "memory" of a digital computer.

What Claude Says It Can Do

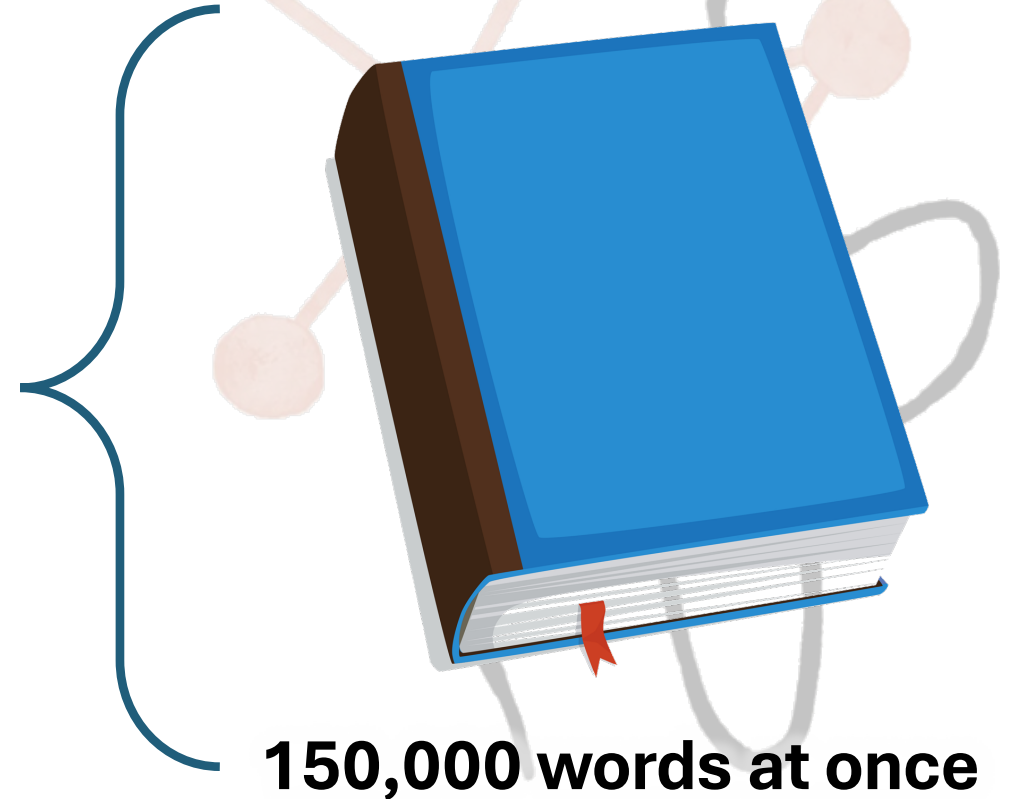
* Good morning

A How does *A Tale of Two Cities* start?

* It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way--in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

ANTHROPIC

“everything from a text message to a screenplay to a novel.”



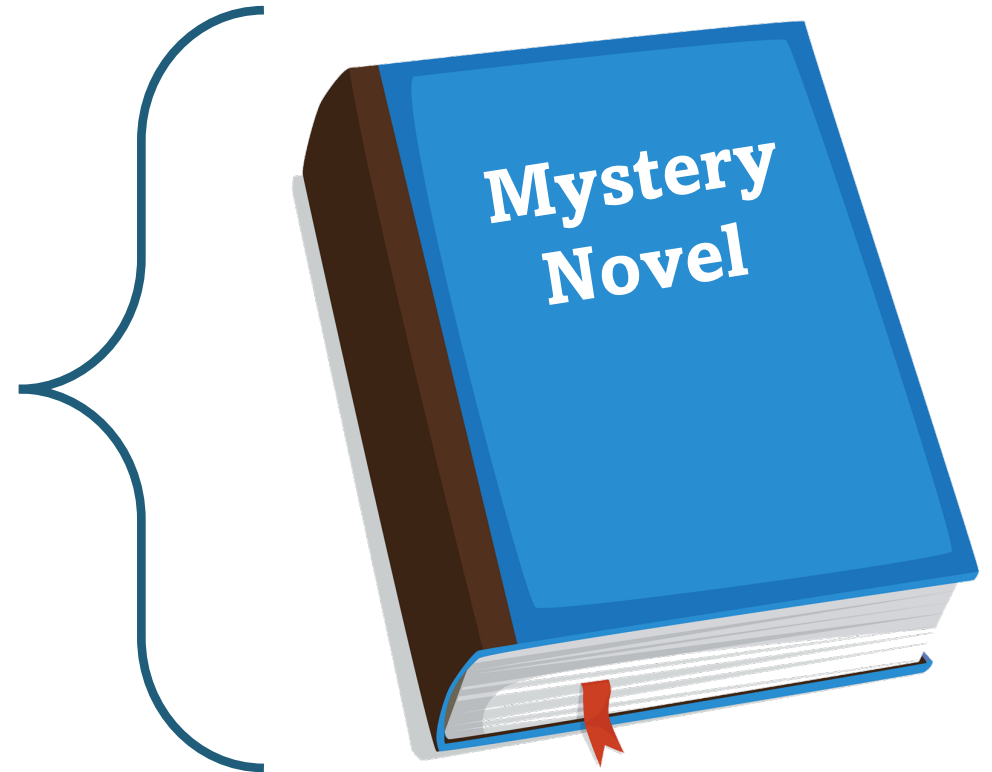
150,000 words at once

What is an LLM?



The murderer is _____.

All 500 pages at once




LLMs have mined from human works



the
decoder

DE  

AI research Jul 11, 2023  Update

GPT-4 architecture, datasets, costs and more leaked



Midjourney prompted by THE DECODER



Maximilian




Max is managing editor at THE DECODER. As a trained philosopher, he writes about consciousness, AI, and the future of technology. He pretends to be a philosopher.

Profile

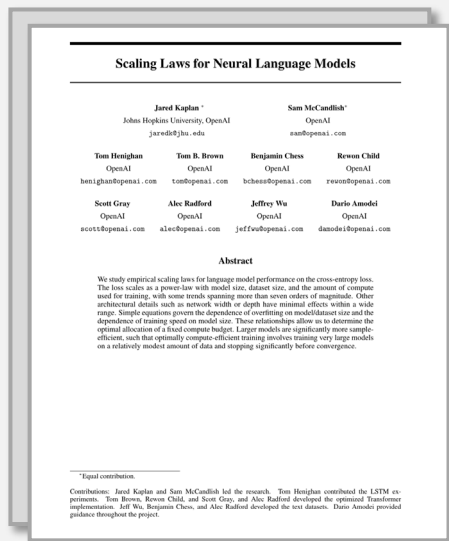
[E-Mail](#)

The latest LLMs have trained on ~9.75 trillion words

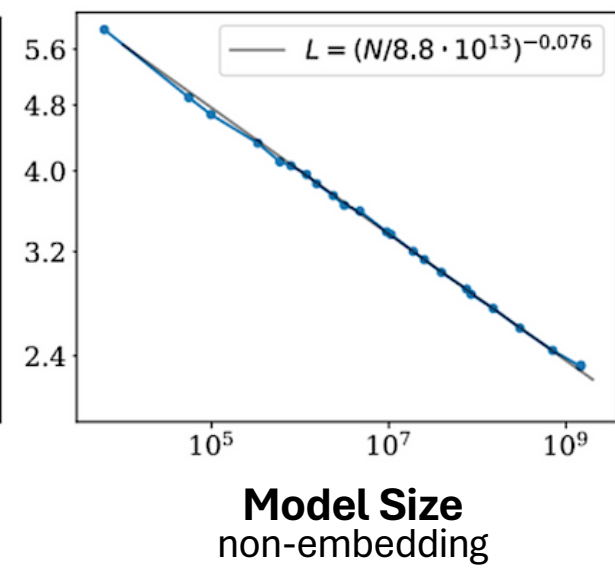
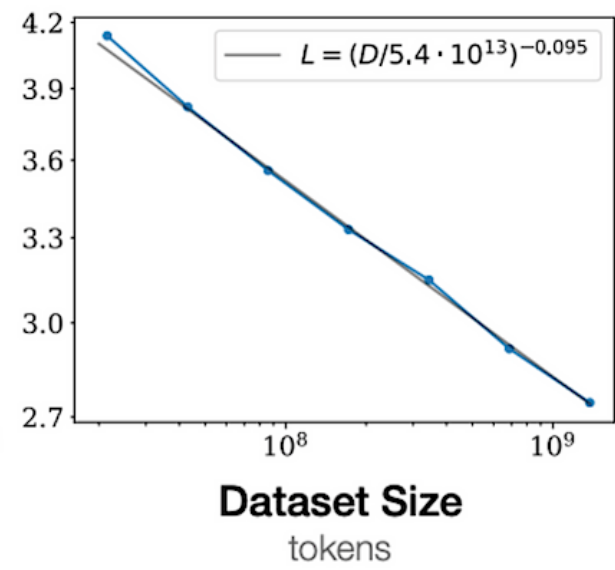
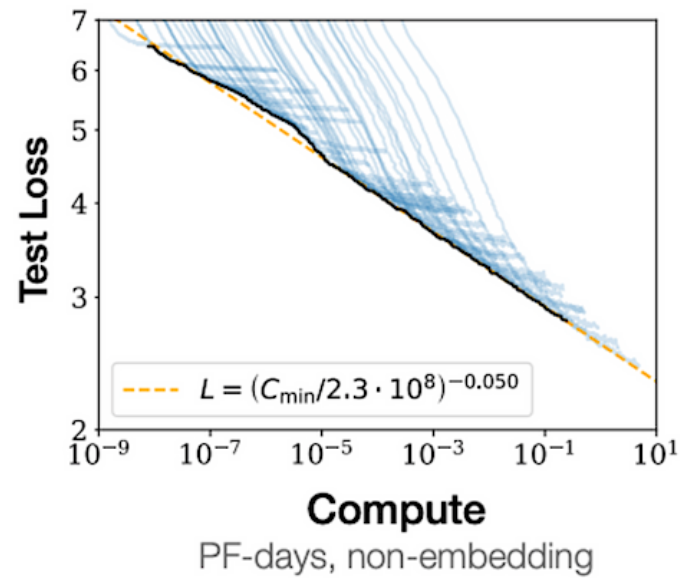
 Content

 Summary

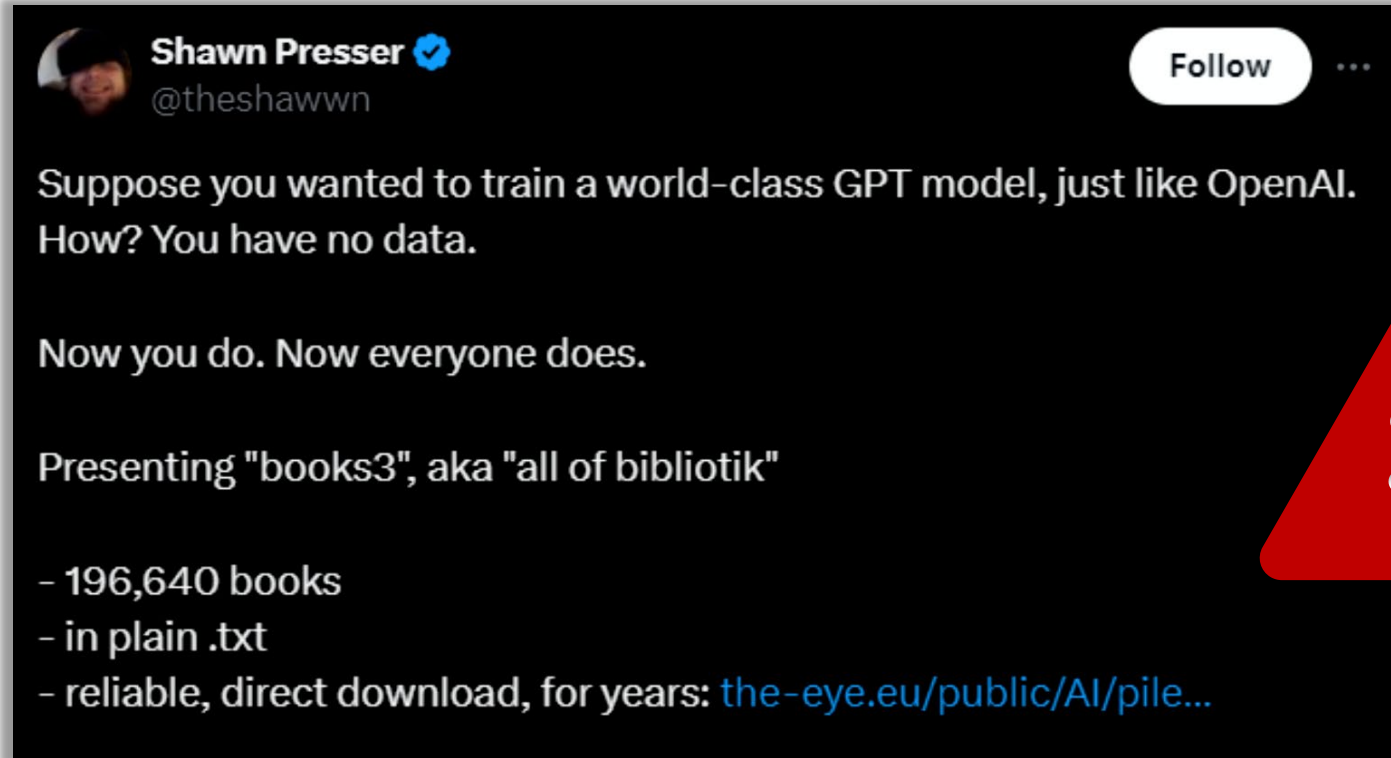
Scaling Laws




Scaling Laws for Neural Language Models, Kaplan et al, 2020



“Large” Data Sets



Shawn Presser 
@theshawwn Follow ...

