Multi-modal GenAI with Transformers

GEMINI 2.5

PROMPT: Tell me a “crossing the road” joke about Transformers.

Why did the Transformer cross the road?

•To get to the other AutoBot.

•To avoid the Decepticons.

GEMINI 3

PROMPT: Tell me a “crossing the road” joke about Transformers.

Q. Why did the Autobot cross the road?

A. To prove to the Decepticons that there was **more to the street than meets the eye!**

PROMPT: Translate the joke to Spanish.

P. ¿Por qué el Autobot cruzó la calle?

R. ¡Para demostrarle a los Decepticons que había **más de lo que el ojo puede ver!**

MIDJOURNEY

PROMPT: a Singapore gen z at a thrift shop in the style of Roy Lichtenstein

Sketch2Sound: My lab (now in Adobe Firefly)

García, H.F., Nieto, O., Salamon, J., Pardo, B. and Seetharaman, P. Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations. *ICASSP 2025*.

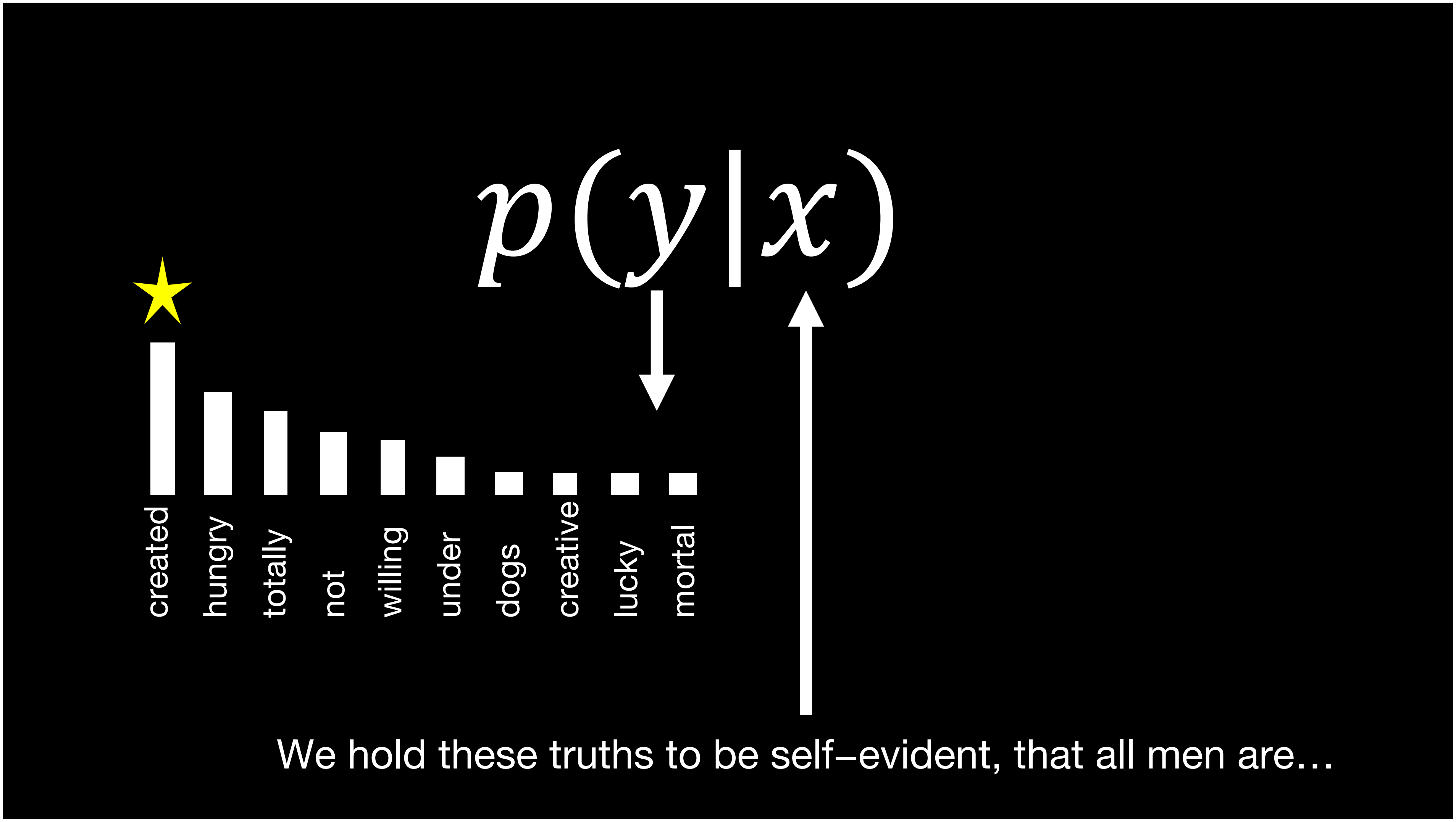
How do they work?

