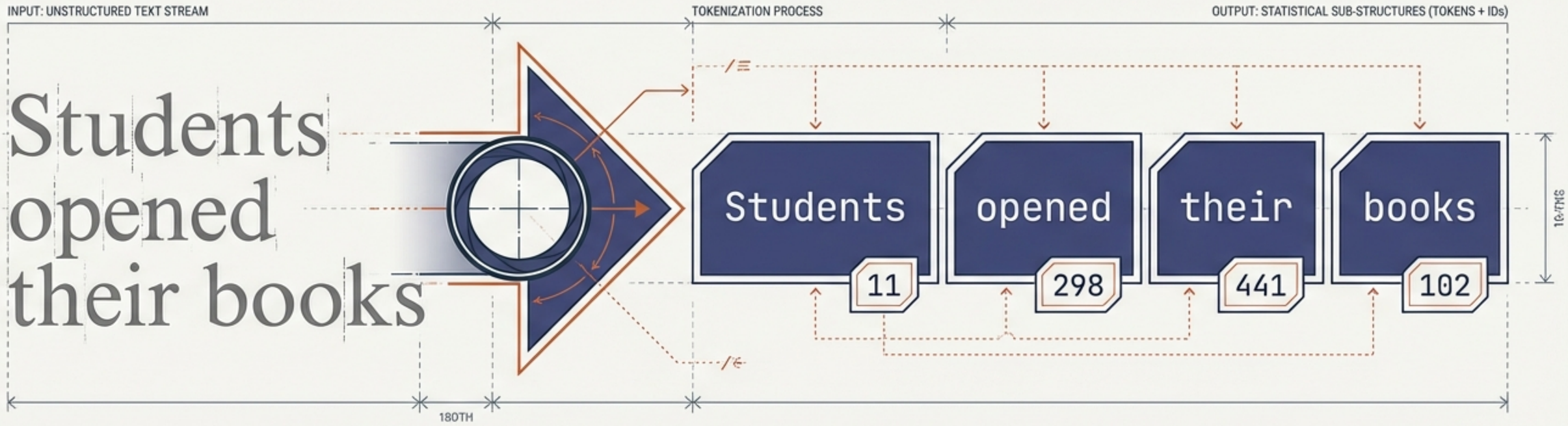
Decoding the Token

The Atomic Units of Modern AI



From raw character streams to the statistical sub-structures that define LLM performance.

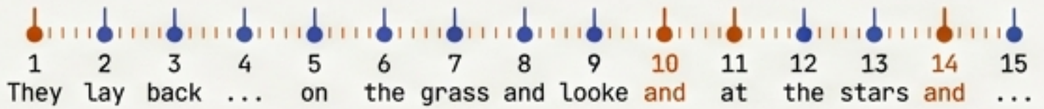
Measuring the Corpus: Tokens vs. Types

They lay back on the grass **and** looked at the stars **and**...

TOKENS

15

Total instances in running text.



TYPES

13

Unique elements in vocabulary.