Suppose you wanted to train a world-class GPT model, just like OpenAI. How? You have no data.

Now you do. Now everyone does.

Presenting "books3", aka "all of bibliotik"

- 196,640 books
- in plain .txt
- reliable, direct download, for years: the-eye.eu/public/AI/pile...



Presentation Roadmap

1 Background of LLMs

2 **How an LLM is trained and how it works**

3 **Why books are so important to LLMs**

How an LLM is Trained

1

Acquiring the Data

- Scraped from across the internet, from public and/or pirating sites
- Books highly valued for diversity, long-context reasoning, high-quality writing, and more.

2

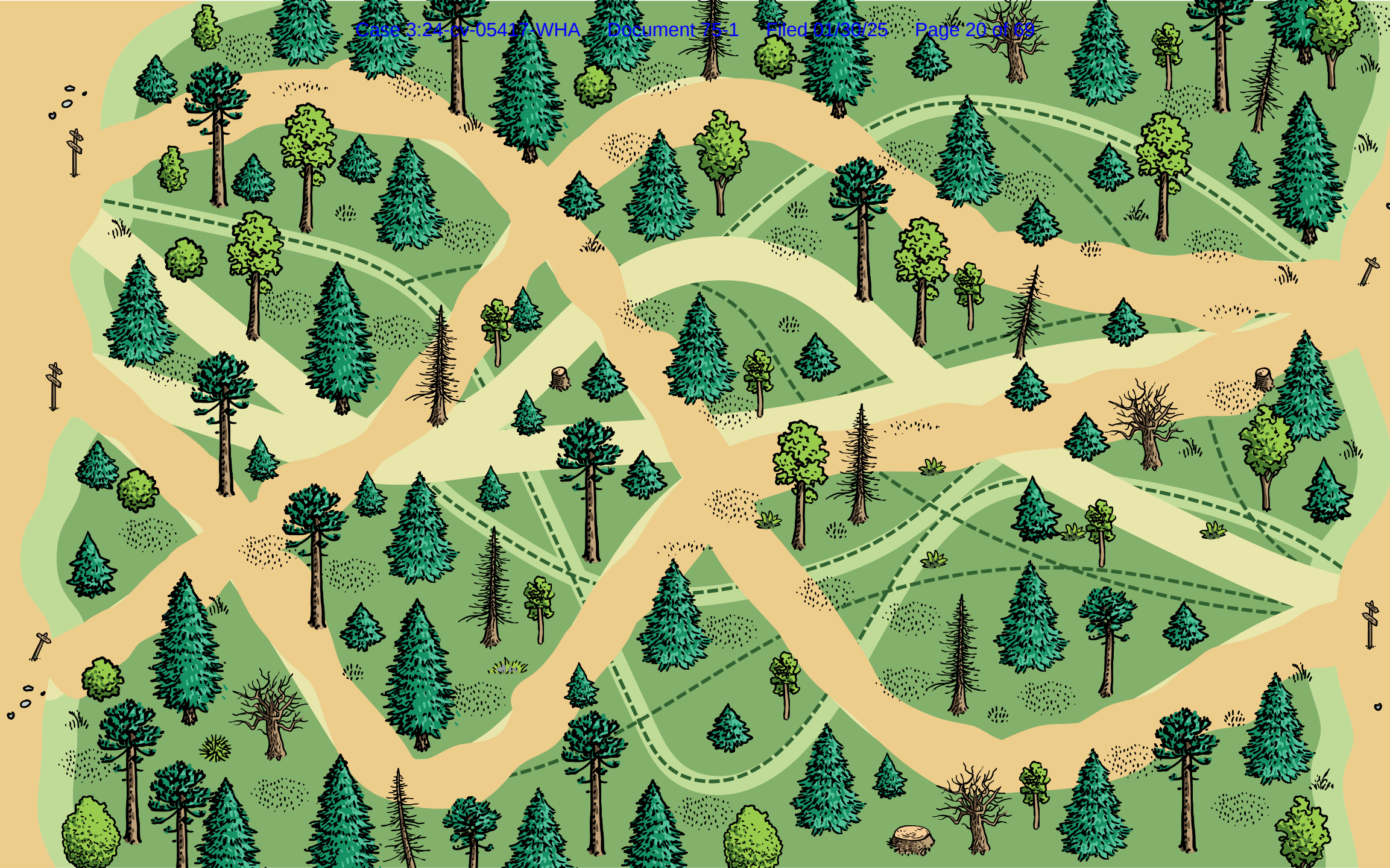
Pre-Training

- **Token**—words or fragments of words represented by a number
- **Vectors**—coordinates locating tokens relative to each other and encoding contextual information
- Predicting against training data in millions of **iterations**
- Ultimately creates a **next-token predictor**

3

Fine Tuning

- Model shown example conversations curated for certain characteristics (e.g. acting like a chatbot) or certain guardrails (e.g. copyright)



The Training Process

1

Acquiring the Data

- Scraped from across the internet, from public and/or pirating sites
- Books highly valued for diversity, long-context reasoning, high-quality writing, and more.

2

Pre-Training

- **Token**—words or fragments of words represented by a number
- **Vectors**—coordinates locating tokens relative to each other and encoding contextual information
- Predicting against training data in millions of **iterations**
- Ultimately creates a **next-token predictor**

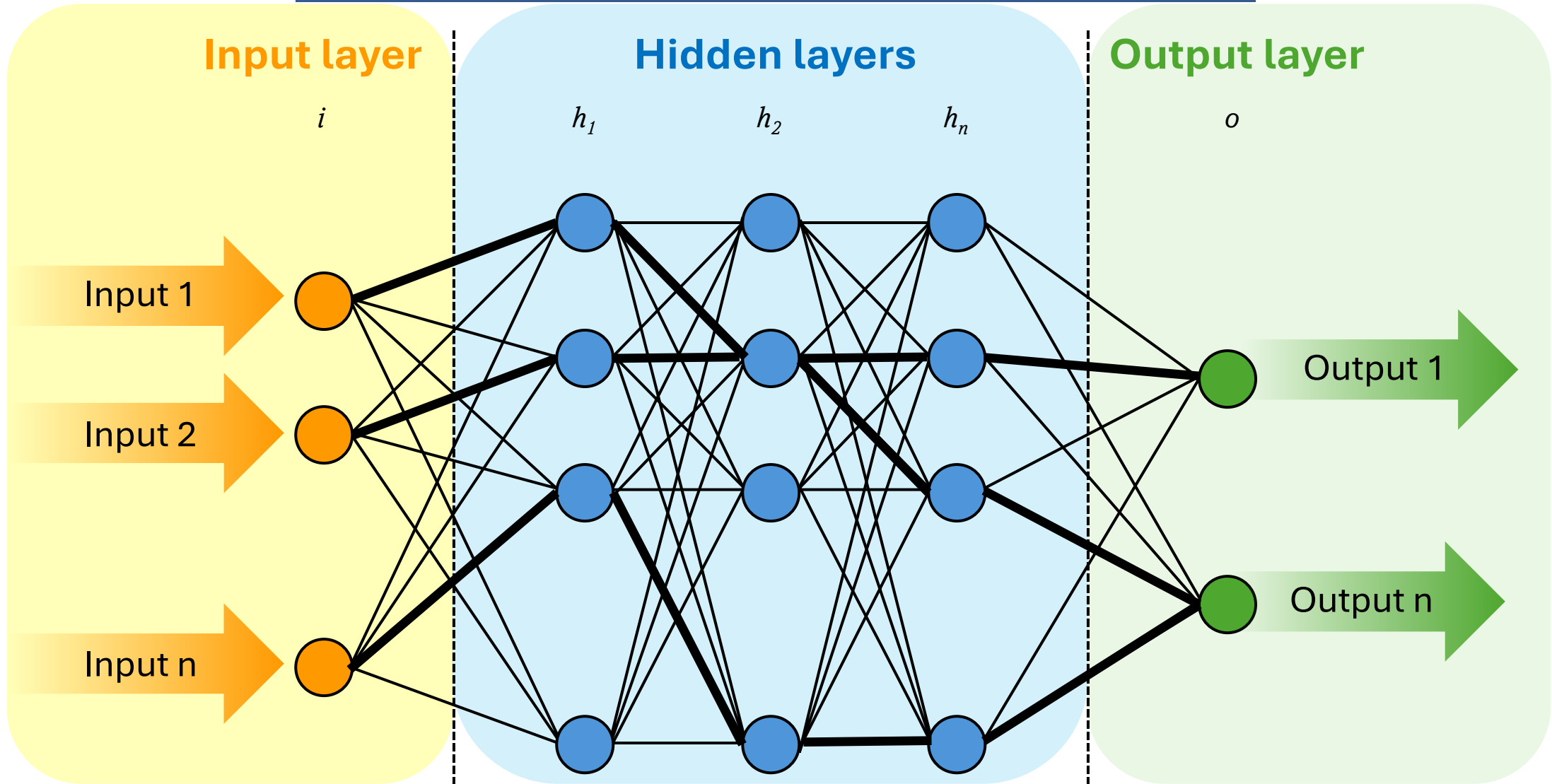
3

Fine Tuning

- Model shown example conversations curated for certain characteristics (e.g. acting like a chatbot) or certain guardrails (e.g. copyright)

LLMs Are a Type of Neural Network

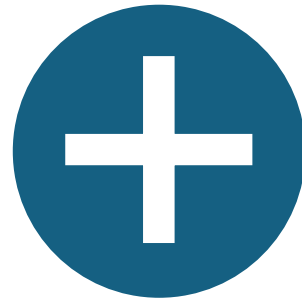
Multi-Layer Perceptron



“Transformer” Architecture Encodes Context to Generate Output

Encoder:

Processes input sequences in parallel to generate numerical representations of context

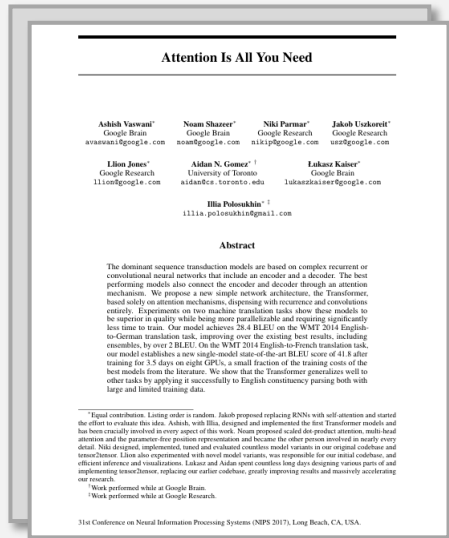


Decoder:

Processes these contextual embeddings in parallel to generate output predictions

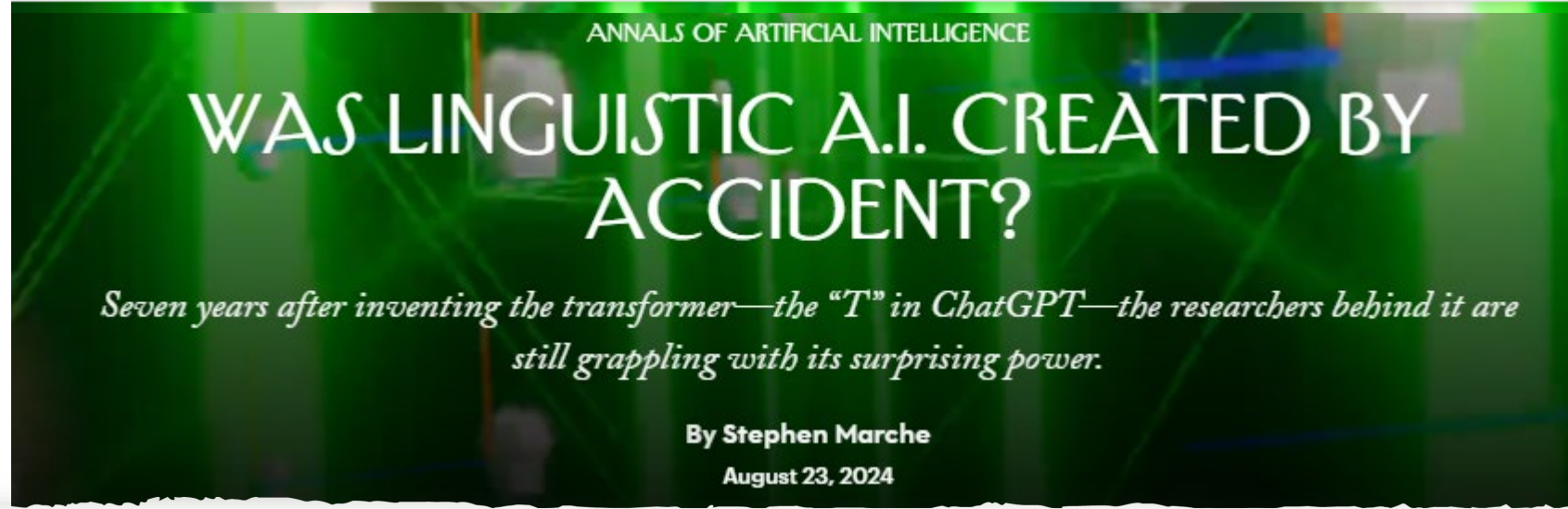
“Transformer”: A Neural Network Architecture

<https://www.newyorker.com/science/annals-of-artificial-intelligence/was-linguistic-ai-created-by-accident>



Attention Is All You Need

By Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, 2017



The element of accident in the transformer’s outsized success has evoked a notable humility in its inventors. When I asked Parmar how she would rate our current understanding of the models developed with the transformer, she said, “Very low.” I asked, How low? Ten per cent? One per cent? She shrugged: “How do we understand other humans? It will be the same with the models.” It’s fitting that the architecture outlined in “Attention Is All You Need” is called the transformer only because Uszkoreit liked the sound of that word. (“I never really understood the name,” Gomez told me. “It sounds cool, though.”)

