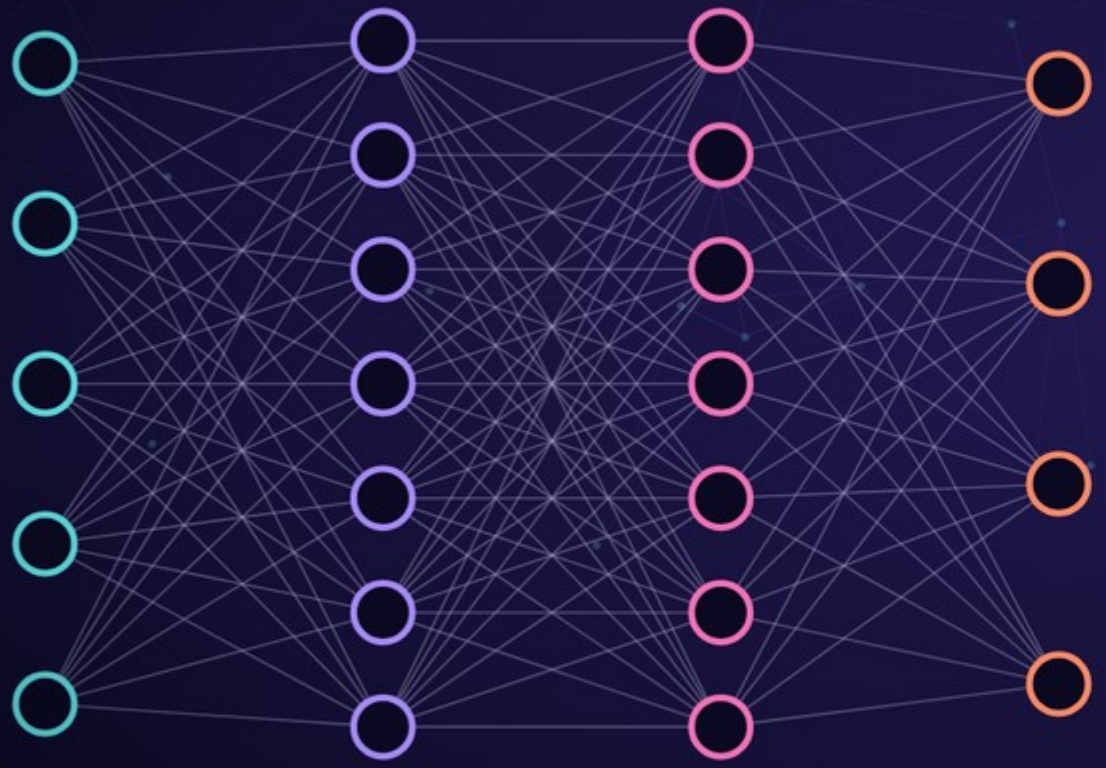
@MyWebUniversity

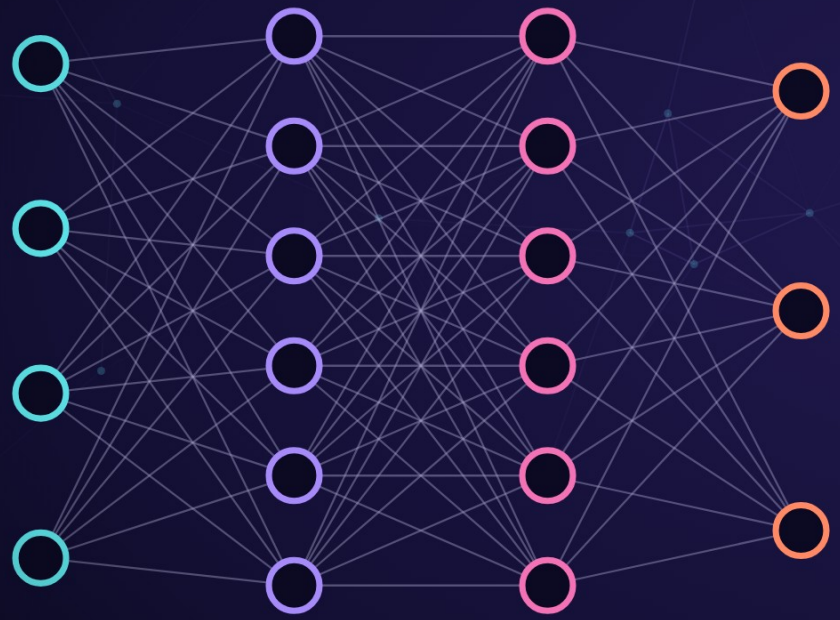
#2



LLM

Under the Hood

AI EXPLAINED · A FREE SERIES FOR STUDENTS AND DEVELOPERS



A MYWEBUNIVERSITY SERIES

AI EXPLAINED

EPISODE 2

How LLMs Work Under the Hood

tokens · neurons · attention