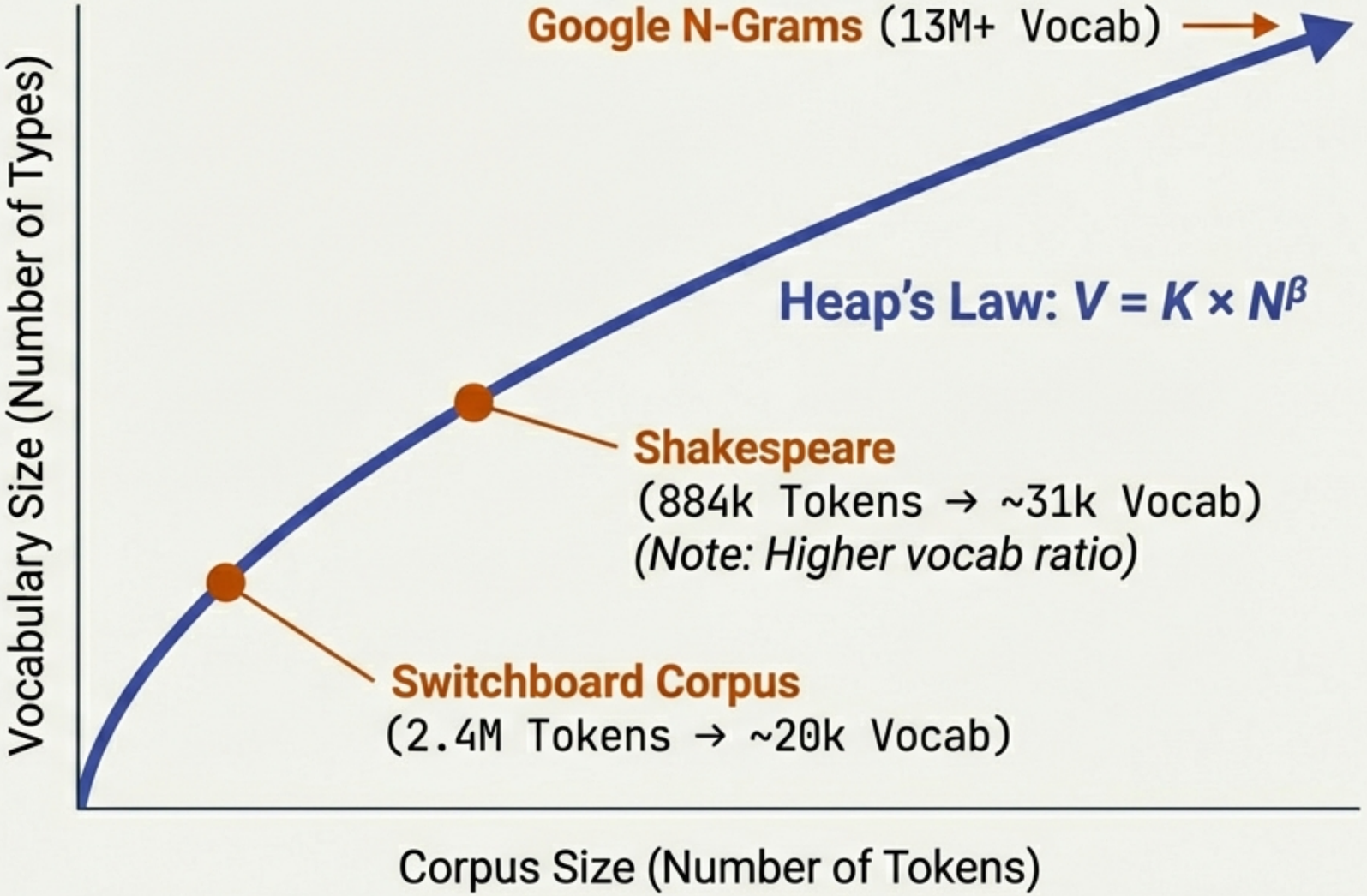
15 (Total) - 2 (Repetitions) = 13 Unique Types

In massive datasets, the gap between Tokens (volume) and Types (vocabulary) defines the computational challenge.

The Limitation of Words: The Vocabulary Explosion

We cannot map every word.

To cover the web—including URLs, dates, and numbers—a word-level vocabulary would require millions of entries. This is computationally impossible for the output layer of a neural network.



The Complexity of Context & Morphology

Mixed Script/Language

He was a friend, **don’t worry.**

Is ‘**or**’ English or Hindi? Ambiguity requires context.

Abbreviations & SMS

I went **2** the store.