Math of LLMs: Matrix Multiplication and the Chain Rule

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} j & k & l \\ m & n & o \\ p & q & r \end{bmatrix} = \begin{bmatrix} (aj + bm + cp) & (ak + bn + cq) & (al + bo + cr) \\ (dj + em + fp) & (dk + en + fq) & (dl + eo + fr) \\ (gj + hm + ip) & (gk + hn + iq) & (gl + ho + ir) \end{bmatrix}$$

Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y} \frac{\partial f}{\partial q}$$

Pre-Training Process



One small step for man, one giant leap for _____.

penguins  52%

mankind  46%

frogs  2%

Pre-Training Process



One small step for man, one giant leap for penguins.

penguins  52%

mankind  46%

frogs  2%

Pre-Training Process



One small step for man, one giant leap for mankind.

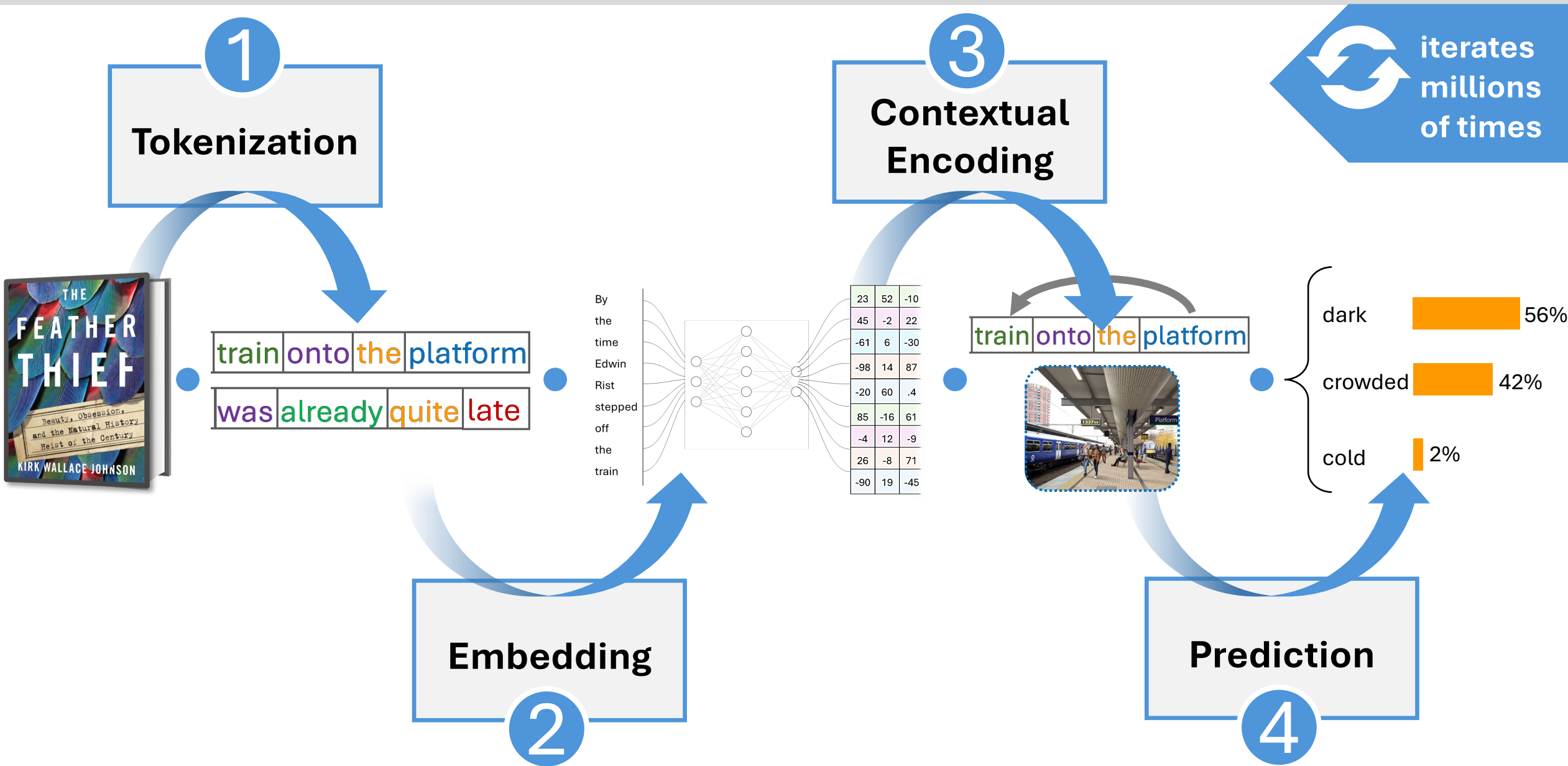
penguins  52%



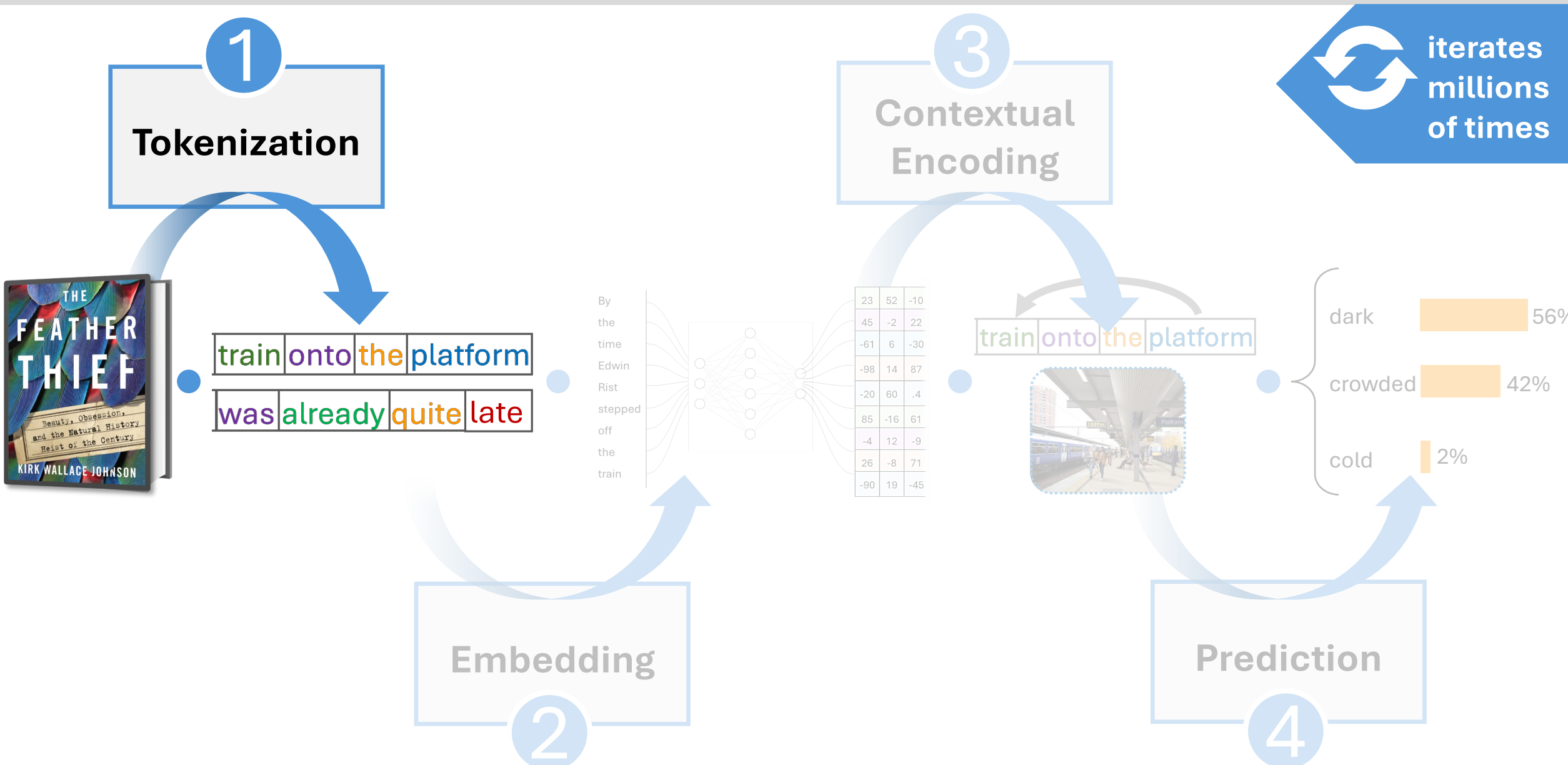
mankind  46%

frogs  2%

The LLM Pre-Training Process



The LLM Pre-Training Process



What is a Token?

2

Pre-Training

Token—words or fragments of words represented by a number

- **Vectors**—coordinates locating tokens relative to each other and encoding contextual information
- Predicting against training data in millions of **iterations**
- Ultimately creates a **next-token predictor**

What is a Token?

GPT-4o & GPT-4o mini

GPT-3.5 & GPT-4

GPT-3 (Legacy)

Many words map to one token, but some don't: indivisible.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🙌

Sequences of characters commonly found next to each other may be grouped together: 1234567890

Clear

Show example

Tokens

53

Characters

252

Many words map to one token, but some don't: indivisible.

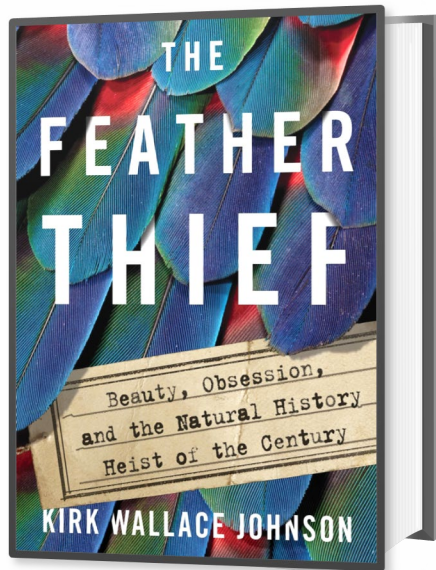
Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🙌

Sequences of characters commonly found next to each other may be grouped together: 1234567890

Text

Token IDs

What is a Token?



GPT-4o & GPT-4o mini

GPT-3.5 & GPT-4

GPT-3 (Legacy)

By the time Edwin Rist stepped off the train onto the platform at Tring, forty miles north of London, it was already quite late.

Clear

Show example

Tokens

29

Characters

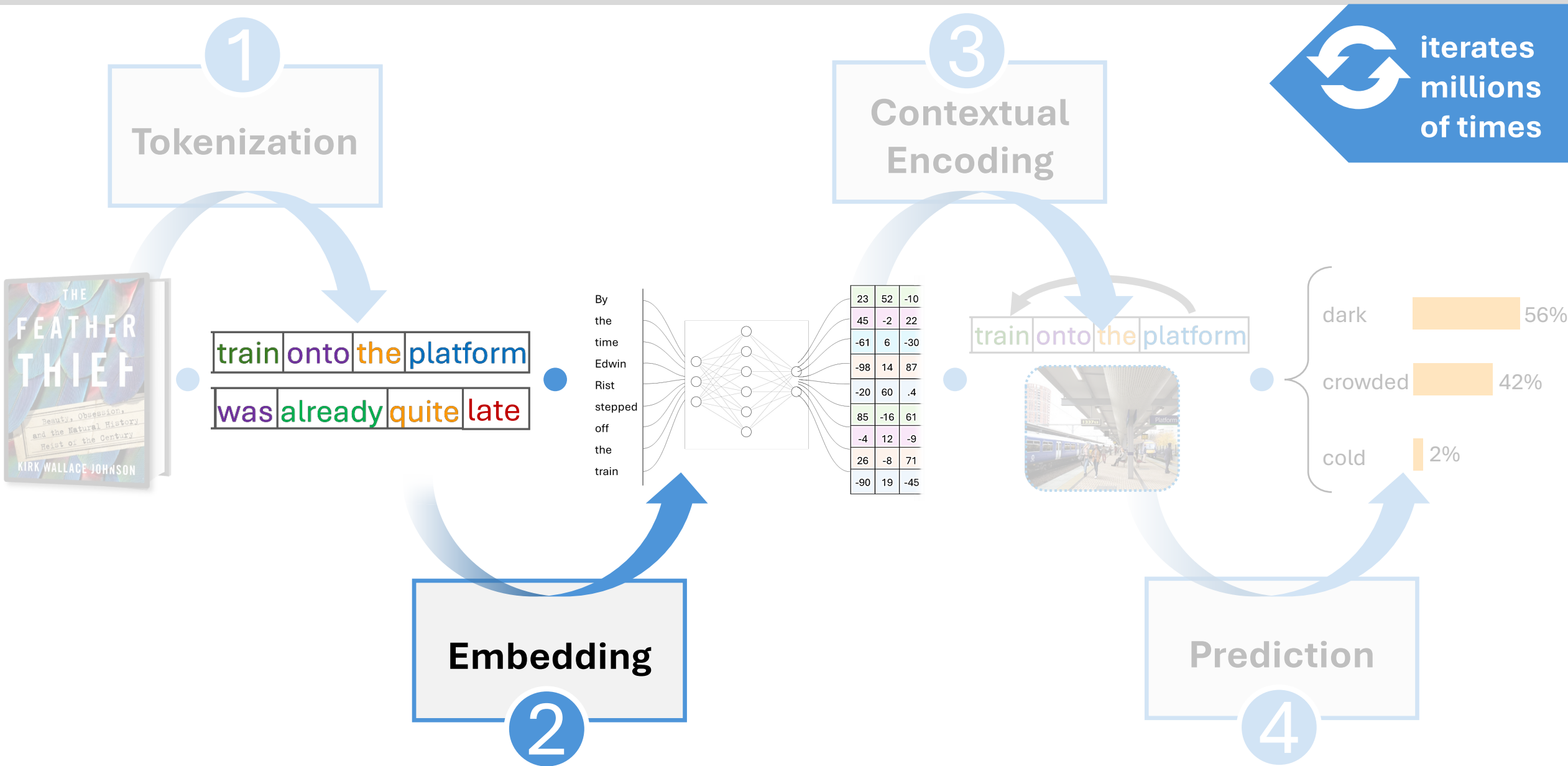
128

```
[1582, 290, 1058, 140958, 460, 421, 45744, 1277, 290, 8513, 13291, 290, 6361, 540, 1514, 289, 11, 56671, 10753, 16173, 328, 9741, 11, 480, 673, 4279, 6752, 7844, 13]
```