One Simple Job: Predict the Next Token



- Core job: predict the next token
- Write answers one token at a time
- Autoregressive generation

01

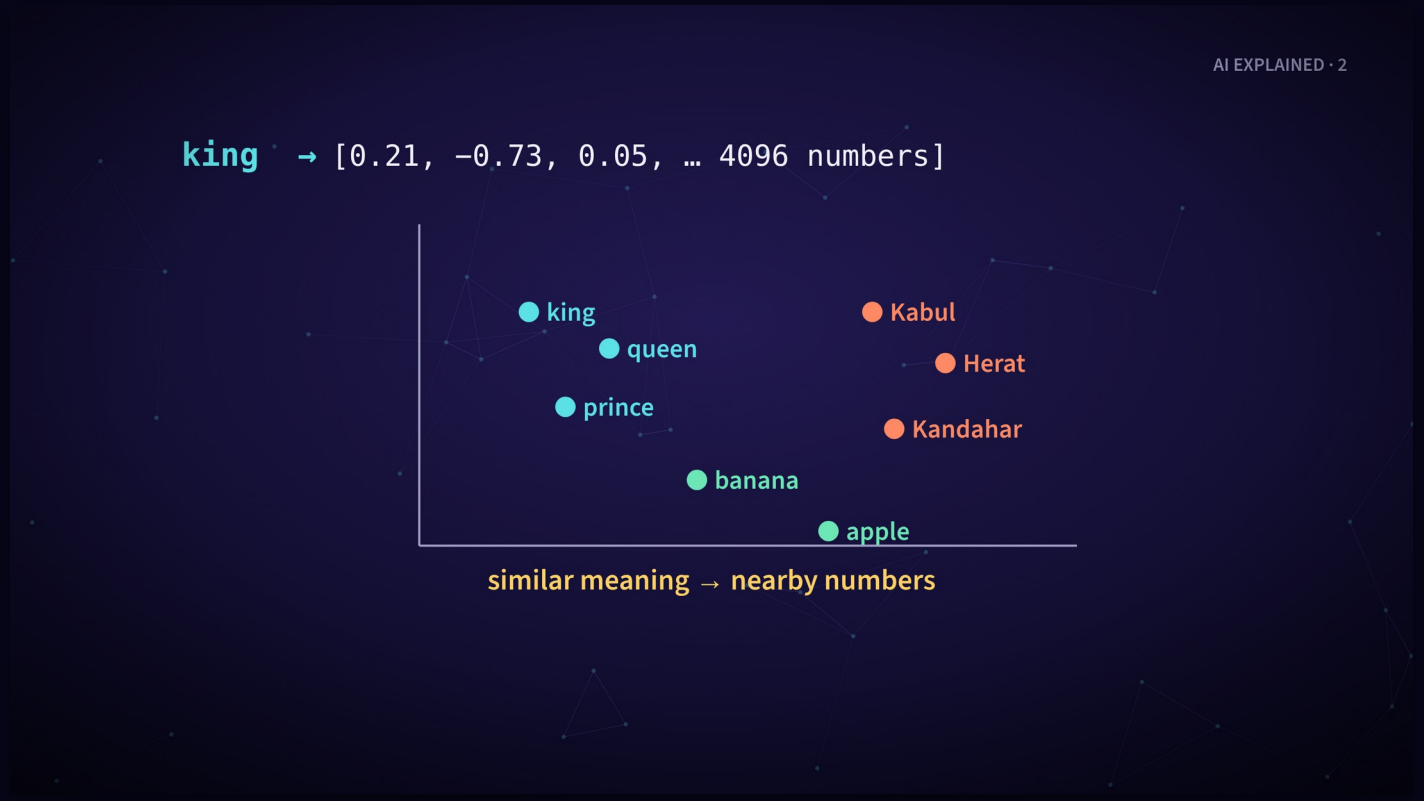
Tokens: How AI Reads Text



- Tokens = words, word pieces, symbols
- ≈ ¾ of an English word per token
- Each token → an ID number