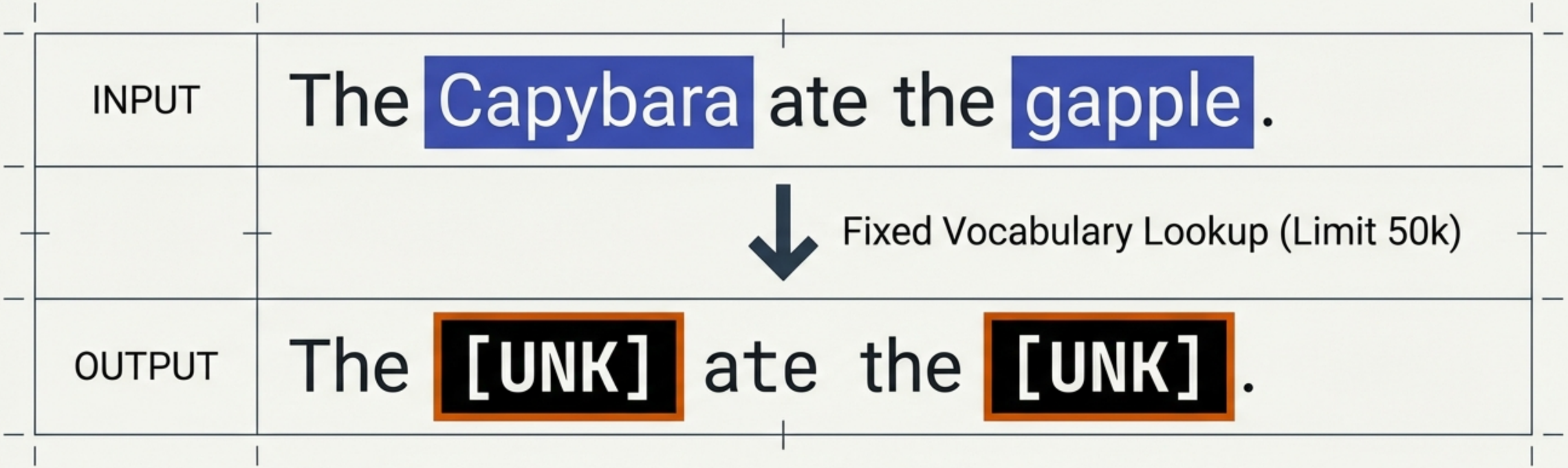
Token ‘**2**’ maps to number, but semantically means preposition ‘**to**’.

Compound Fusion (Sanskrit)

Shveto + Dhavati → → Shvetodhavati

Word boundaries disappear. Simple whitespace tokenization fails.

The Destroyer of Information: The [UNK] Token



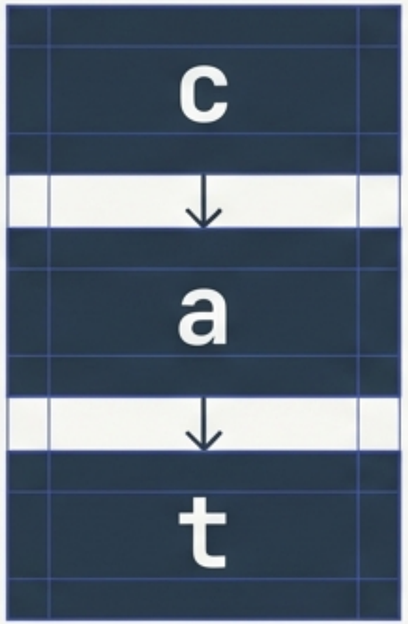
Total Information Loss.

Rare words are erased. The model treats a rodent and a fruit exactly the same. We cannot generate what we cannot see.