Text

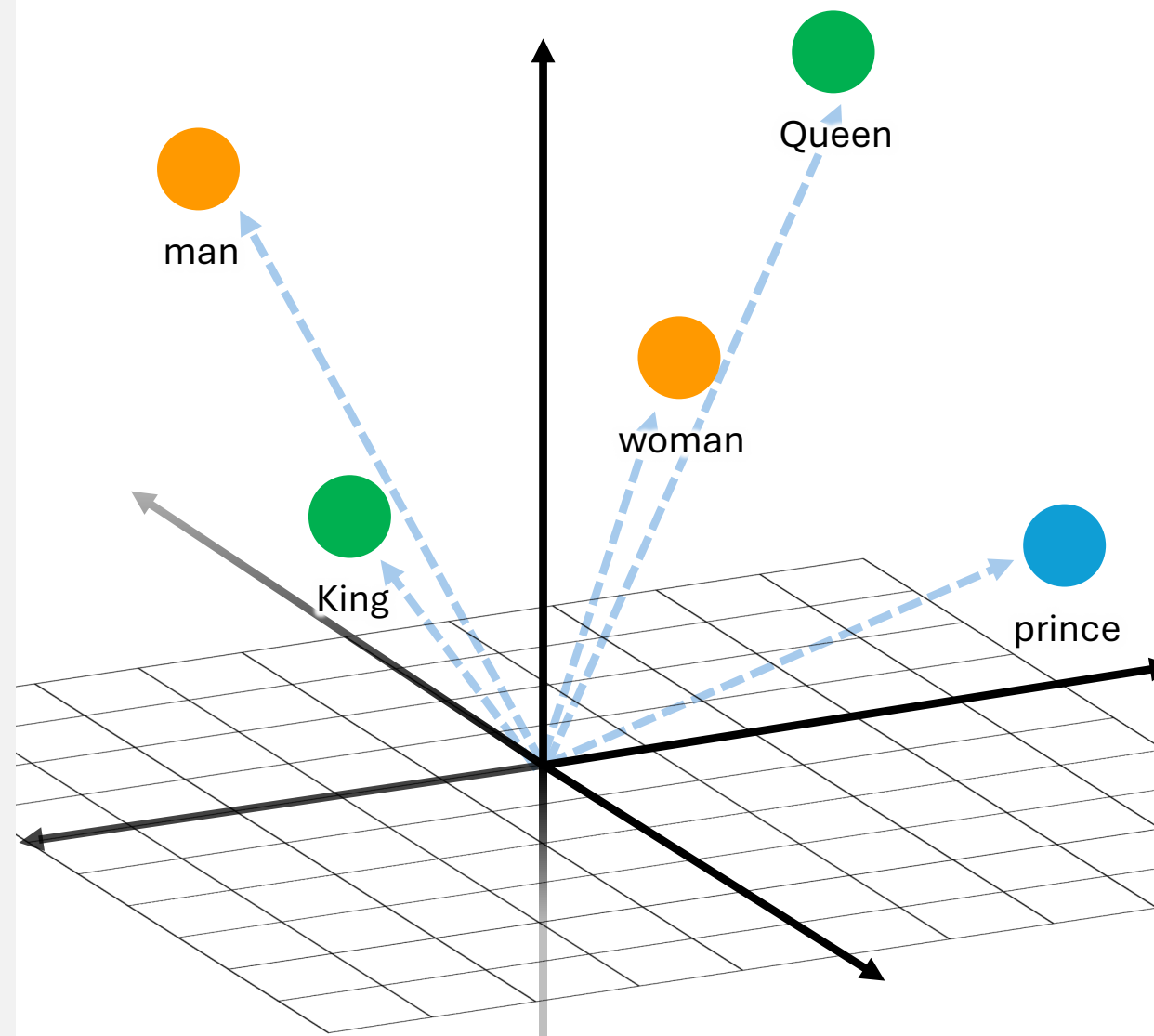
Token IDs

The LLM Pre-Training Process



Embedding Assigns 'Tokens' to 'Vectors'—Like Points on a Map

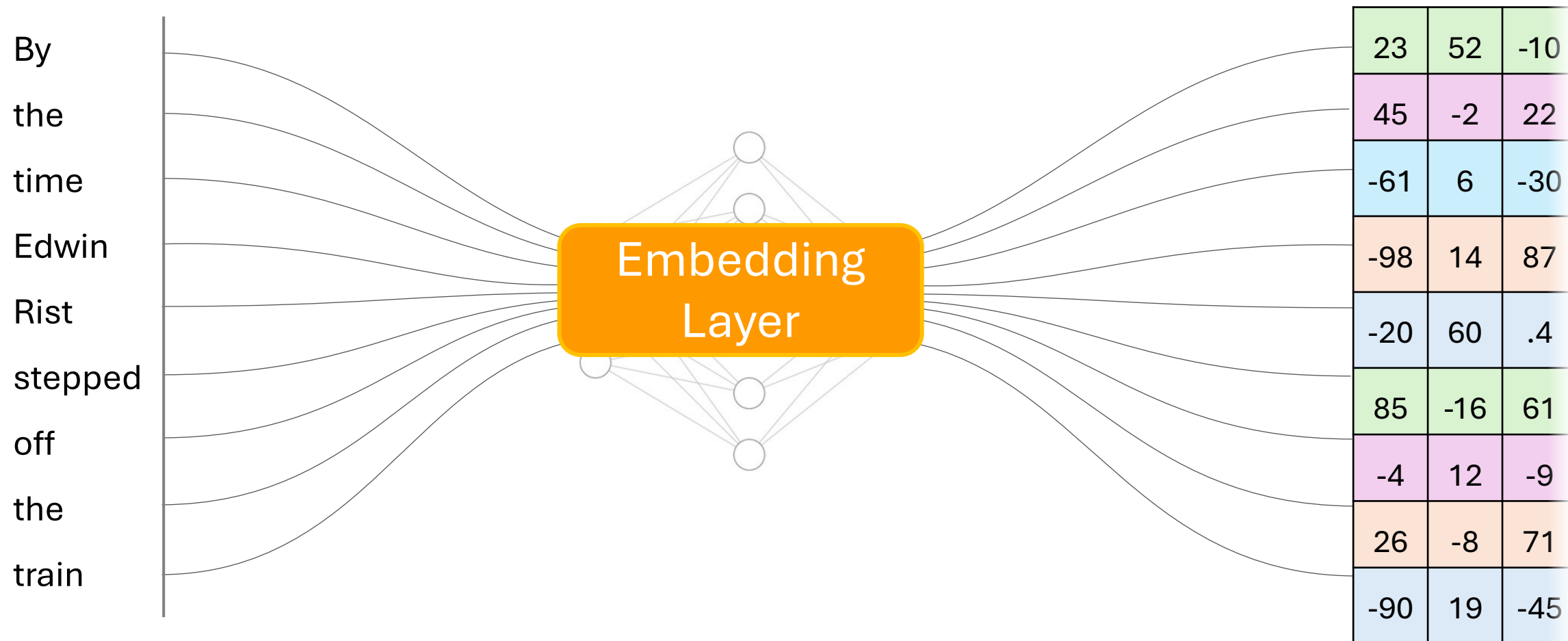
- Vectors are long lists of numbers that act like coordinates in very-high dimensional space.
- In the “Embedding” step, the tokens are assigned vectors.
- At first the vectors are essentially random, though they “improve” as the model trains.