Everything is conditional

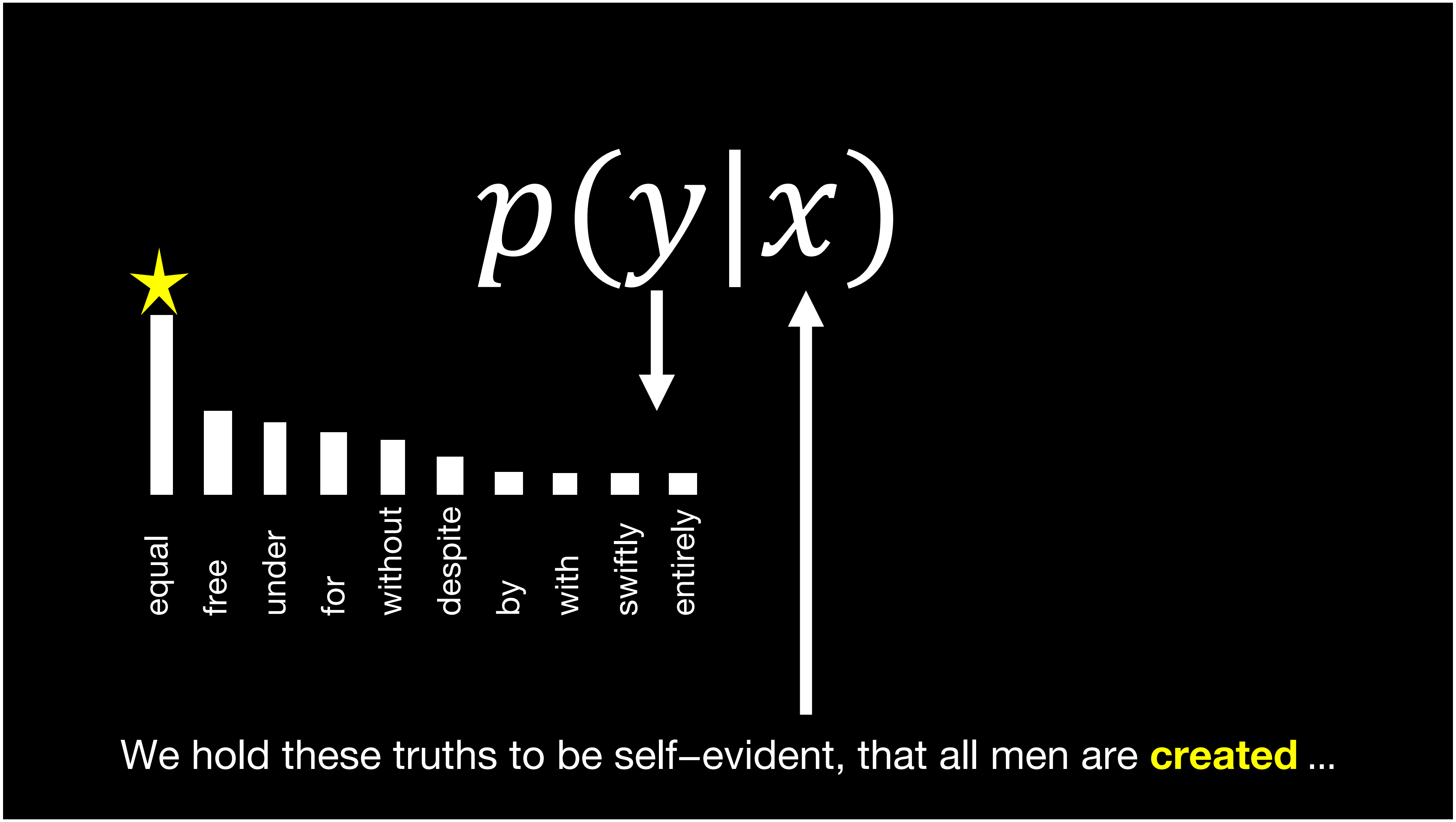
$$p(y|x)$$

The output

The input
(aka “conditioning”)



$p(y|x)$
created
hungry
totally
not
willing
under
dogs
creative
lucky
mortal
We hold these truths to be self–evident, that all men are…


$p(y|x)$
equal
free
under
for
without
despite
by
with
swiftly
entirely
We hold these truths to be self–evident, that all men are **created** ...

Labels? We don’t need no stinking labels

$$p(y|x)$$

$\overbrace{\text{We hold these}}^{x}\ \overbrace{\text{truths}}^{y}$ to be self–evident, that all men are created equal