The Limitation of Characters

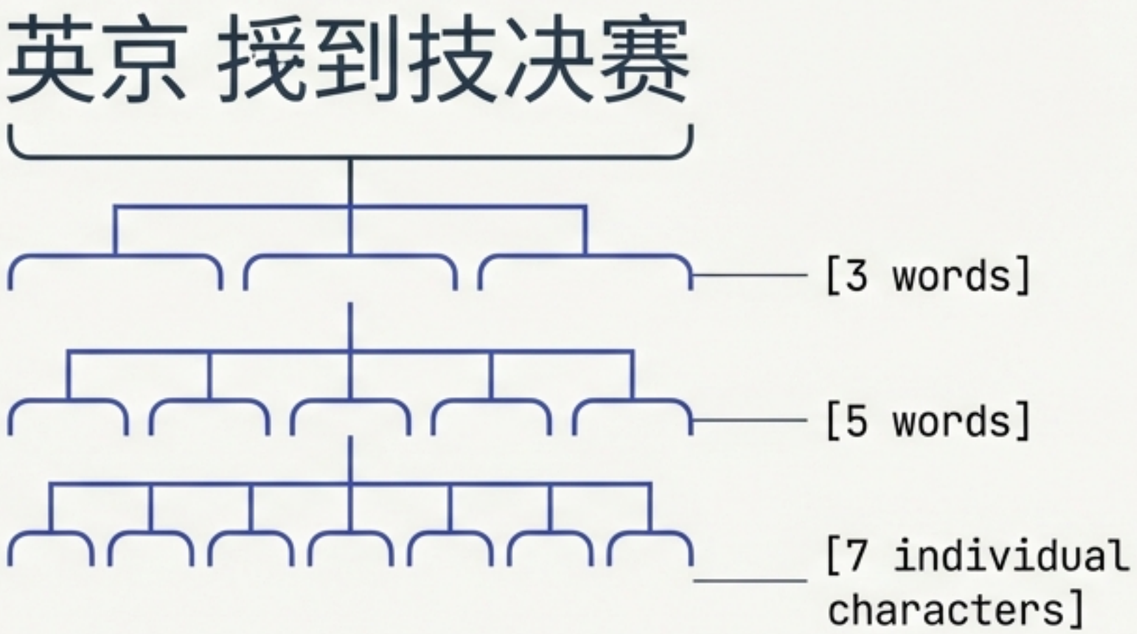
Why not just tokenize every letter?

English



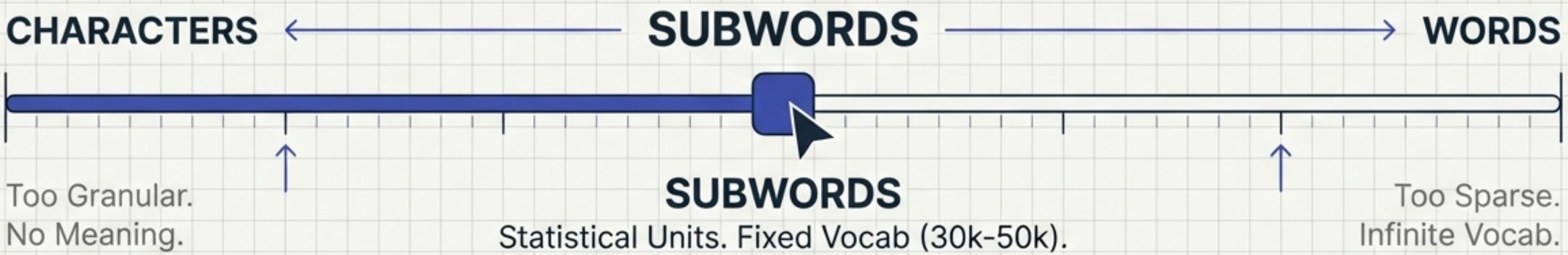
No [UNK] tokens, but the model must re-learn that “c-a-t” means a furry animal.
Loss of semantic shortcut.

Chinese



Character-level tokenization solves the vocabulary size problem but creates extremely long sequences, increasing computational cost and diluting context.

The Goldilocks Solution: Subwords



Unlikeliest

[un] + [likely] + [est]

Paradigm Shift: From Linguistic Units (defined by dictionaries) to Statistical Units (defined by data frequency).

Byte Pair Encoding (BPE): The Logic

Core Rule Text: Iteratively merge the most frequent pair of adjacent characters.
Don't use linguistic insight; use data frequency.