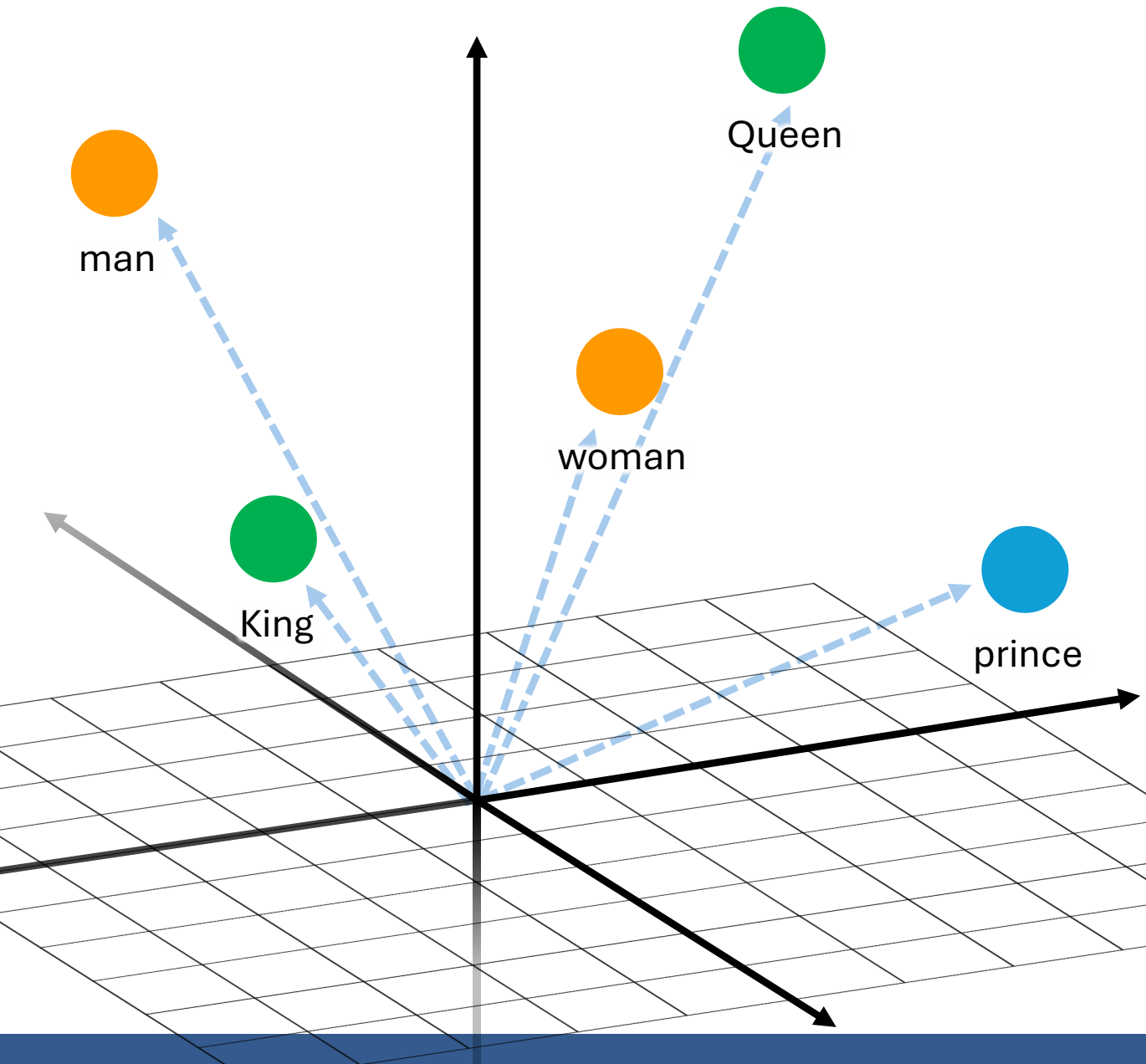
Phase 2: Embedding

Input

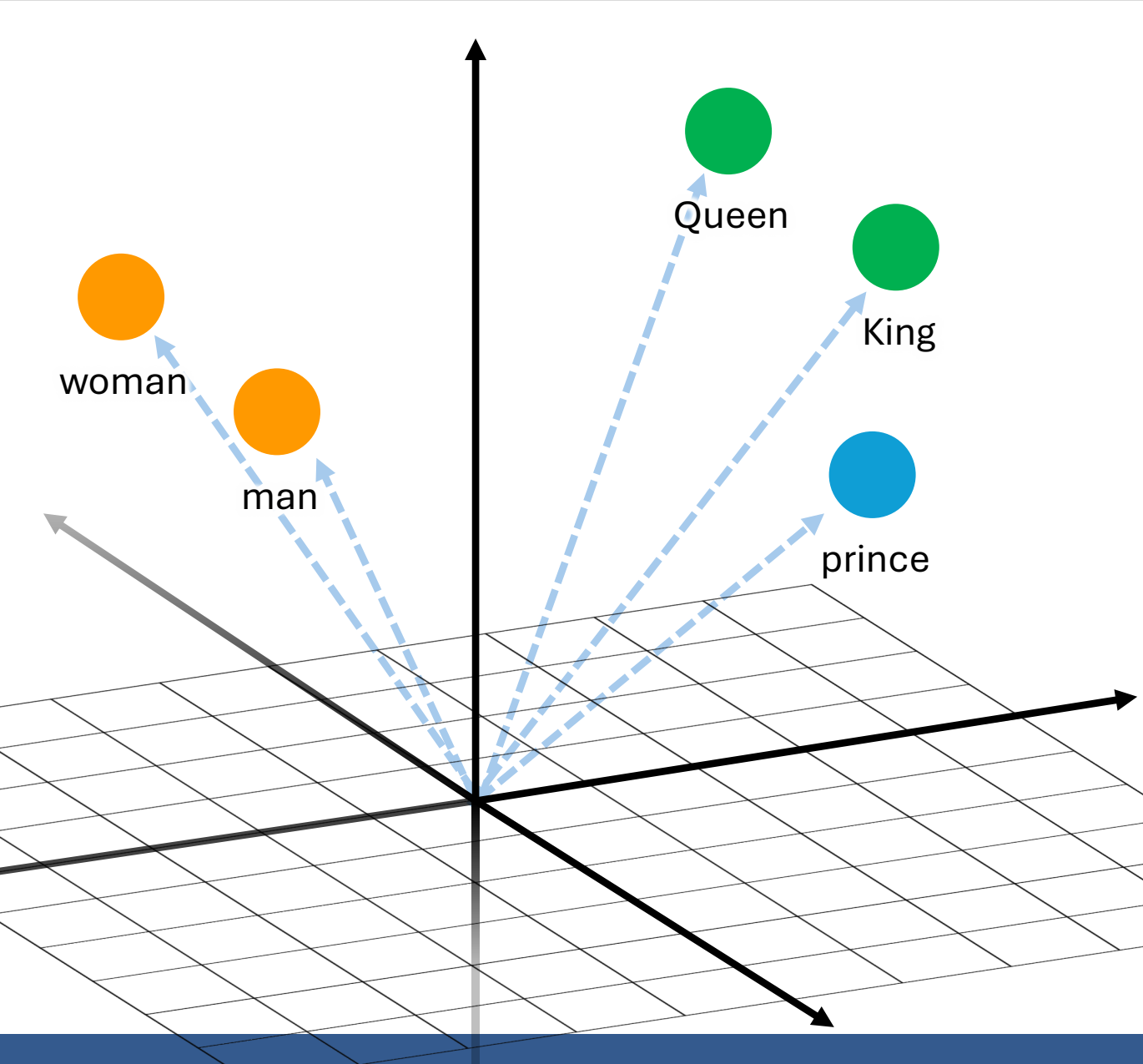
Vectors



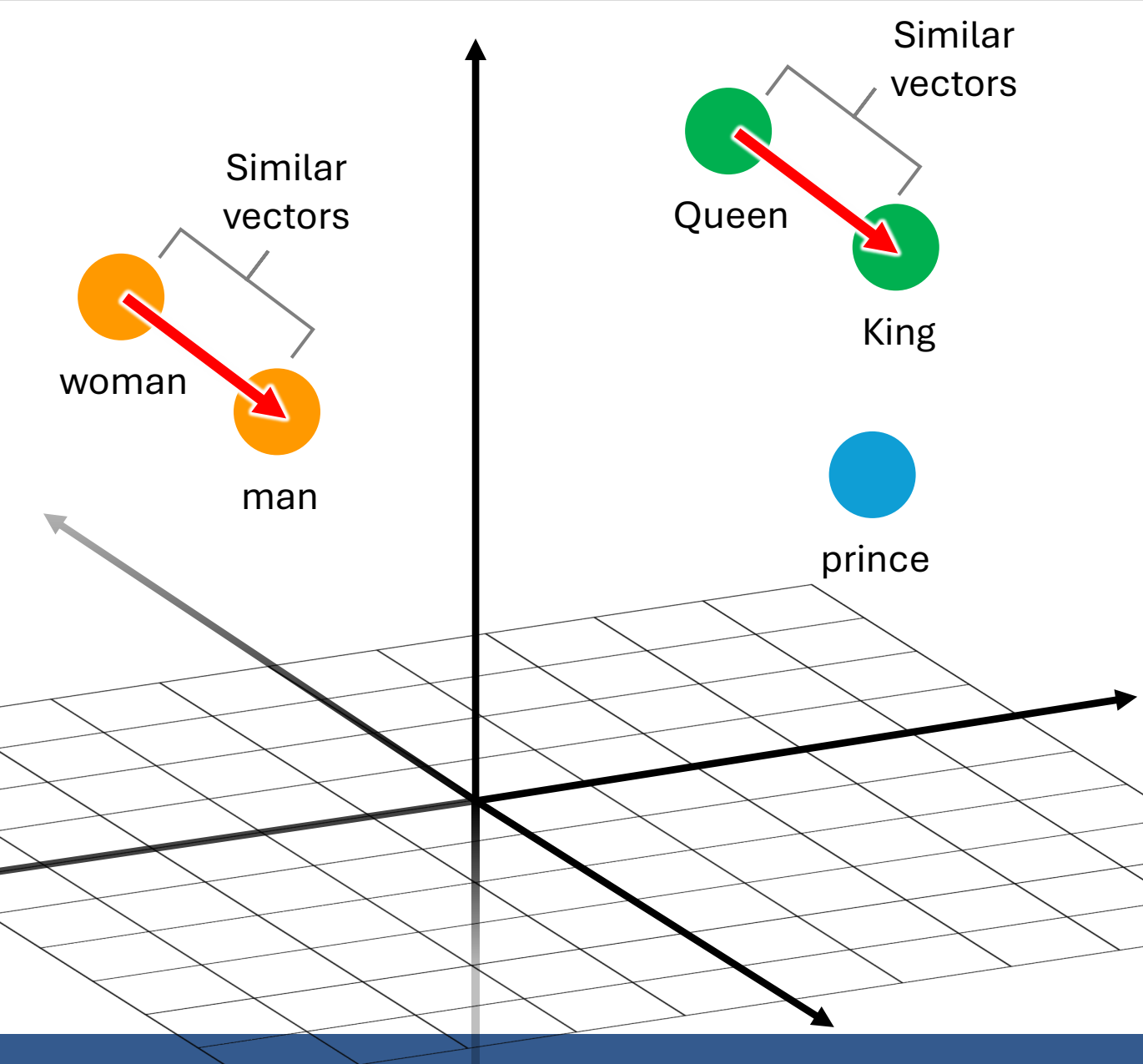
Phase 2: Embedding



Phase 2: Embedding



Phase 2: Embedding

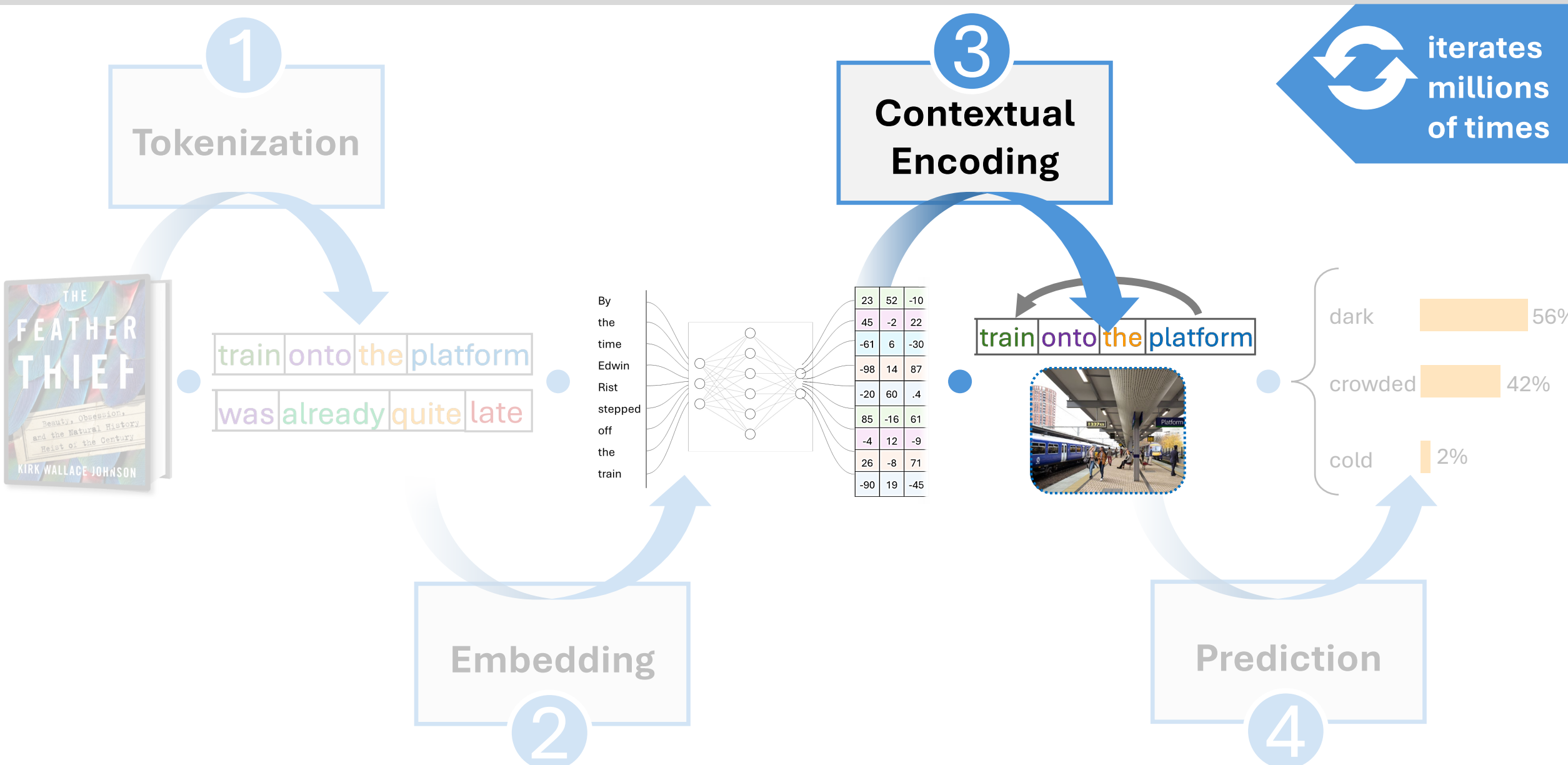


King – man \approx **royalty**

King – man + woman \approx **queen**

King – Queen \approx man - woman

The LLM Pre-Training Process

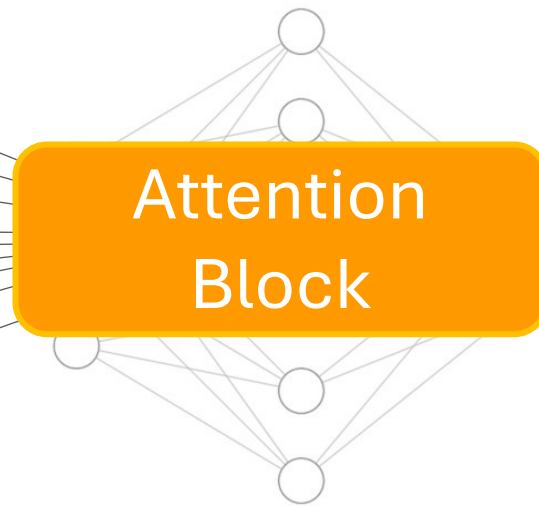


Input

23	52	-10
45	-2	22
-61	6	-30
-98	14	87
-20	60	.4
85	-16	61
-4	12	-9
26	-8	71
-90	19	-45

Vectors

94	25	-5
85	-88	10
92	-48	-2
-65	84	74
-14	58	15
42	-8	47
-97	-56	7
14	35	-78
52	66	-44



By the time Edwin Rist stepped off the train onto the platform
at Tring, forty miles north of London, it was already quite late

platform



Phase 3: Contextual Encoding

By the time Edwin Rist stepped off the train onto the platform
at Tring, forty miles north of London, it was already quite late