02

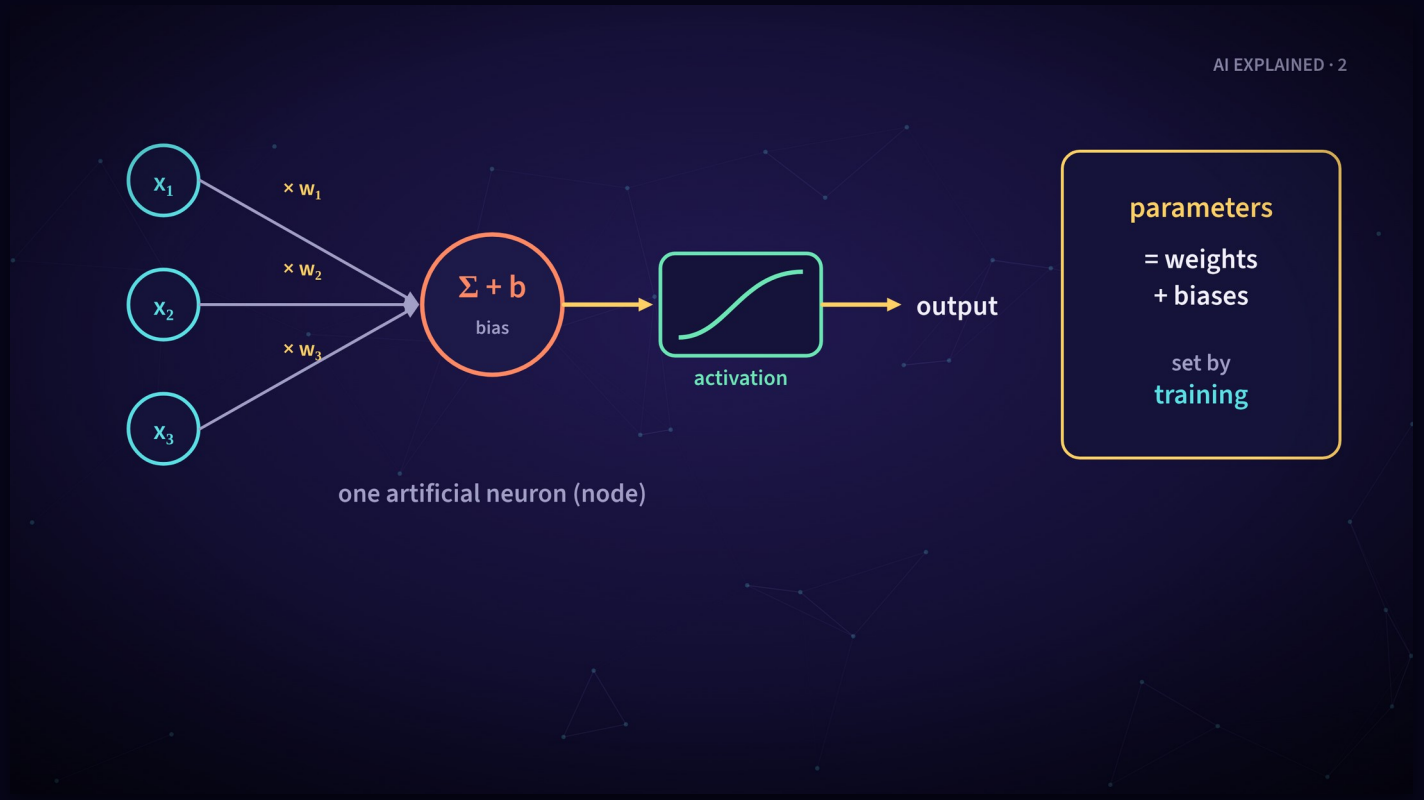
Embeddings: Meaning as Numbers



- Embedding = vector of numbers
- Similar meaning → nearby vectors
- Thousands of dimensions