Phase 1: Training

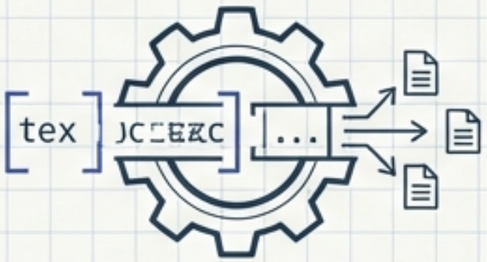
The Token Learner

Scans corpus to build merge rules.

Generates Vocabulary

Phase 2: Inference

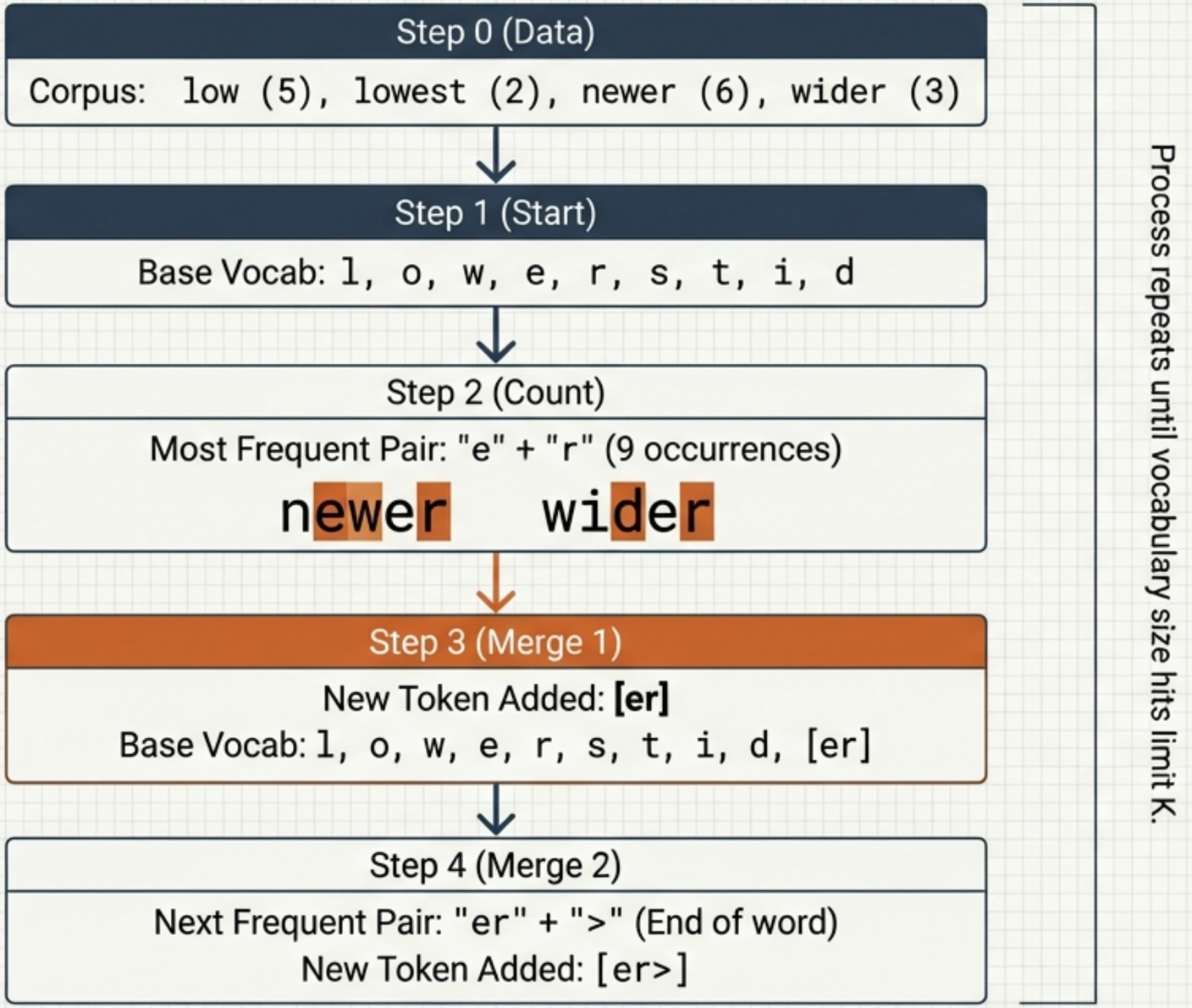
The Token Segmenter



Applies rules to slice new text.