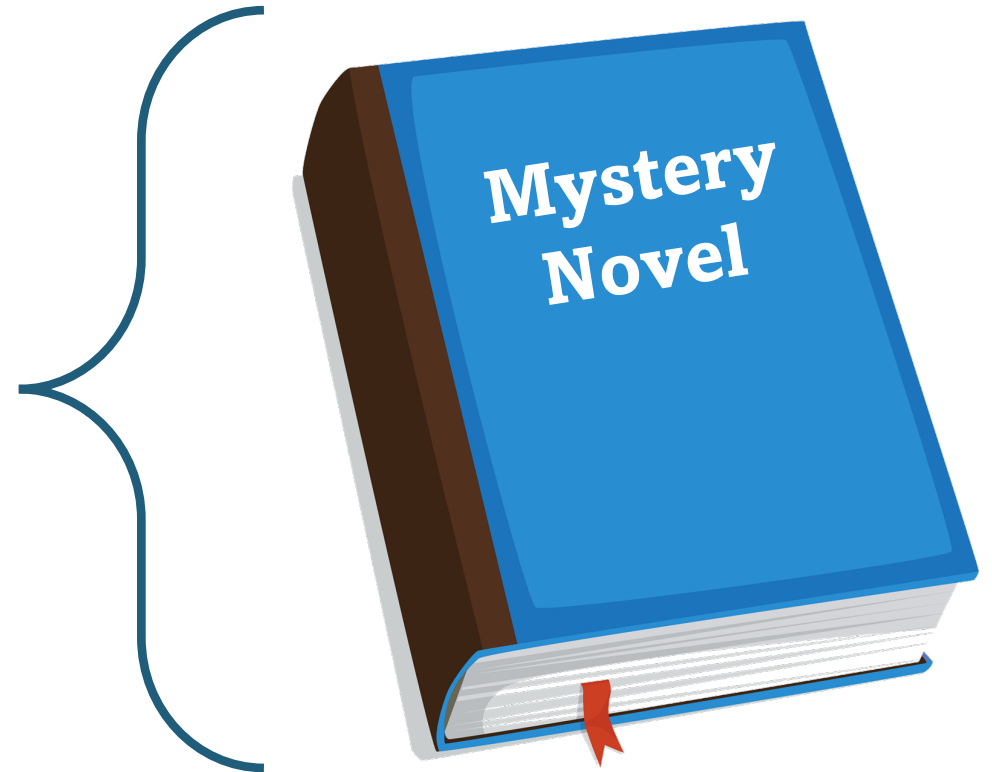
Tring



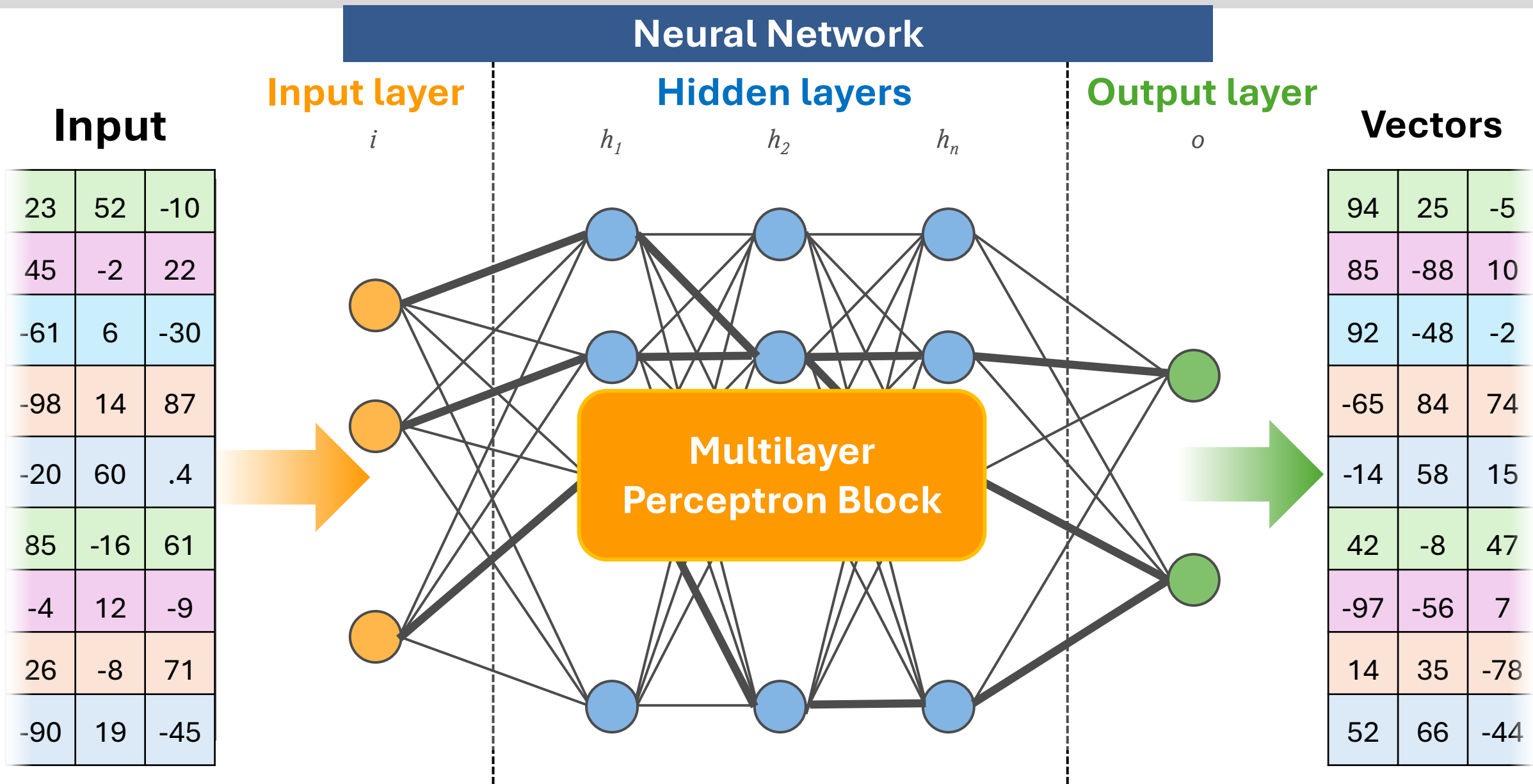
Phase 3: Contextual Encoding



“plant”



Phase 3: Contextual Encoding



Phase 3: Contextual Encoding

By the time Edwin Rist stepped off the train onto the platform
at Tring, forty miles north of London, it was already quite late

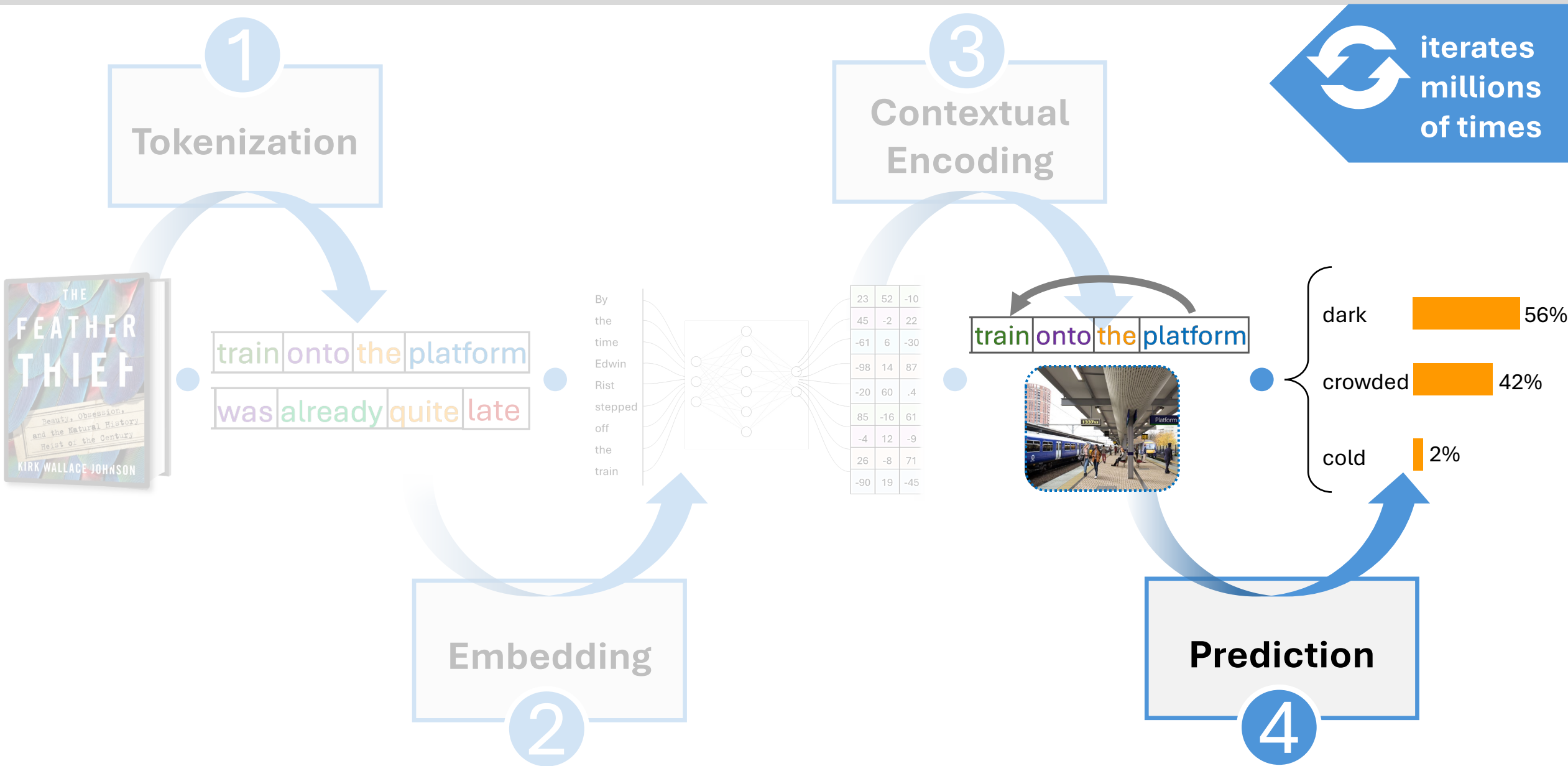


“time”
“already”
“quite”



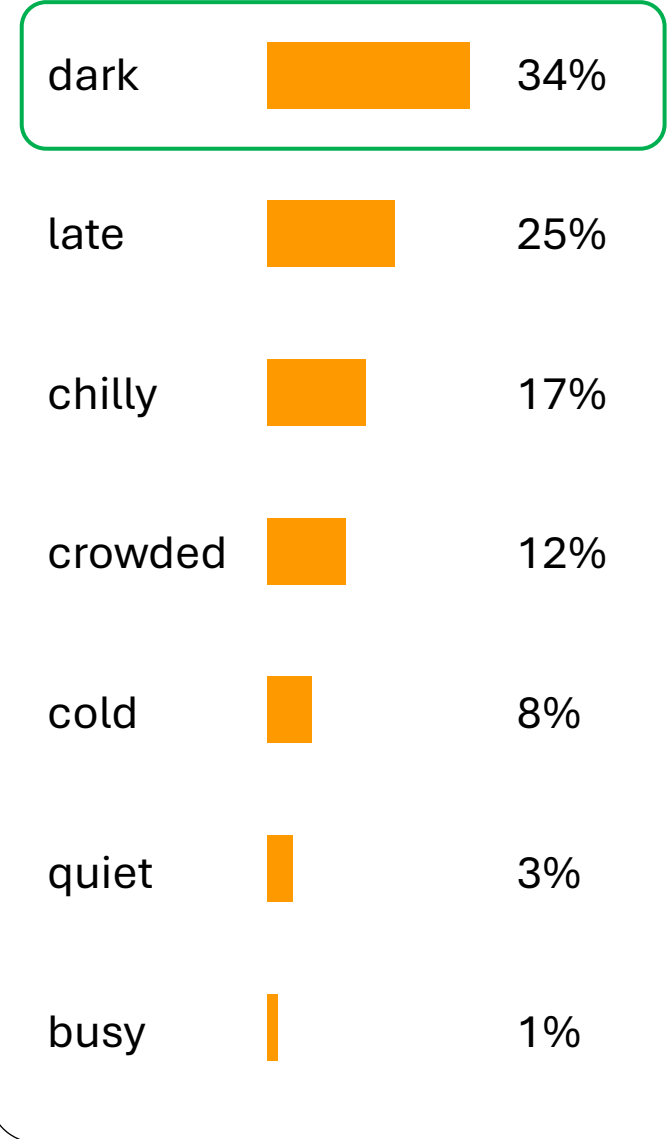
“negligence”
“breach”
“plaintiff”