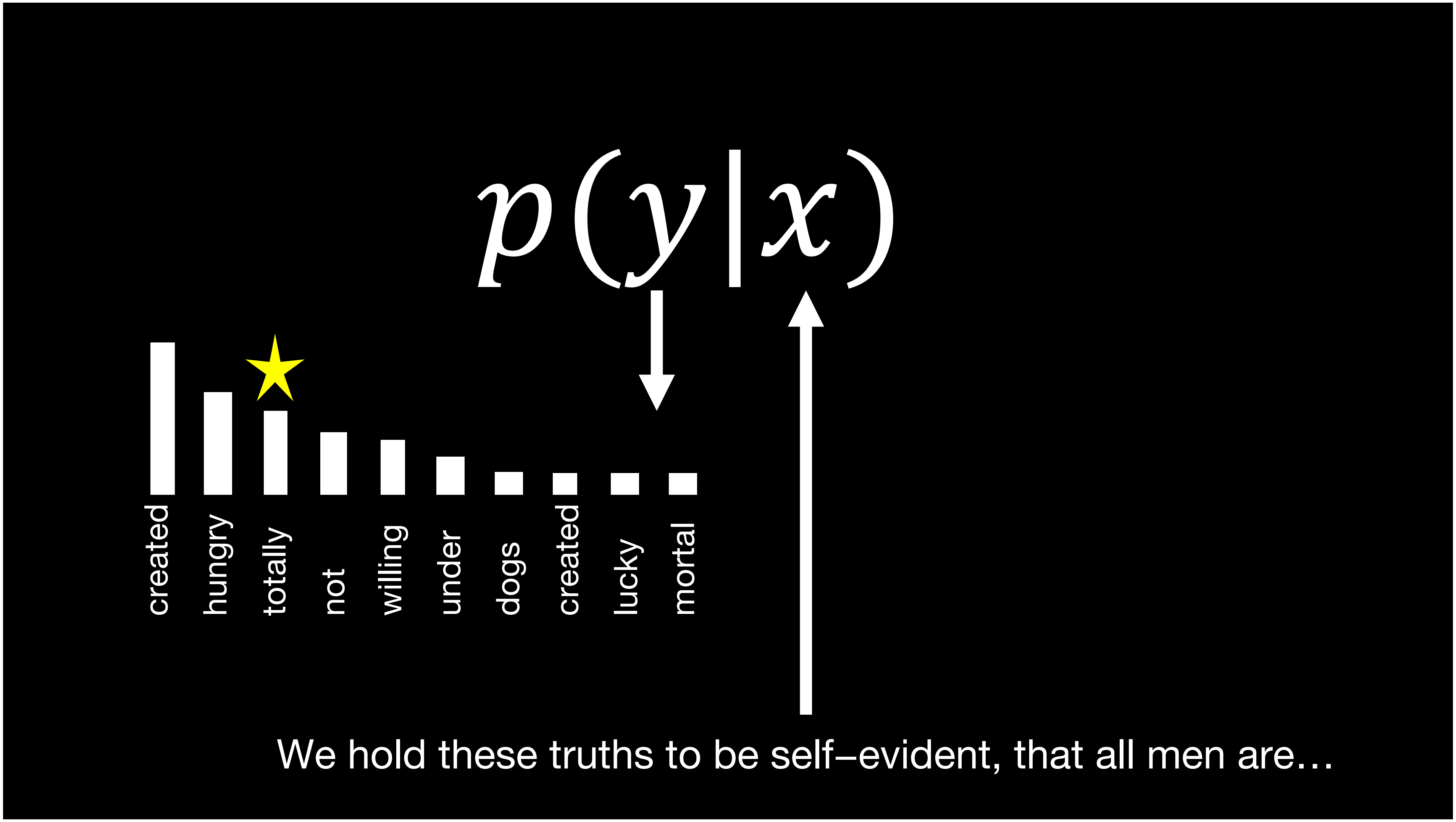
$$p(y|x)$$

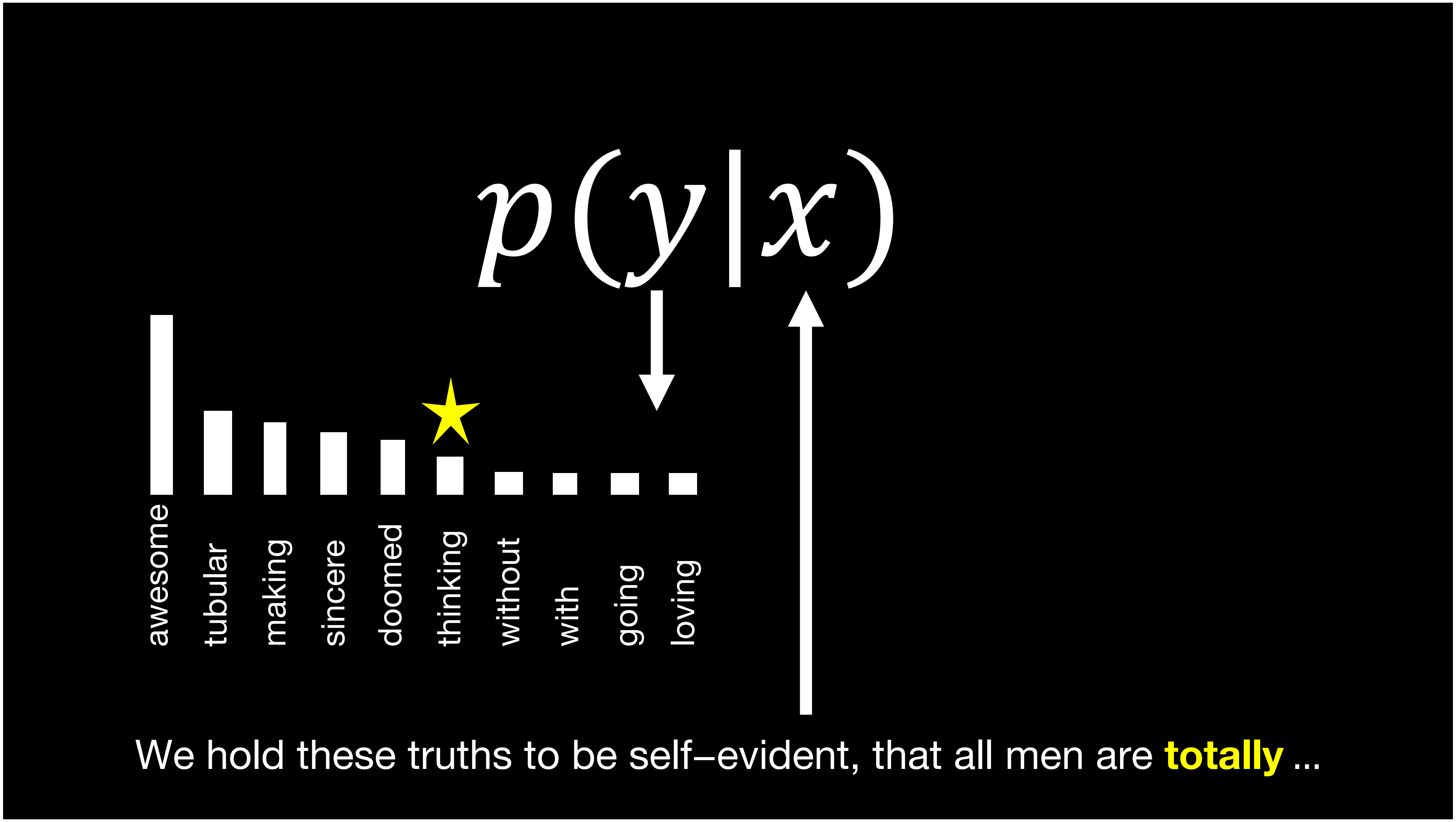
$\overbrace{\text{We hold these truths}}^{x}\ \overbrace{\text{to}}^{y}$ be self–evident, that all men are created equal

$$p(y|x)$$

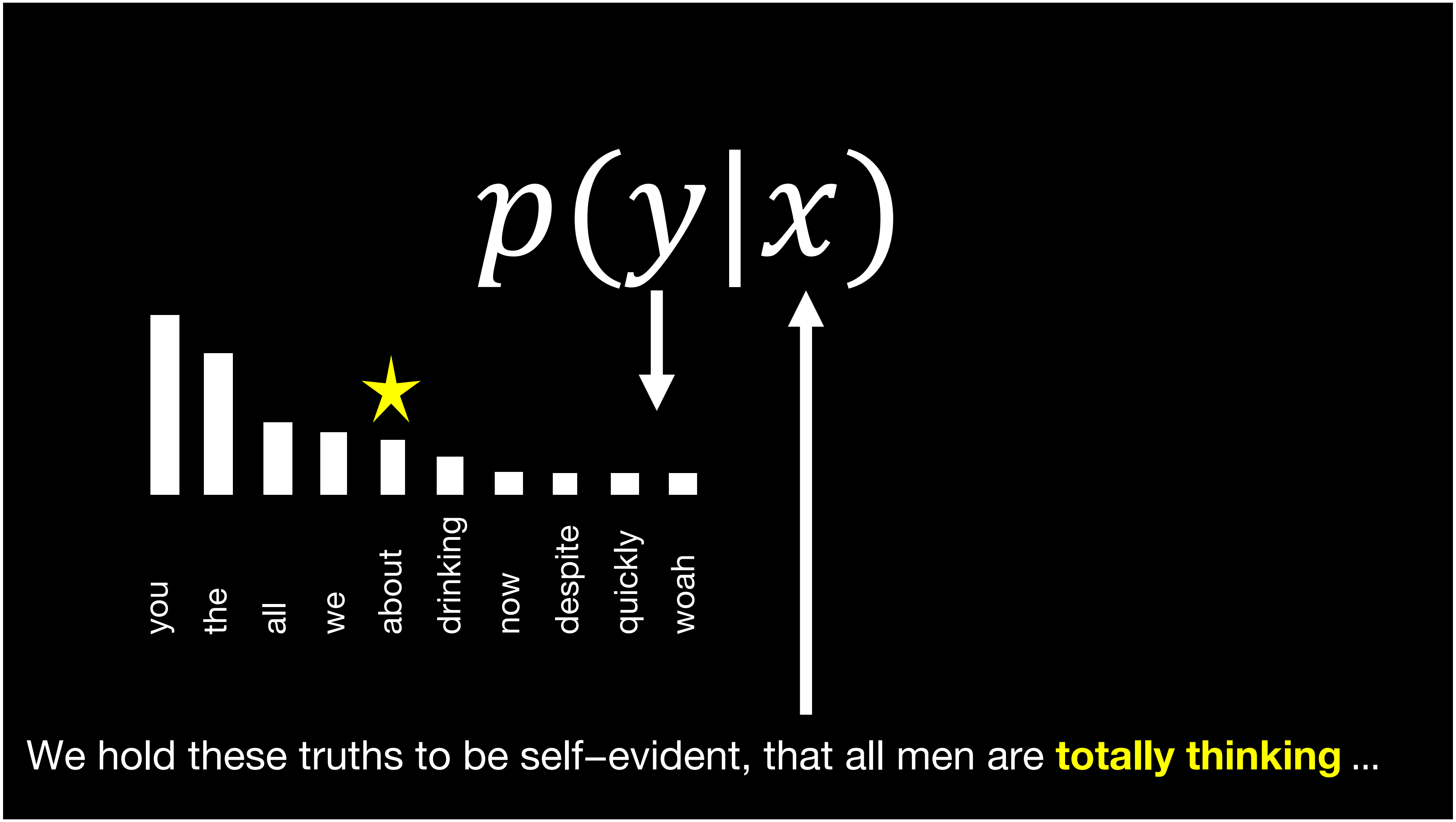
$$\overbrace{\text{We hold these truths to}}^{x}\ \overbrace{\text{be}}^{y}\ \text{self–evident, that all men are created equal}$$