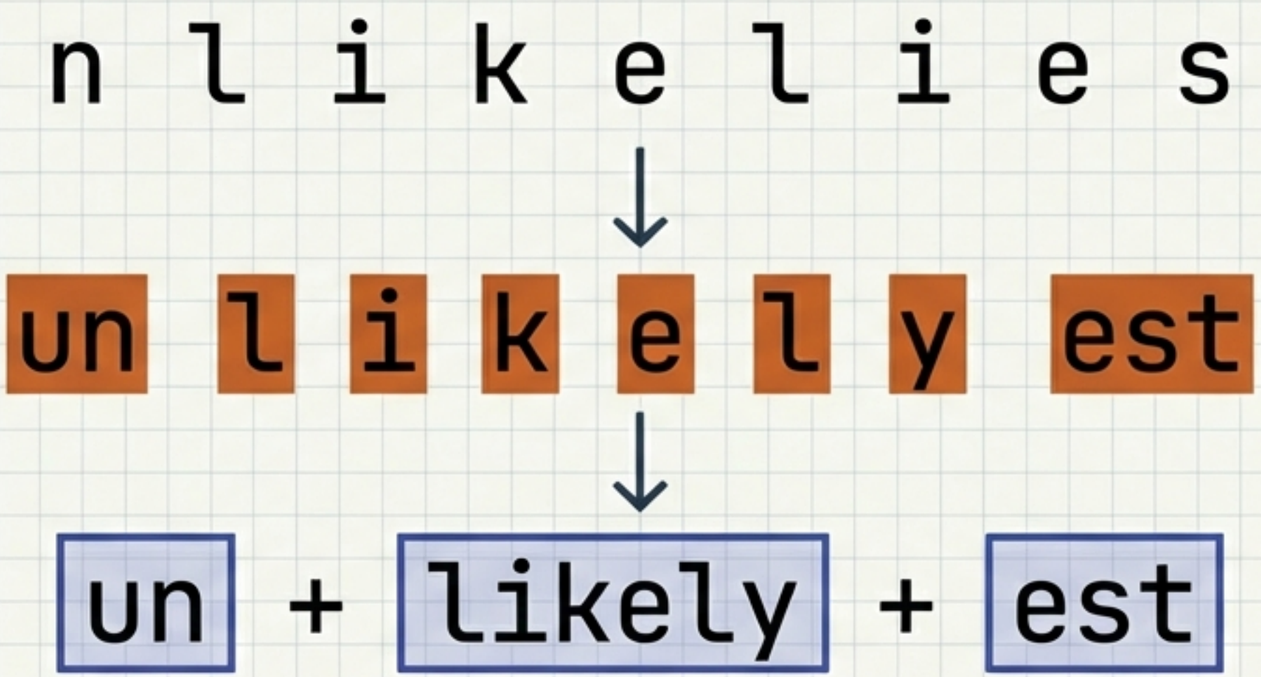
Key Insight
The vocabulary is learned once, then applied forever.

Inside the Learner: Building the Vocabulary



The Segmenter: Handling the Unknown

Scenario: Input Word: ‘Unlikeliest’ (Assume unseen in training).