03

Neurons, Weights and Biases



- Neuron: Σ(inputs × weights) + bias
- Activation function decides output
- Weights + biases = parameters

04

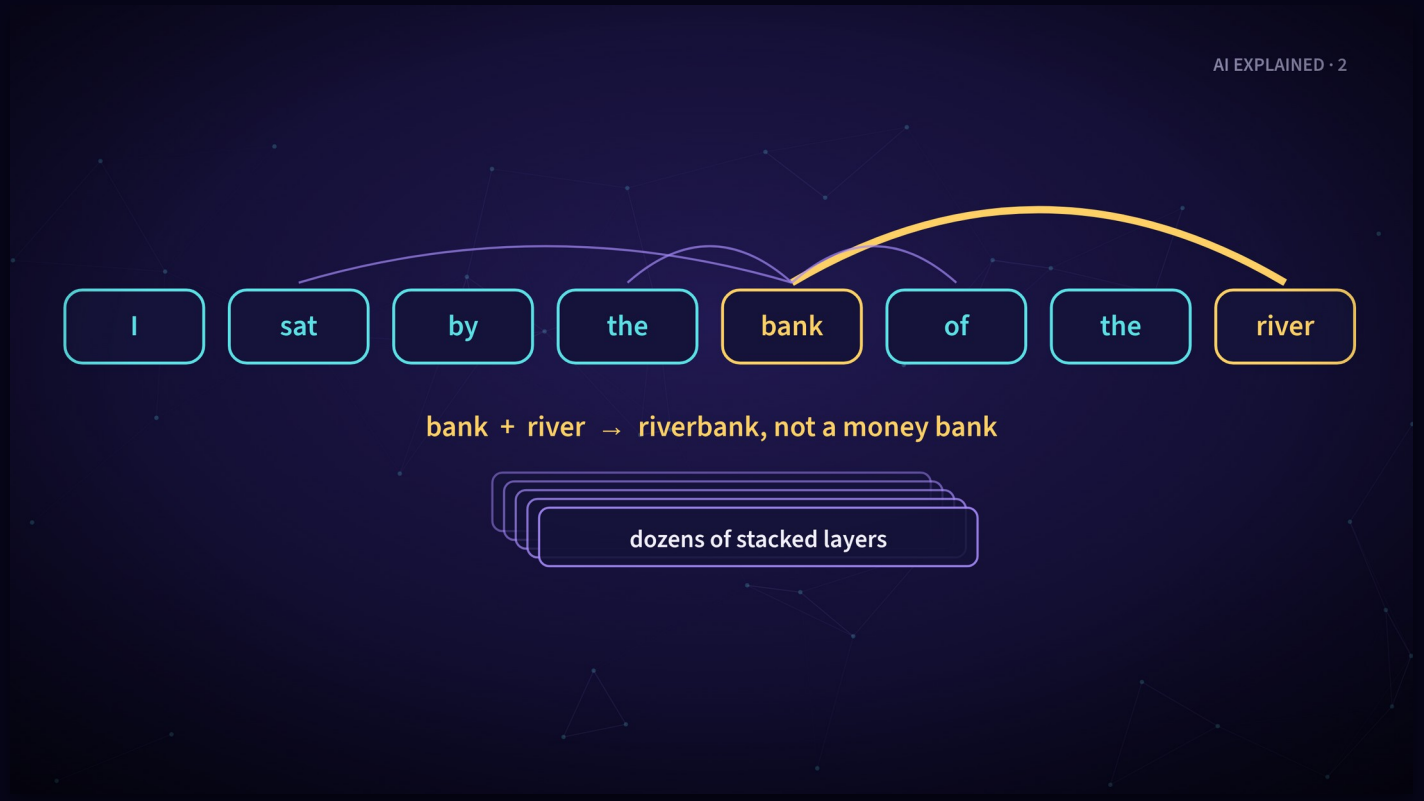
Billions of Parameters



- GPT-3: 175B parameters (2020)
- Small open models: 1–8B · largest: 100s of B to >1T
- Mixture of experts: only some experts per token