Random walk = creativity?

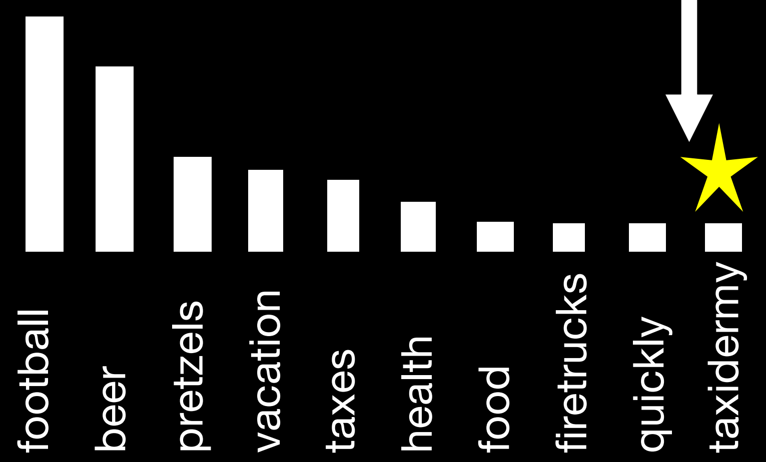

$p(y|x)$
created
hungry
totally
not
willing
under
dogs
created
lucky
mortal
We hold these truths to be self–evident, that all men are…



$p(y|x)$
awesome
tubular
making
sincere
doomed
thinking
without
with
going
loving
We hold these truths to be self–evident, that all men are **totally** ...


$p(y|x)$
you
the
all
we
about
drinking
now
despite
quickly
woah
We hold these truths to be self–evident, that all men are **totally thinking** ...

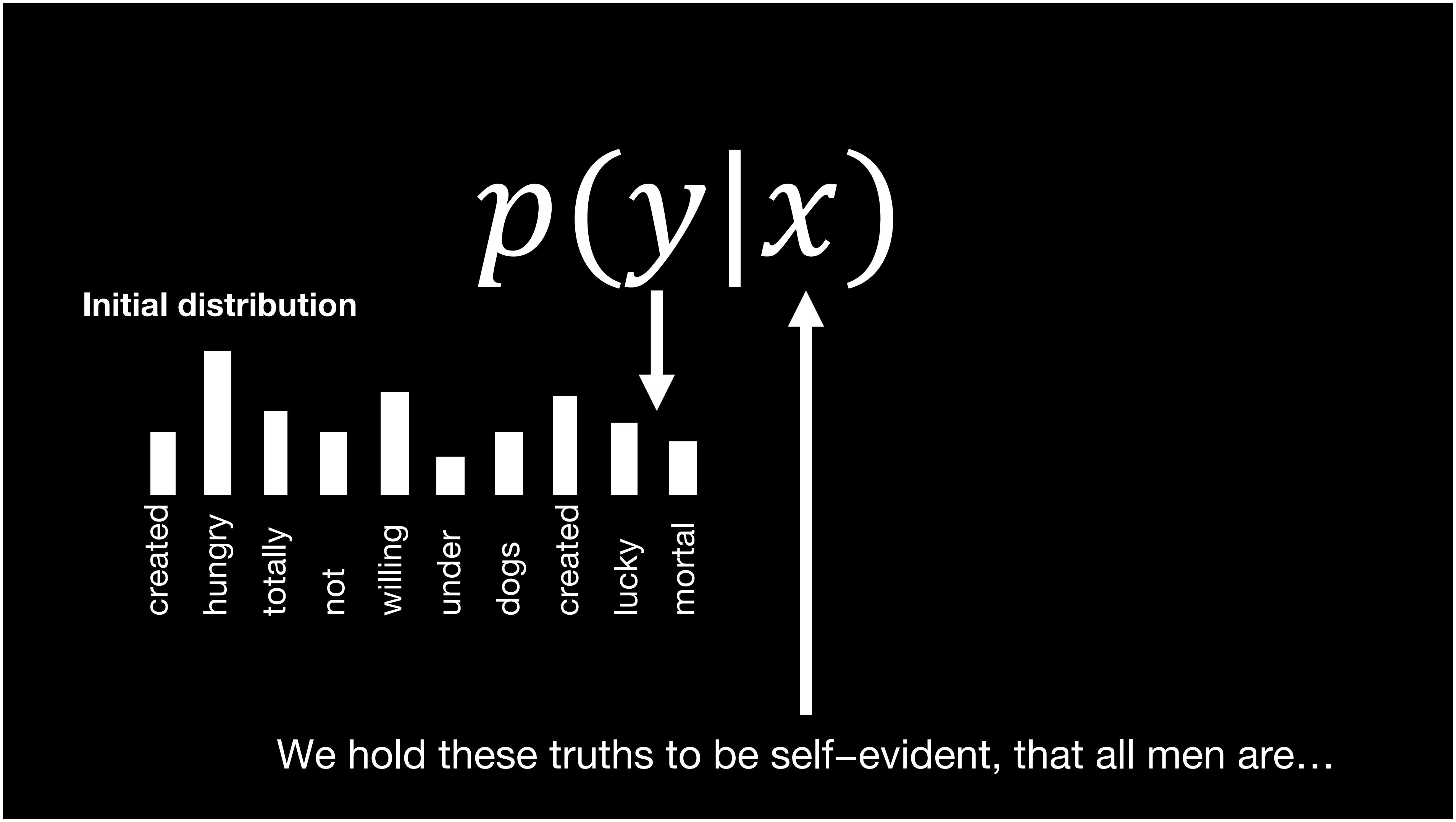
$$p(y|x)$$



We hold these truths to be self–evident, that all men are **totally thinking about**…