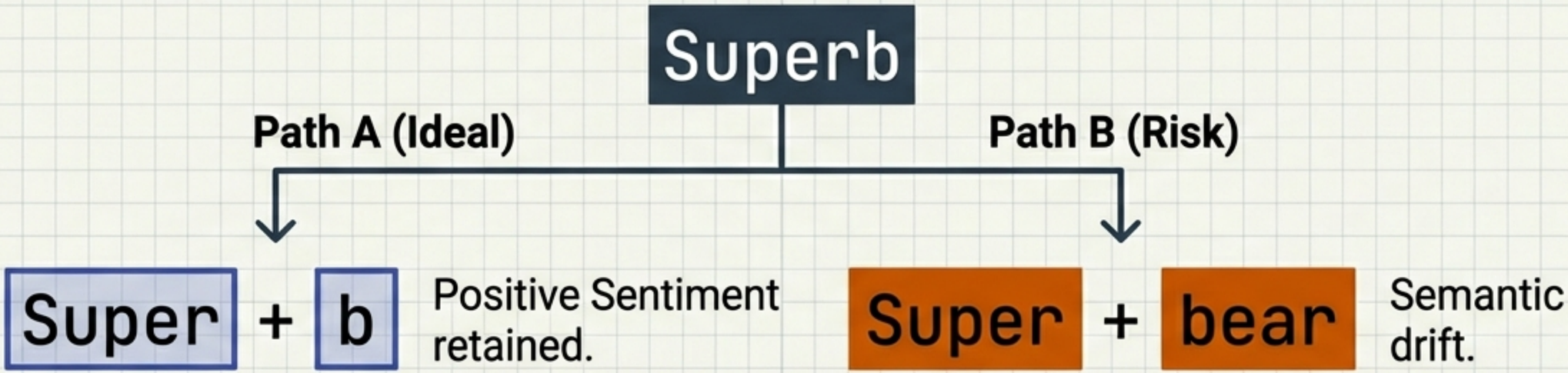
Takeaway: Greedy application of learned rules allows the model to process words it has never seen by breaking them into familiar sub-components (morphemes).

The Modern Tokenization Landscape

Model Family	Algorithm	Typical Vocab Size	Notes
GPT-2, GPT-4, LLaMA	Byte Pair Encoding (BPE)	50,000+	Standard for generation.
BERT	WordPiece	~30,000	Maximizes likelihood vs. frequency.
T5 / Multi-lingual	SentencePiece	32,000 – 250,000	Treats input as raw stream (no whitespace dependency).

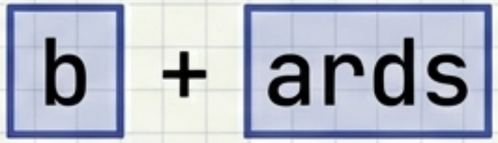
When Statistics Fail Linguistics

Quirks of the BPE approach.



Real World Example:

GPT-3.5 Tokenization of ‘bards’:

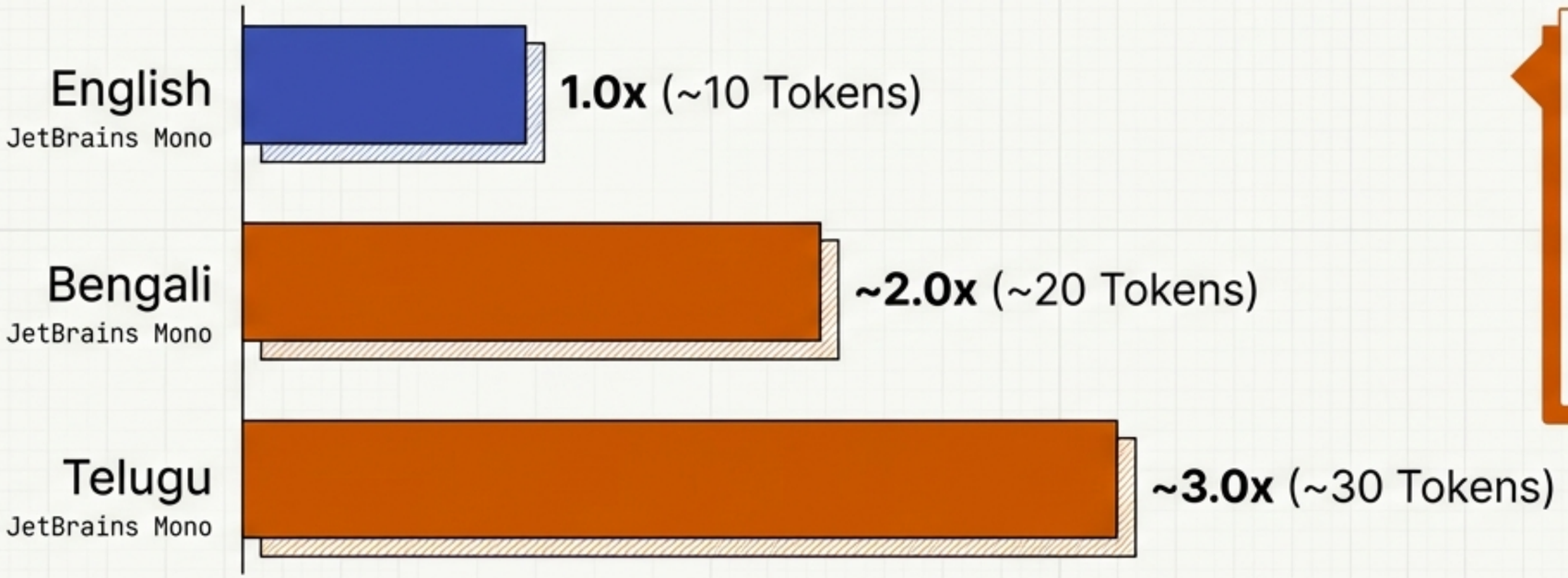


Not `bard` + `s`. BPE optimizes for compression, not human readability.

The Hidden Tax: Language Fairness

Tokens required to represent the same sentence

Roboto

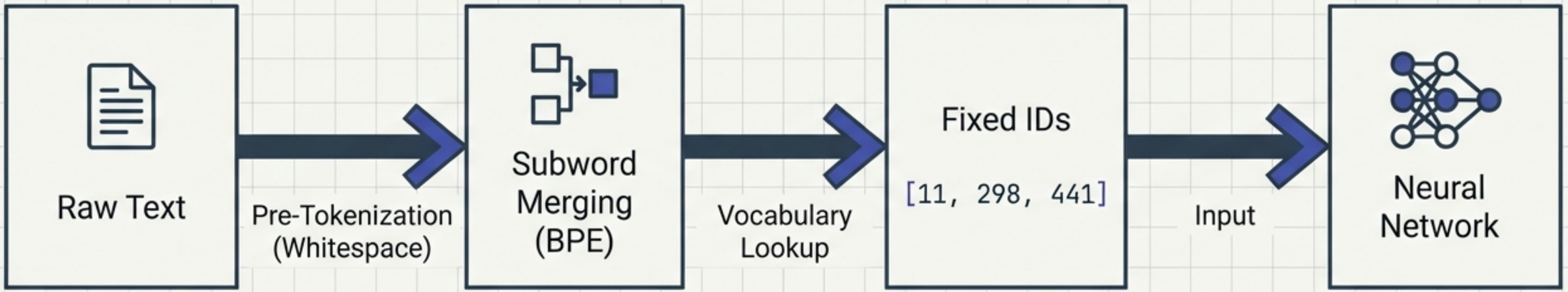


Cost & Latency: Processing Telugu costs ~3x more than English for the exact same information.

Rare languages are pulverized into characters because pairs aren’t frequent in training data.

Roboto

Summary: The Pipeline to Intelligence



Tokenization defines what the model sees, how well it generalizes, and how much it costs to run.

Beyond the Token?

TOKEN BLOCK
JetBrains Nono

> “Deep Learning has given tokenization a new birth.

As models scale, the definition of the ‘atomic unit’ may evolve.
But today, the **Subword** is the **currency of the AI economy**.