05

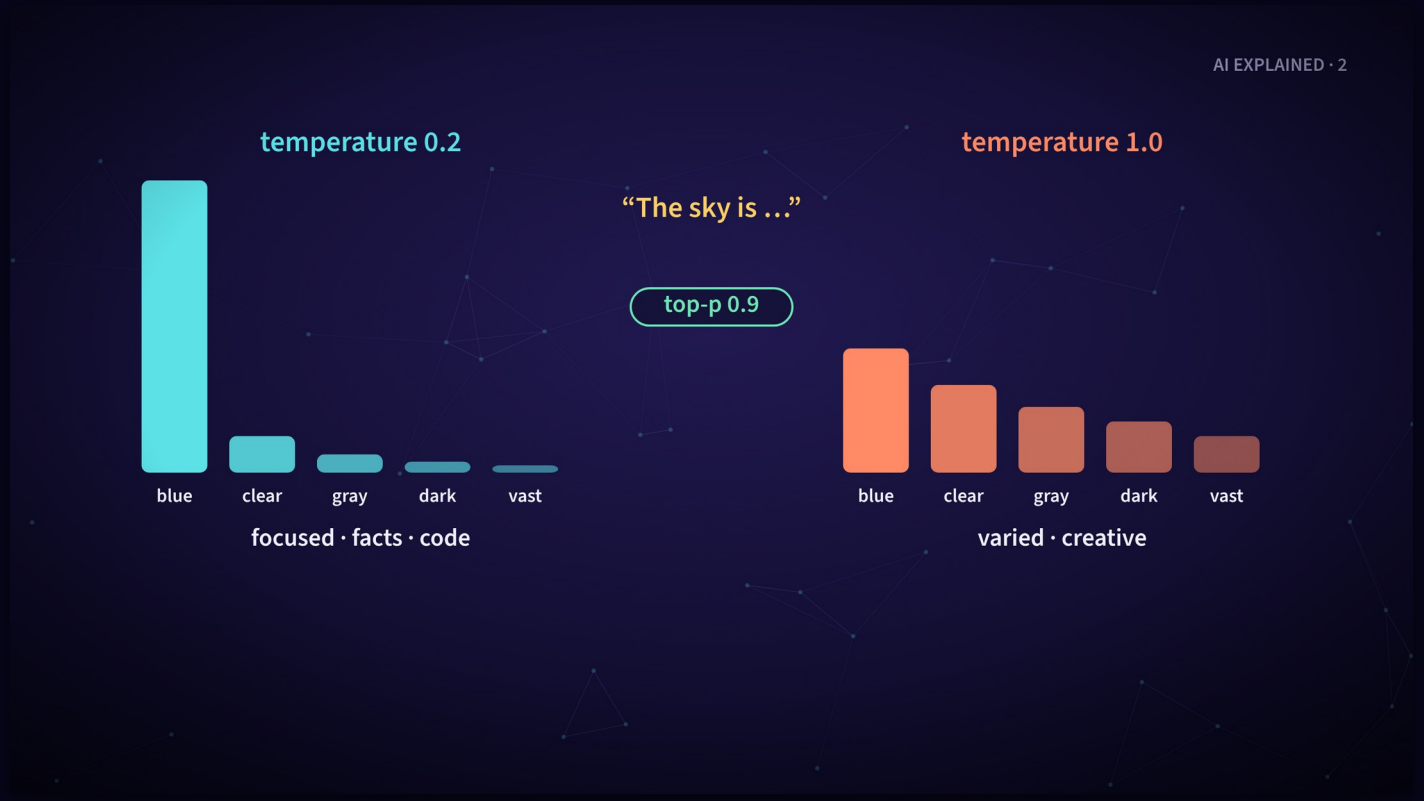
Attention: Which Words Matter?