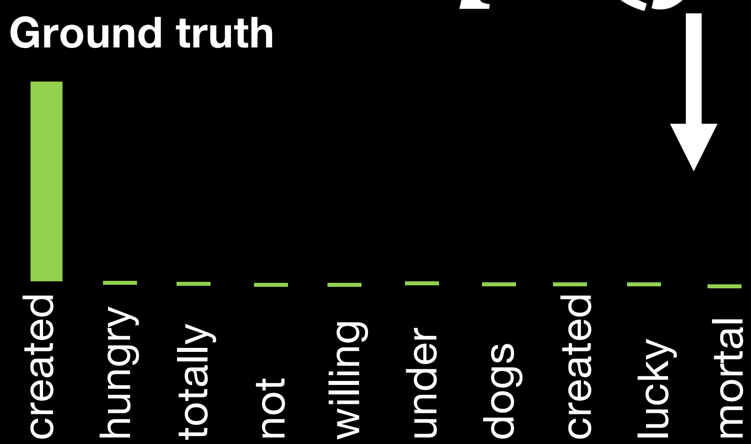
We hold these truths to be self–evident, that all men are **totally thinking about taxidermy.**

Their loss is our gain

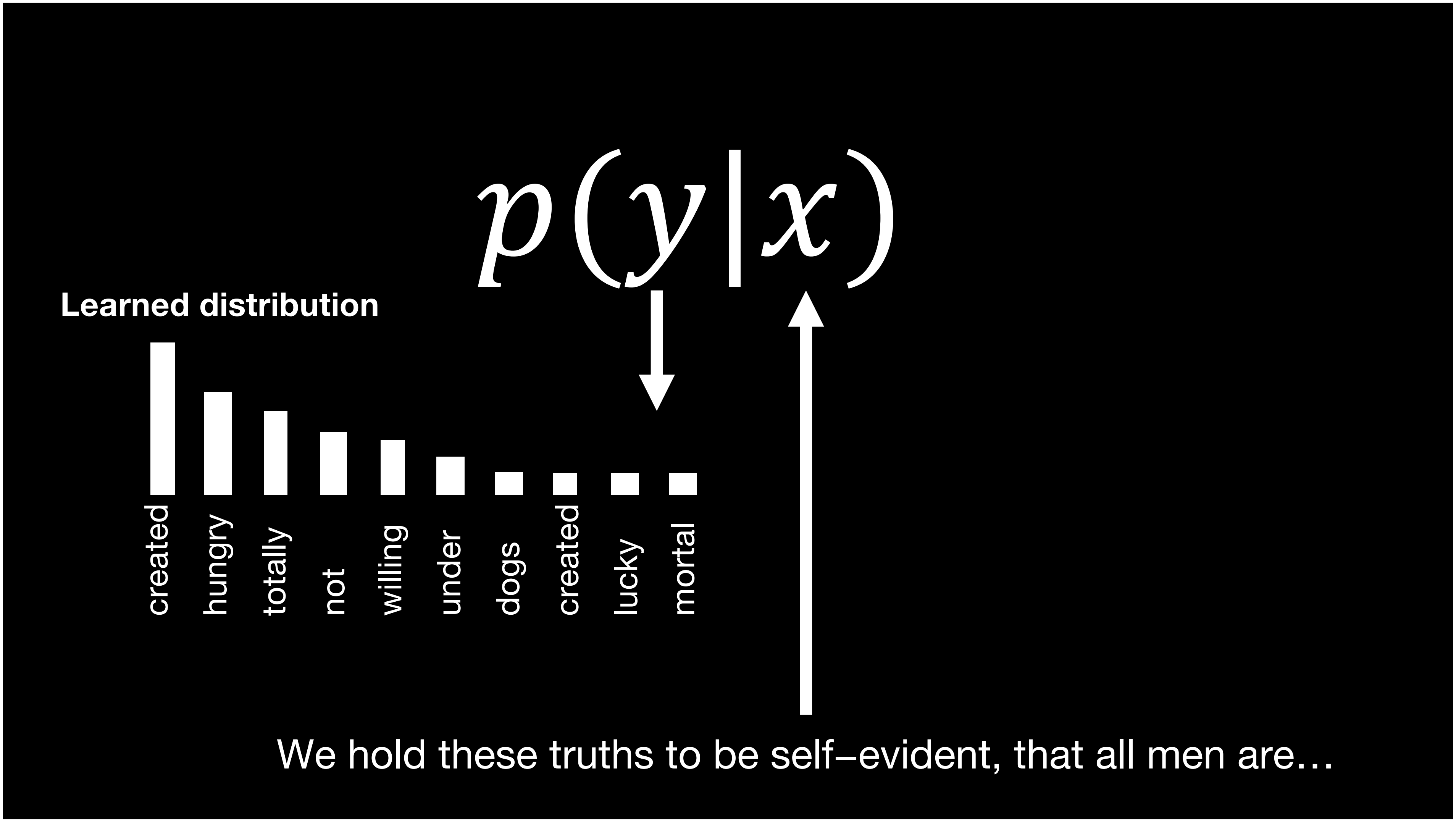


$p(y|x)$
Initial distribution
created
hungry
totally
not
willing
under
dogs
created
lucky
mortal
We hold these truths to be self–evident, that all men are…

$$p(y|x)$$



We hold these truths to be self–evident, that all men are…


$p(y|x)$
Learned distribution
created
hungry
totally
not
willing
under
dogs
created
lucky
mortal
We hold these truths to be self–evident, that all men are…

$$p(y|x)$$

Overfit (mode collapse)

We hold these truths to be self–evident, that all men are…

We hold these truths to be self–evident, that all men are…

Gibberish: …but are not but are but are not but are…

Creative: …totally thinking about taxidermy.