The LLM Pre-Training Process



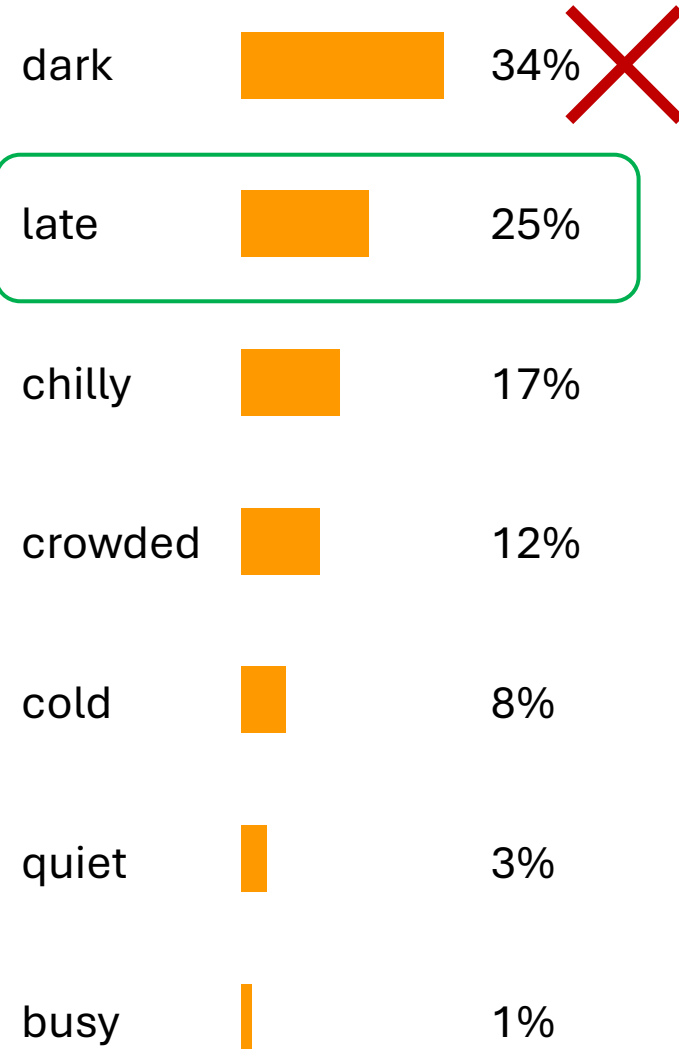
Phase 4: Prediction and Learning

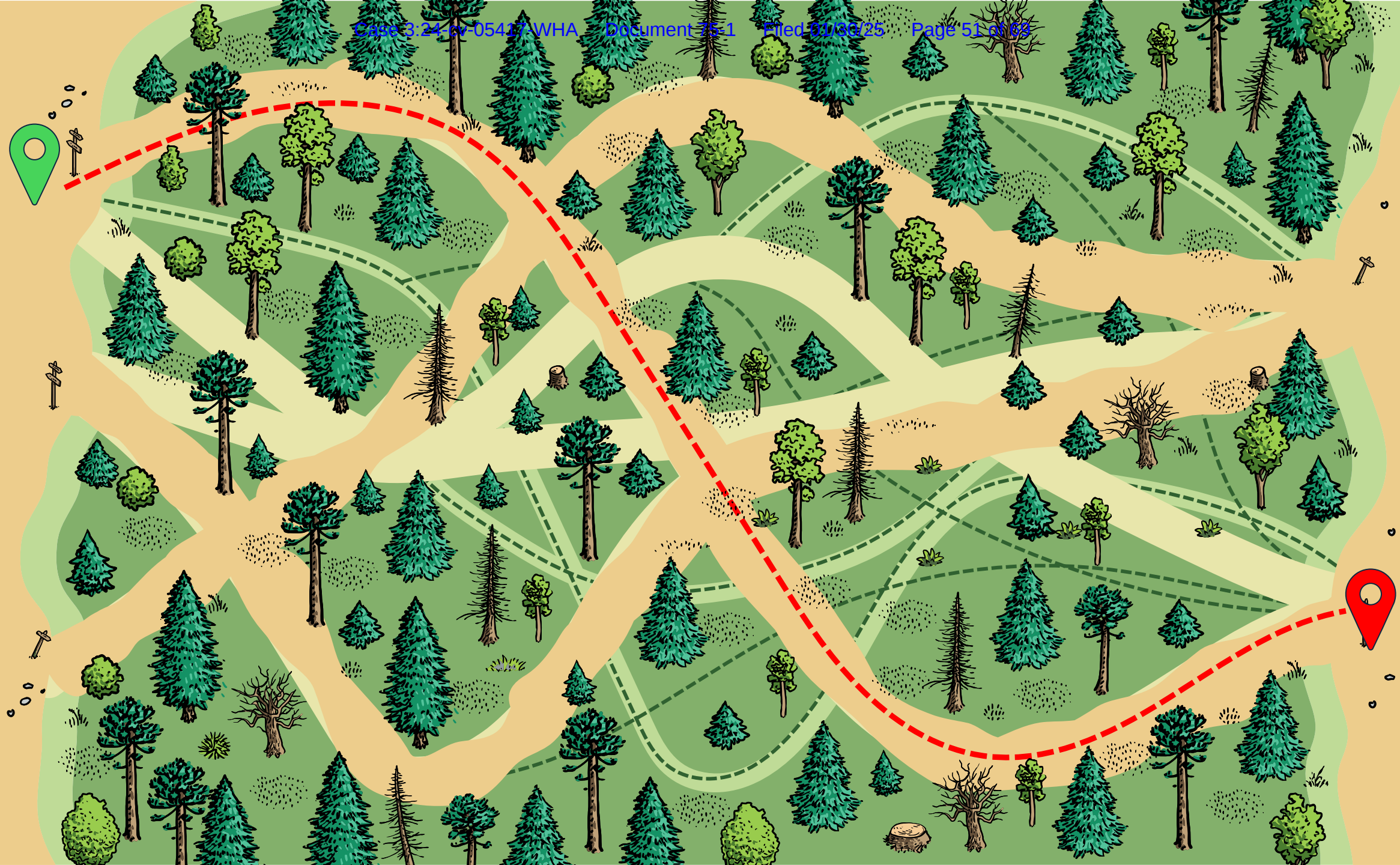
By the time Edwin Rist stepped off the train onto the platform
at Tring, forty miles north of London, it was already quite



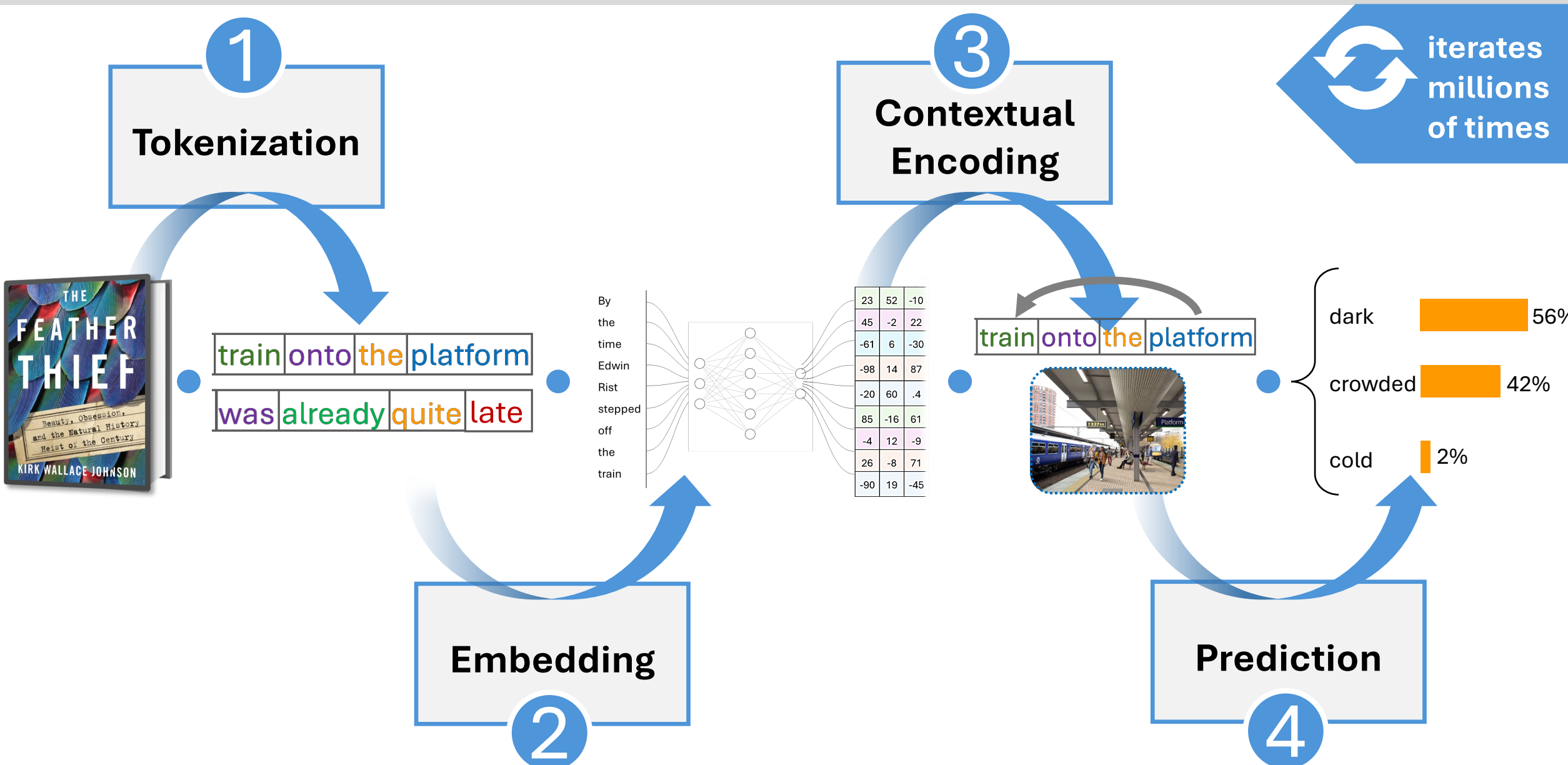
Phase 4: Prediction and Learning

By the time Edwin Rist stepped off the train onto the platform
at Tring, forty miles north of London, it was already quite late





The LLM Pre-Training Process



The Training Process

1

Acquiring the Data

- Scraped from across the internet, from public and/or pirating sites
- Books highly valued for diversity, long-context reasoning, high-quality writing, and more.

2

Pre-Training

- **Token**—words or fragments of words represented by a number
- **Vectors**—coordinates locating tokens relative to each other and encoding contextual information
- Predicting against training data in millions of **iterations**
- Ultimately creates a **next-token predictor**

3

Fine Tuning

- Model shown example conversations curated for certain characteristics (e.g. acting like a chatbot) or certain guardrails (e.g. copyright)

Different Types of Guardrails



1

User Side Blocking

looks at the *input* and sanitizes it

2

Alignment Training


changes the *model* itself based on human feedback

3

Post-Interference Filtering

looks at the model's *output* while it's being generated

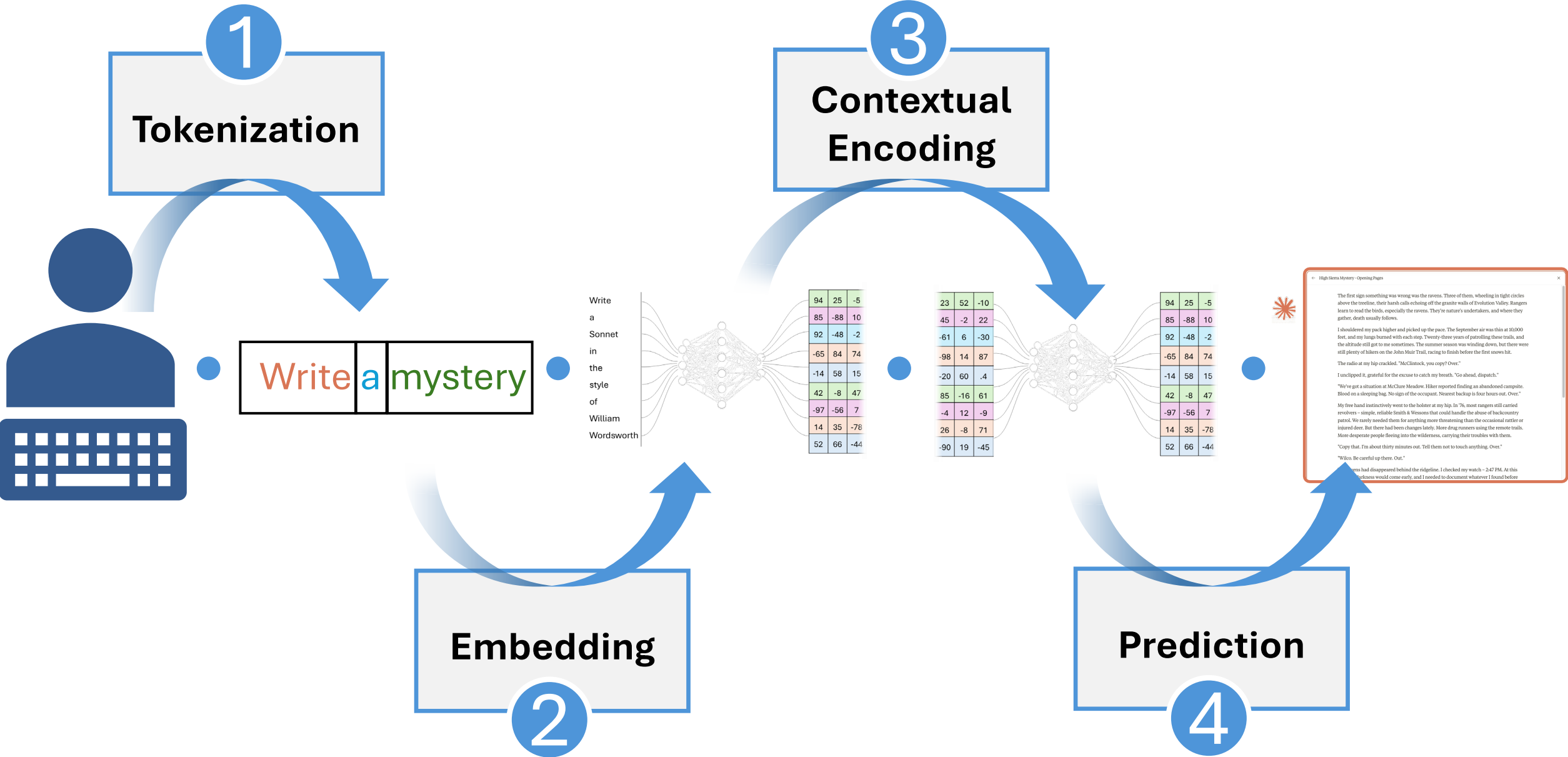
How does an LLM work? “The Inference Stage”



The Inference Stage

JC Write me the opening two pages of a mystery novel, set on the John Muir trail in the 1970s.

The LLM Inference Process



LLM Hallucinations

AI Generated Case Cites

Varghese v. China Southern Airlines Co Ltd,
925 F.3d 1339 (11th Cir. 2019).

Does not exist

Shaboon v. Egyptair, 2013 IL App (1st) 111279-
U (Ill. App. Ct. 2013).

Does not exist

Petersen v. Iran Air, 905 F. Supp. 2d 121
(D.D.C. 2012).