- Attention: which tokens matter to each token
- “bank” + “river” → riverbank
- Dozens of stacked transformer layers

06

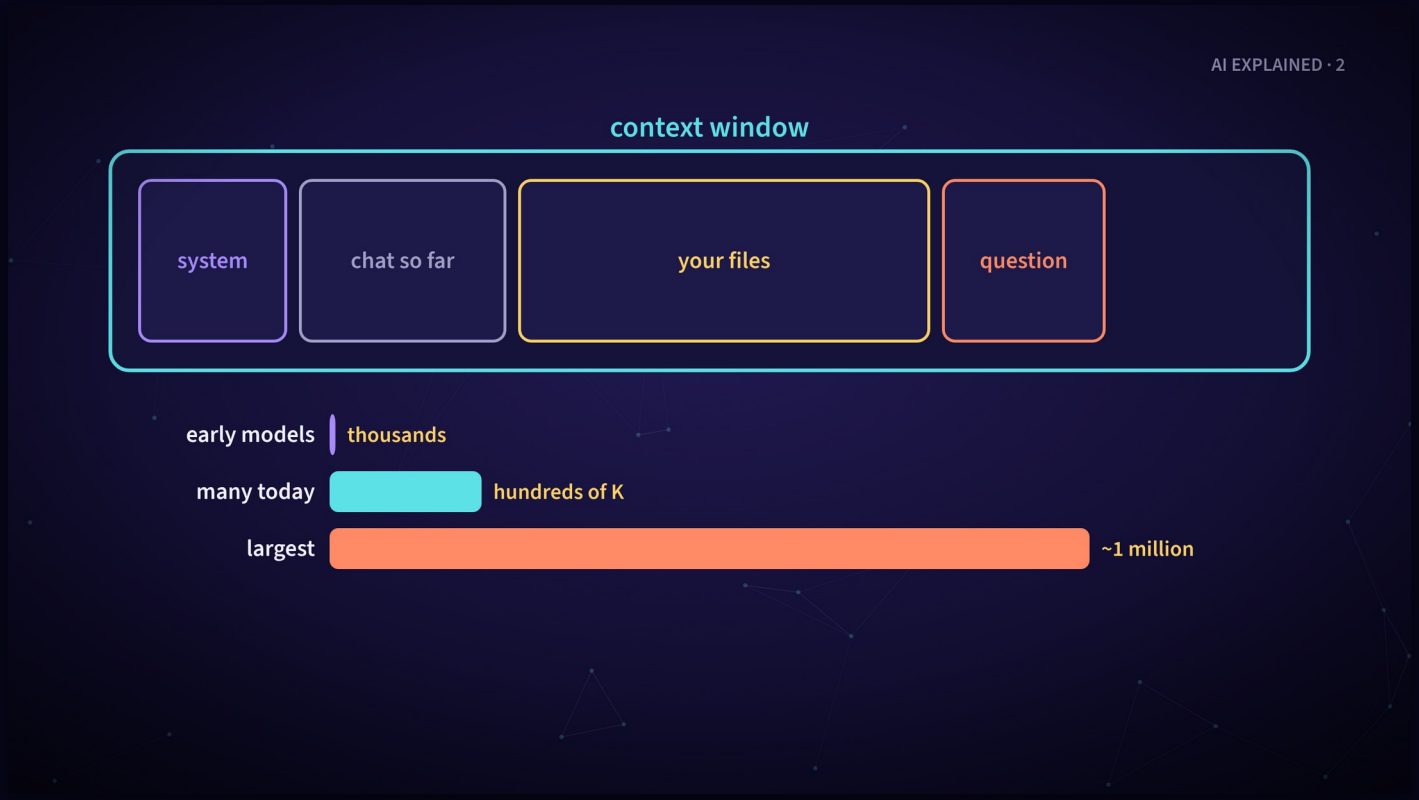
Probabilities, Temperature and Top-p



- Probability for every vocabulary token
- Temperature: low = focused, high = creative
- Top-p: sample from the most likely set

07

The Context Window



- Context window = tokens seen at once
- Thousands → hundreds of thousands → ~1M
- Older text summarized or dropped