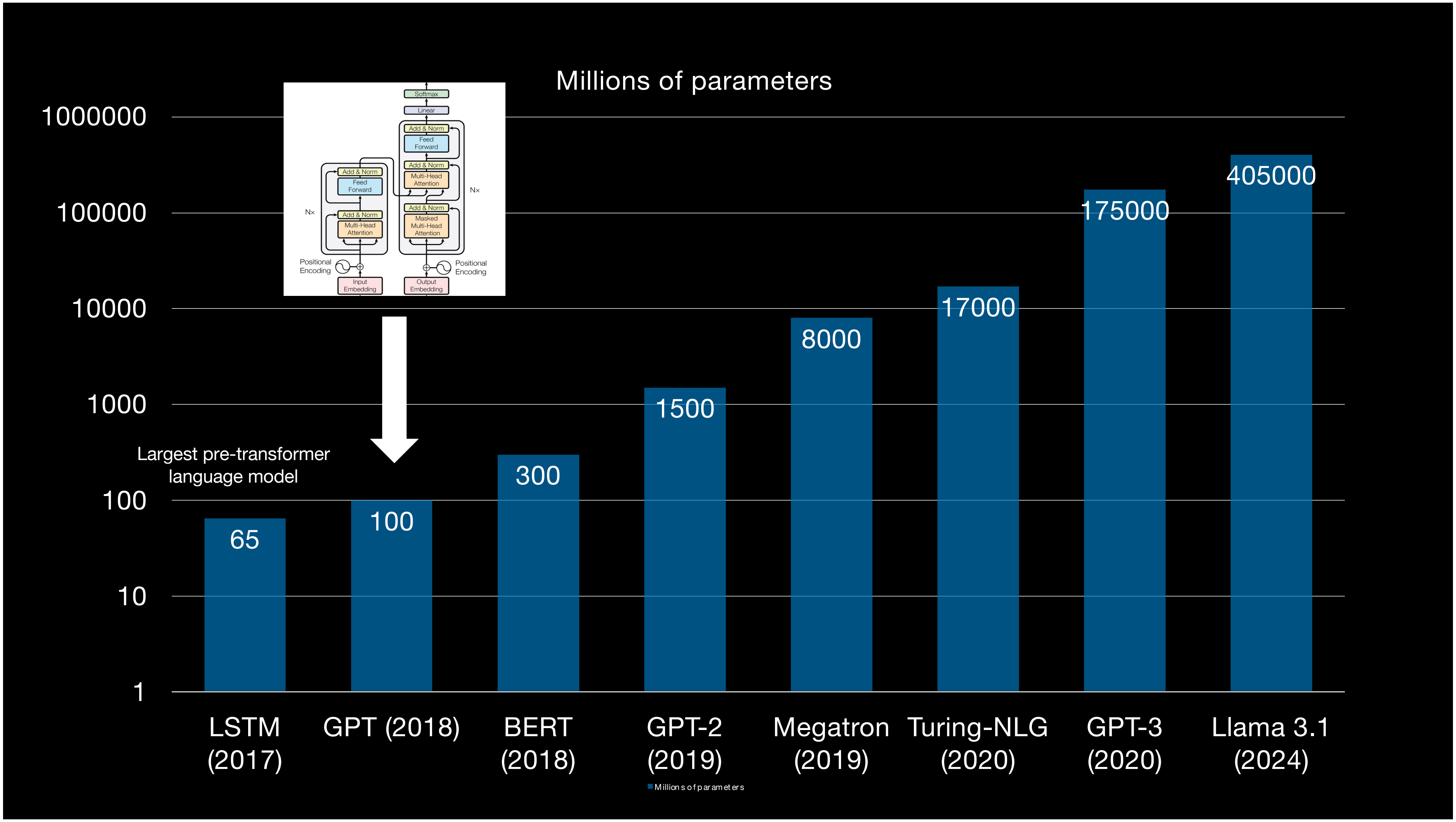
Copyright violating: ...that all men are created equal, but some are more equal than others.

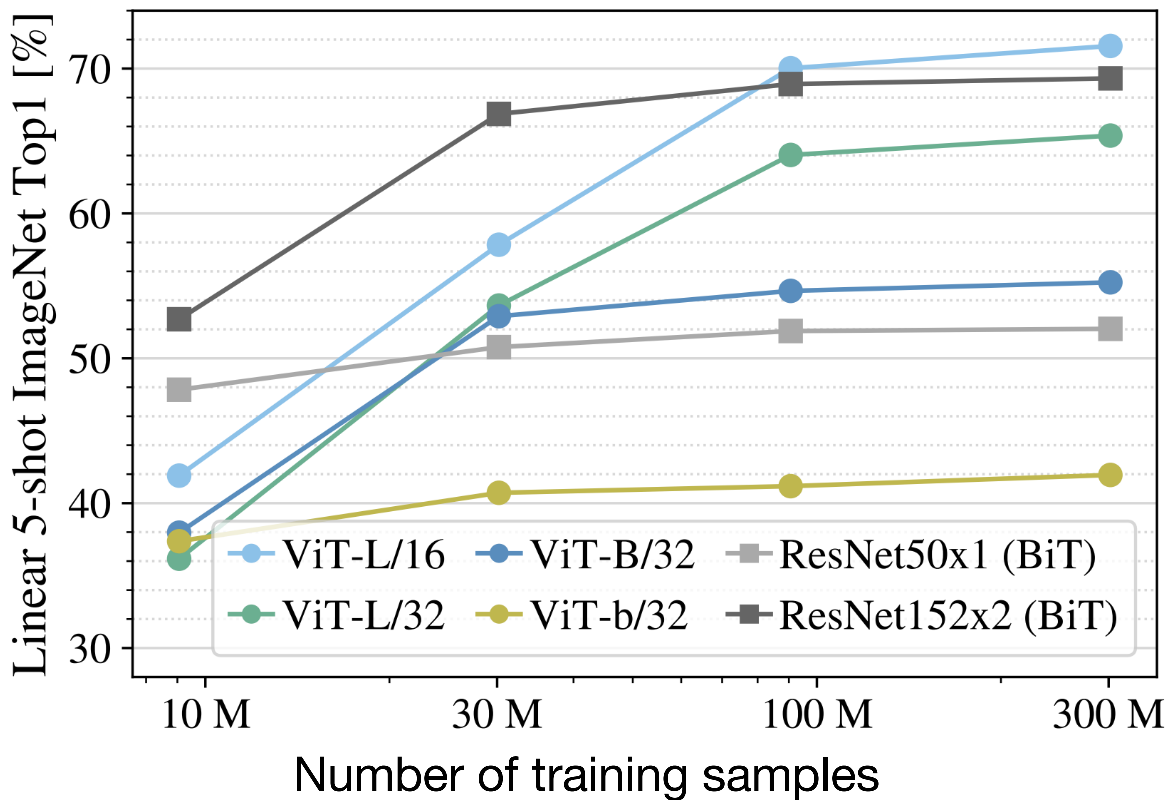
Correct?: …created equal, that they are endowed, by their Creator, with certain unalienable rights, that among these are Life, Liberty, and the pursuit of Happiness.

It’s a balance

- Learn the distribution too well and only reproduce your training data
- Learn it too poorly and your output is gibberish
- Sample with argmax and violate copyright
- Sample more randomly and risk hallucination

Bigger is better


Millions of parameters
1000000
100000
10000
1000
100
10
1
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Largest pre-transformer language model
65
100
300
1500
8000
17000
175000
405000
LSTM (2017)
GPT (2018)
BERT (2018)
GPT-2 (2019)
Megatron (2019)
Turing-NLG (2020)
GPT-3 (2020)
Llama 3.1 (2024)
Millions of parameters



ViT = Transformer ResNet = CNN

Dosovitskiy et al, ICLR 2021

Tokenizing is transformative

Words are too hard to model

- Consider all variants: “talk”, “talking”, “talks”, “talked”
- Can we break things into functional units?