Does not exist

Martinez v. Delta Airlines, 2019 WL 4639462
(Tex. App. Sept. 25, 2019)

Does not exist

Estate of Durden v. KLM Royal Dutch Airlines,
2017 WL 2418825 (Ga. Ct. App. June 5, 2017).

Does not exist

Miller v. United Airlines, Inc., 174 F.3d 366 (2d.
Cir. 1999)

Does not exist



Presentation Roadmap

1 Background of LLMs

2 How an LLM is trained and how it works

3 Why books are so important to LLMs

What is Valuable Data?



“Models are truly approximating their datasets.”

The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

I've been at OpenAI for almost a year now. In that time, I've trained a lot of generative models. More than anyone really has any right to train. As I've spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It's becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don't matter, like what photos humans are likely to take or words humans commonly write down

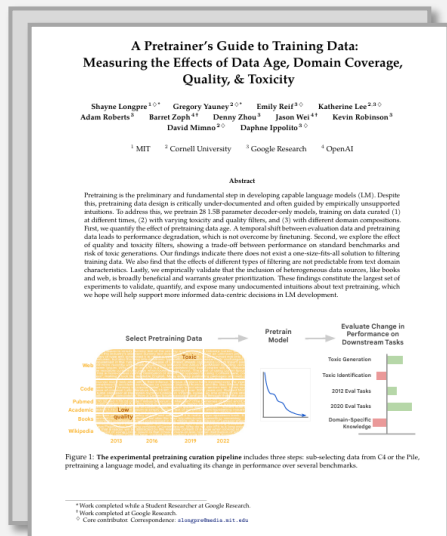
“[Model behavior is] determined by your dataset, nothing else. Everything else is a means to an end in efficiently deliver[ing] compute to approximating that dataset.”

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it's not the model weights that you are referring to. It's the dataset.



Understanding LLMs:
A Comprehensive
Overview from
Training to Inference
 By: Yiheng Liua, Hao Hea, Tianle Hana, Xu Zhanga, Mengyuan Liua, Jiaming Tiana, Yutong Zhangb, Jiaqi Wangc, Xiaohui Gaod, Tianyang Zhongd, Yi Pane, Shaochen Xue, Zihao Wue, Zhengliang Liue, Xin Zhangb, Shu Zhangc, Xintao Hud, Tuo Zhangd, Ning Qianga, Tianming Liue and Bao Ge,
 2024

“Books . . . contribute to the models’ training by exposing them to a diverse array of textual genres and subject matter, fostering a more comprehensive understanding of language across various domains.”

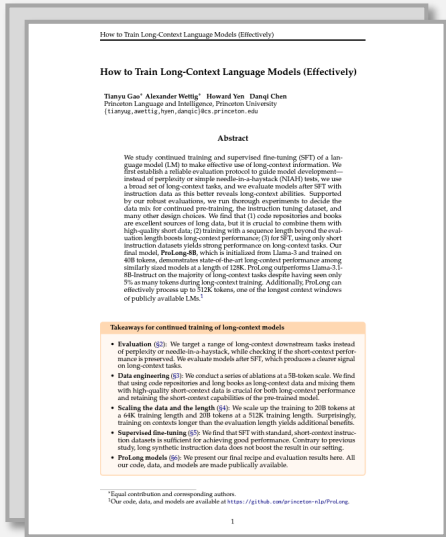


A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

By: Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, Daphne Ippolito, 2023

“The best performing domains comprise high-quality (Books) and heterogeneous (Web) data”

Long-form Coherence



Princeton AI researchers concluded that using “long books as long-context data” was “crucial” for “long-context performance.”

How to Train Long-Context Language Models (Effectively)

By: Tianyu Gao, Alexander Wettig, Howard Yen, Danqi Chen, 2024



The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao Stella Biderman Sid Black Laurence Golding
 Travis Hoppe Charles Foster Jason Phang Horace He
 Anish Thite Noa Nabeshima Shawn Presser Connor Leahy

EleutherAI
 contact@eleuther.ai

Abstract

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present the Pile, an 825 GiB English text corpus we gathered by training large-scale language models. The Pile is constructed from 22 diverse, high-quality subsets—both existing and newly constructed—many of which derive from academic or professional sources. Our evaluation of the in-sample performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.

venity leads to better downstream generalization capability (Boscuti, 2019). Additionally, large-scale language models have been shown to effectively acquire knowledge in a novel domain with only relatively small amounts of training data from that domain (Boscuti, 2019; Brown et al., 2020; Carlini et al., 2020). These results suggest that by mixing together a large number of smaller, high-quality, diverse datasets, we can improve the general cross-domain knowledge and downstream generalization capabilities of the model compared to models trained on only a handful of data sources.

To address this need, we introduce the Pile, a 825.18 GiB English text dataset designed for training large scale language models. The Pile is composed of 22 diverse and high-quality datasets, including both established natural language processing datasets and several newly introduced ones. In addition to its utility in training large language models, the Pile can also serve as a broad coverage benchmark for cross-domain knowledge and generalization ability of language models.

1 Introduction

Recent breakthroughs in general-purpose language modeling have demonstrated the effectiveness of training massive models on large text corpora for downstream applications (Radford et al., 2019; Shwartz et al., 2019; Raffel et al., 2019; Brown et al., 2020; Brown et al., 2020; Lepikhin et al., 2020). As the field continues to scale up language model training, the demand for high-quality massive text data will continue to grow (Kaplan et al., 2020).

We introduce new datasets derived from the following sources: PubMed Central, ArXiv, GitHub, the FreeLaw Project, Stack Exchange, the US Patent and Trademark Office, PubMed, Usenet, BBC, HackerNews, SciDirect, PhilPapers, and NIH ExPorter. We also introduce OpenWebText2 and BookCorpus2, which are extensions of the original OpenWebText (Gokulan and Cohen, 2019) and BookCorpus (Zhu et al., 2015; Kobayashi, 2018) datasets, respectively.

The growing need for data in language modeling has caused most existing large-scale language models to rely on the Common Crawl for most or all of their data (Brown et al., 2020; Raffel et al., 2019). While training on the Common Crawl has been effective, recent work has shown that dataset di-

In addition, we incorporate several existing high-quality datasets: Books3 (Presser, 2020), Project Gutenberg (PG19) (Ike et al., 2019), OpenSubtitles (Fleermann, 2016), English Wikipedia, DM Mathematics (Sutton et al., 2019), EuroParl (Beaulieu, 2005), and the Enron Emails corpus (Klein and Yang, 2004). To supplement these, we also in-

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3[†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB

“We included Bibliotik because **books are invaluable** for long-range context modeling research and coherent storytelling.”

Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails [†]	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao, Travis Hoppe, Anish Thite, Stella Biderman, Charles Foster, Noa Nabeshima, Sid Black, Jason Phang, Shawn Presser, Laurence Golding, Horace He, Connor Leahy, 2020



EleutherAI

Long-form Coherence

Input

Machine Learning

Vectors

By

th



cr in

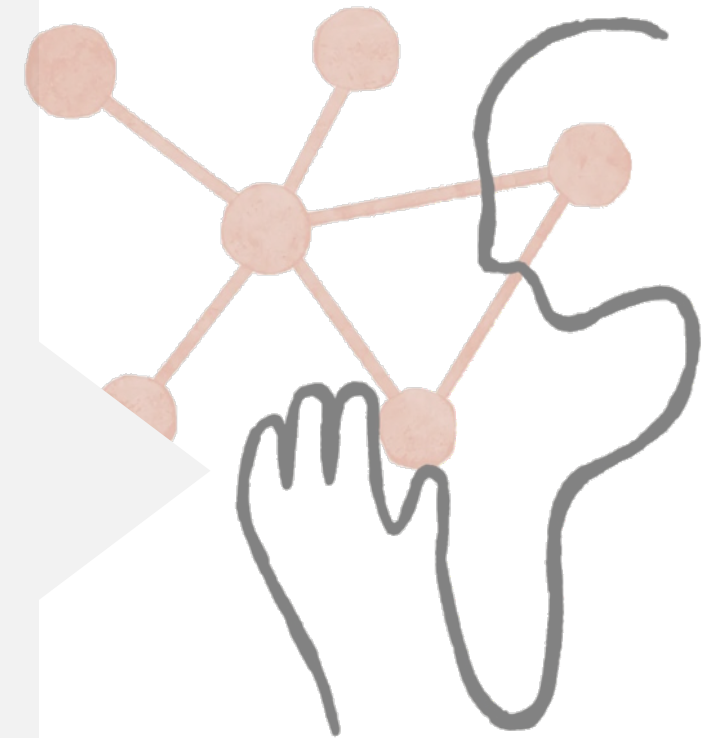
Attention Block

23	52	-10
45	-2	22
-61	6	-30
-98	14	87
-20	60	.4
85	-16	61
-4	12	-9
26	-8	71
-90	19	-45

Books are especially valuable training data

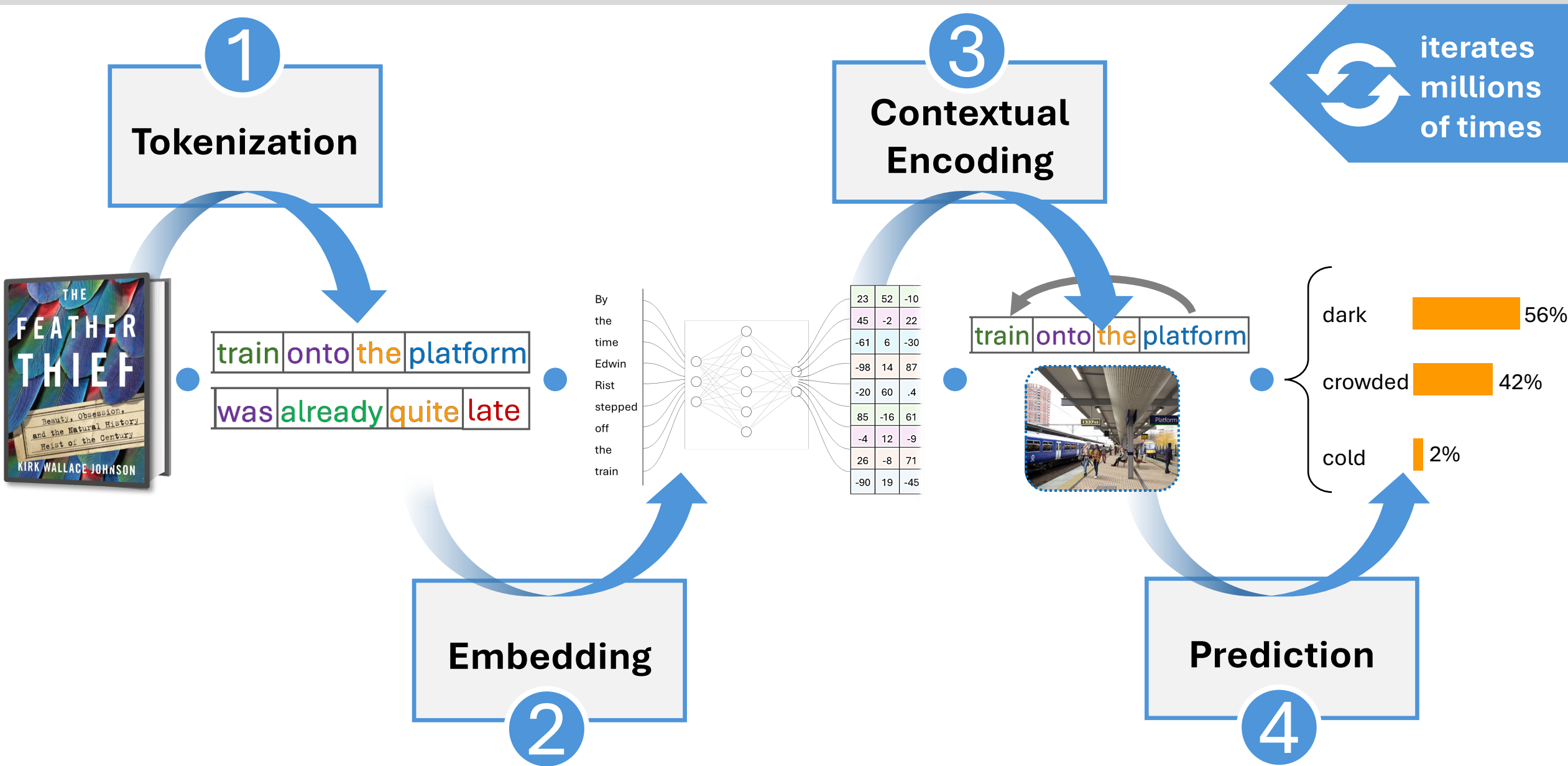
Books are indeed a particularly valuable source of training data for several key reasons:

- **High signal-to-noise ratio:** Books generally represent carefully crafted, edited, and curated content. Unlike social media posts or informal web content, books typically go through extensive review and editing processes, resulting in higher quality information and expression.
- **Complex reasoning and extended arguments:** Books allow authors to develop sophisticated arguments and ideas over hundreds of pages. This extended format enables deep exploration of topics and complex chains of reasoning that shorter formats can't support.
- **Rich contextual relationships:** Books often contain complex networks of references between concepts, characters, and ideas. This interconnected nature provides rich semantic relationships for learning.



ANTHROPIC
* Claude

The LLM Pre-Training Process



LLMs Would Not Work Without High-Quality Data

“Infinite Monkey Theorem” vs. “H.I. — Human Imitation”



A numerical evaluation of the Finite Monkeys Theorem

Stephen Woodcock,
Jay Fallett, 2024