08

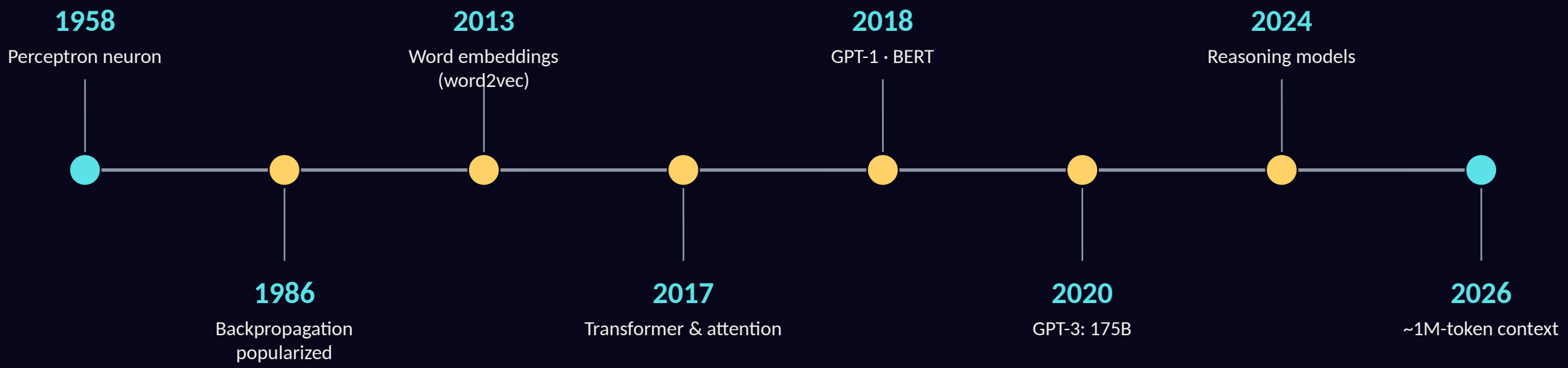
Why AI Can Be Wrong



- Not a database: can hallucinate
- Reasoning models think step by step
- Verify important facts and code

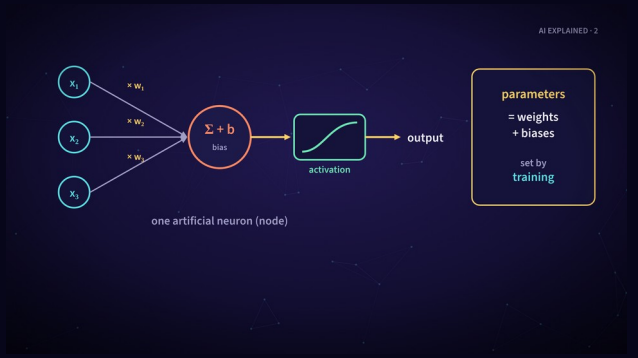
09

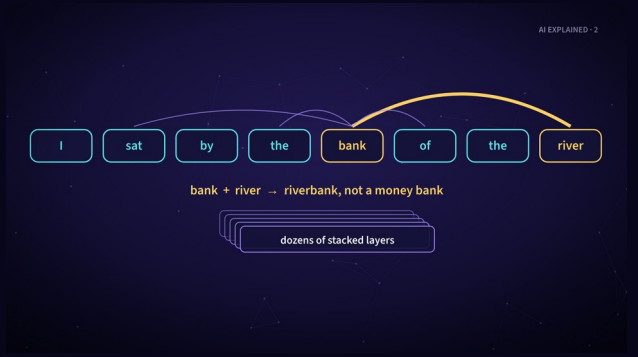
Timeline: 1958 to 2026



Sources and Further Reading

1. Vaswani et al., “Attention Is All You Need,” 2017
2. Mikolov et al., “Efficient Estimation of Word Representations” (word2vec), 2013
3. Brown et al., “Language Models are Few-Shot Learners” (GPT-3), 2020
4. Anthropic, “Tokens and context windows” — platform.claude.com/docs
5. 3Blue1Brown, “Neural Networks” video series — 3blue1brown.com





Thank you for watching

MyWebUniversity.com

Free technical education since 2000 — English and Dari

@MyWebUniversity · @MyWebUniversity-Dari