“The extraterrestrial walked or is walking or maybe walks autonomously”

[The extra# terrestrial walk# #ed or is walk# #ing or maybe walk# #s autonomous# #ly]

What Gemini Says

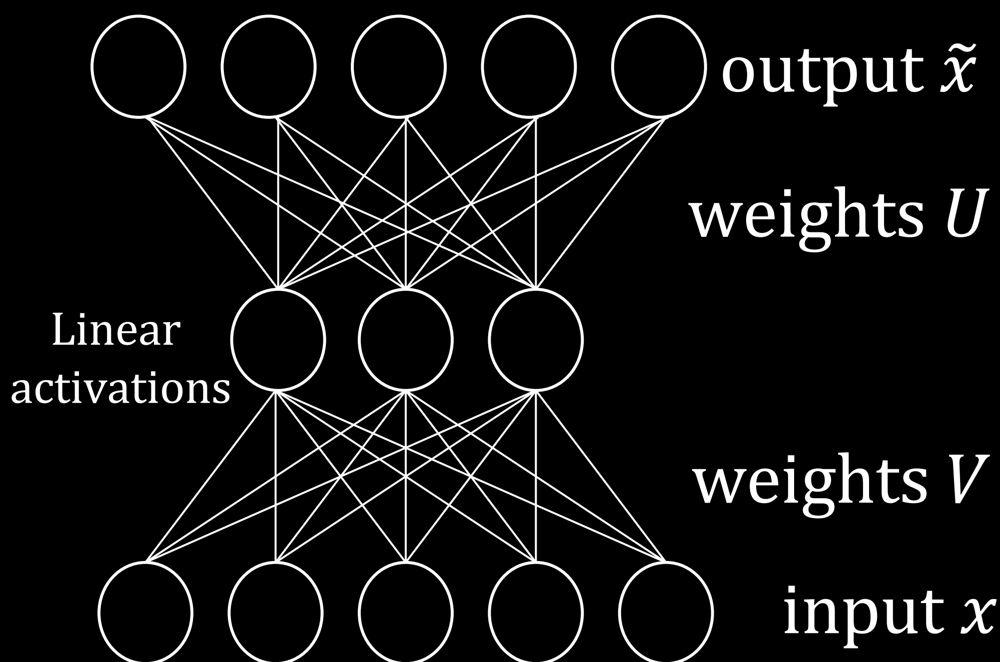
Feature	Word-based Tokenization	Subword Tokenization
Vocabulary size	Creates a huge vocabulary to account for every unique word and its variations, like plurals or different tenses.	Keeps vocabulary size manageable by breaking down rare words into shared subword units.
Compactness	Less compact, especially when dealing with morphologically rich languages or text with many rare words.	More compact, as it reuses common subword units across different words. For example, "unexpectedly" is tokenized into "un", "##expect", and "##edly".
OOV handling	Treats any word not seen during training as an unknown (`<UNK>`) token, causing information loss.	Can process unseen words by breaking them into known subwords, allowing the model to make informed predictions.

You can tokenize any kind of data

How, exactly?

(Variational) Autoencoders

A simple autoencoder

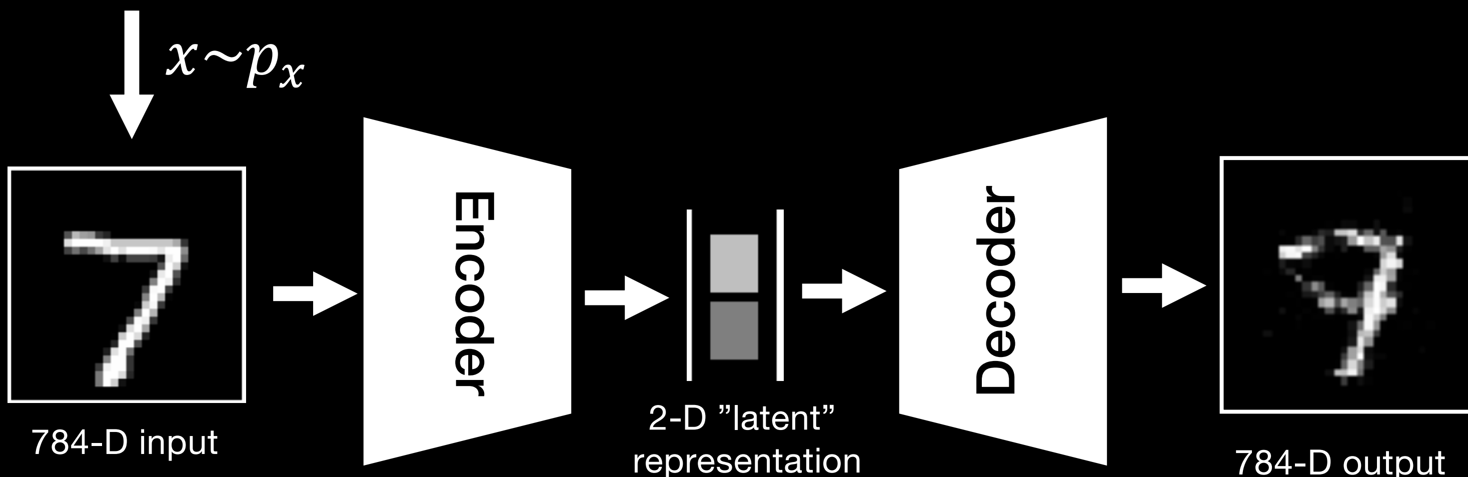


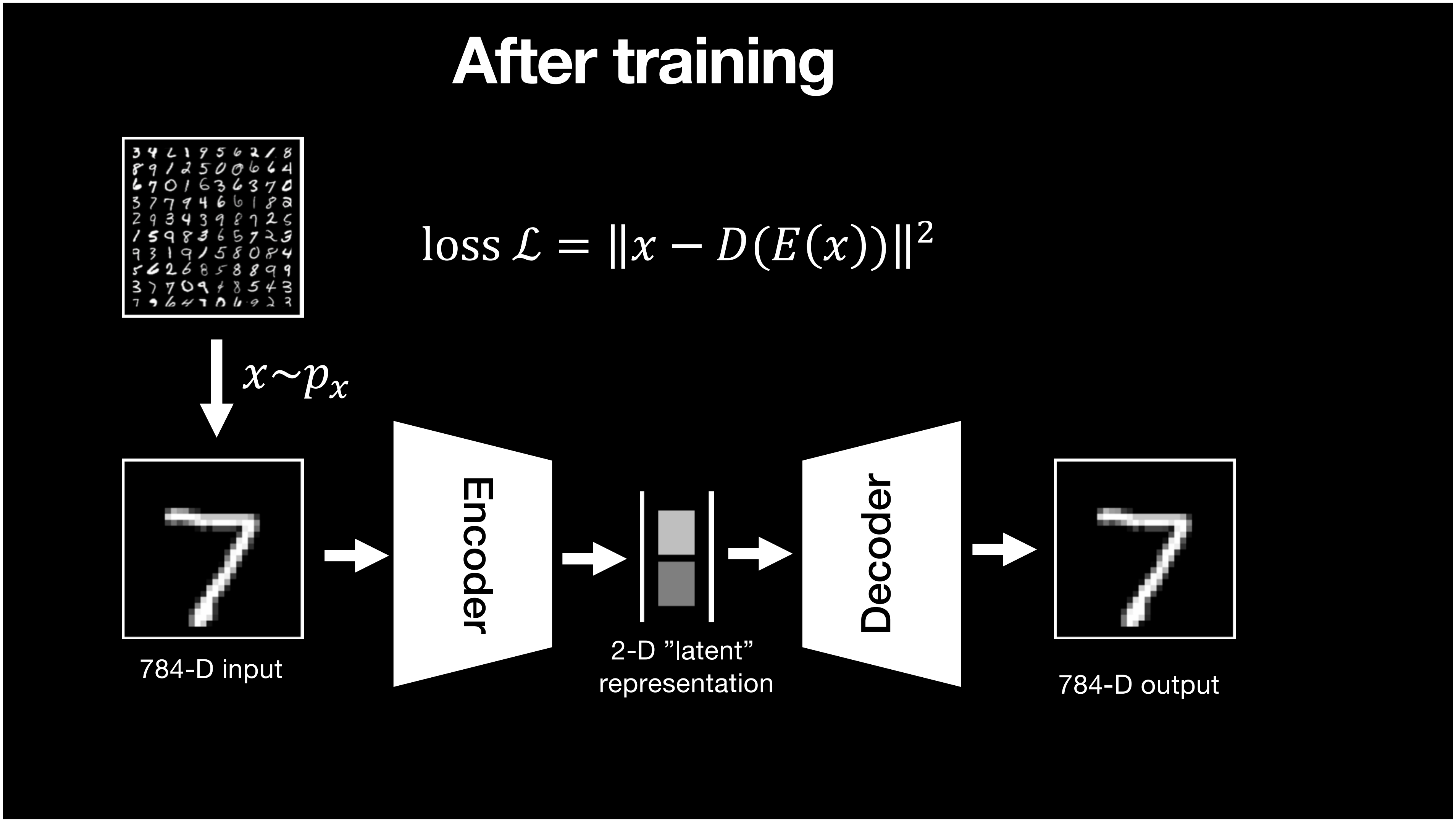
$$\text{loss } \mathcal{L} = \|x - \tilde{x}\|^2$$
$$\tilde{x} = UVx$$

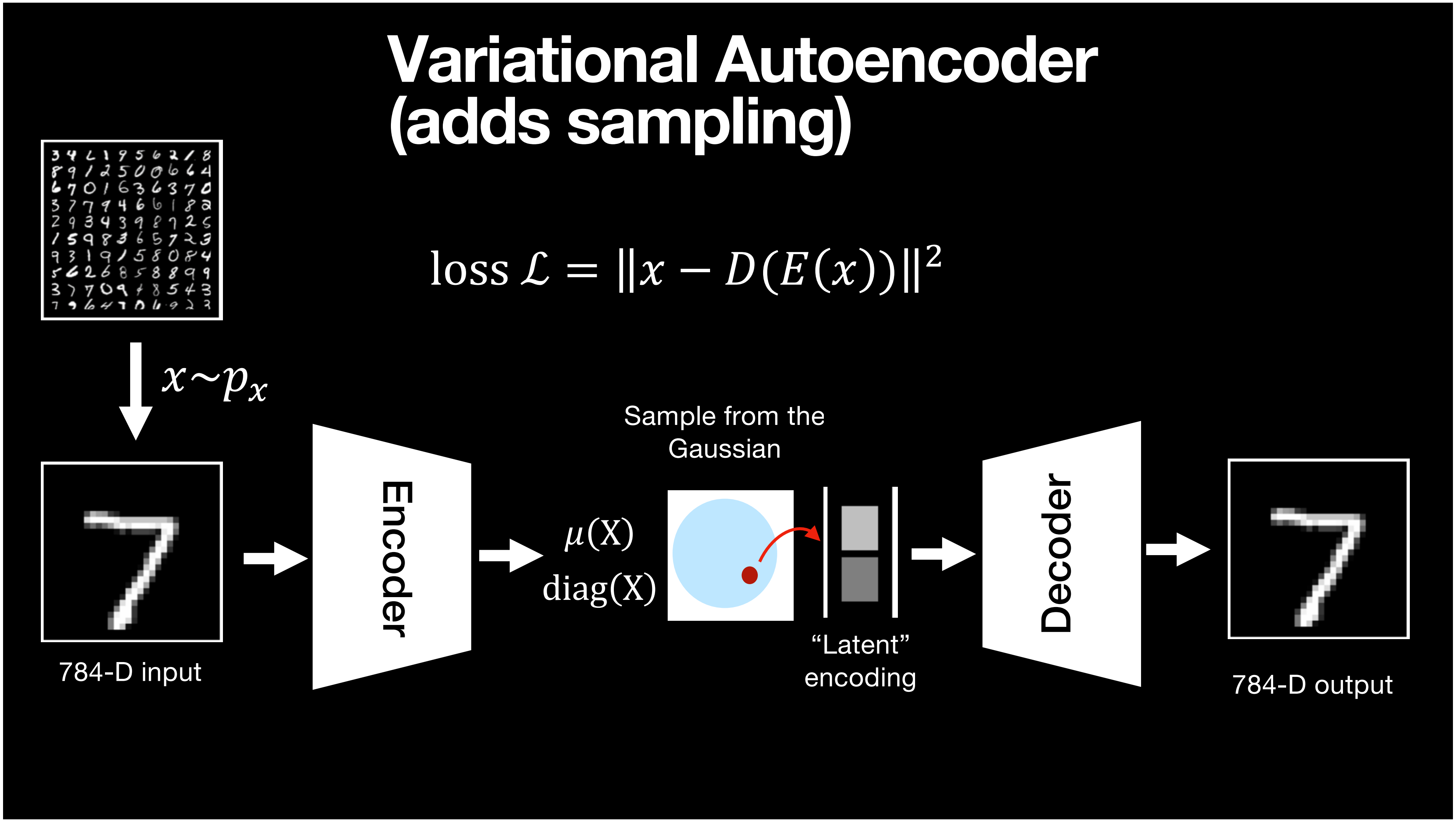
Maps d dimensional input x to k dimensional embedding subspace S

Autoencoding MNIST

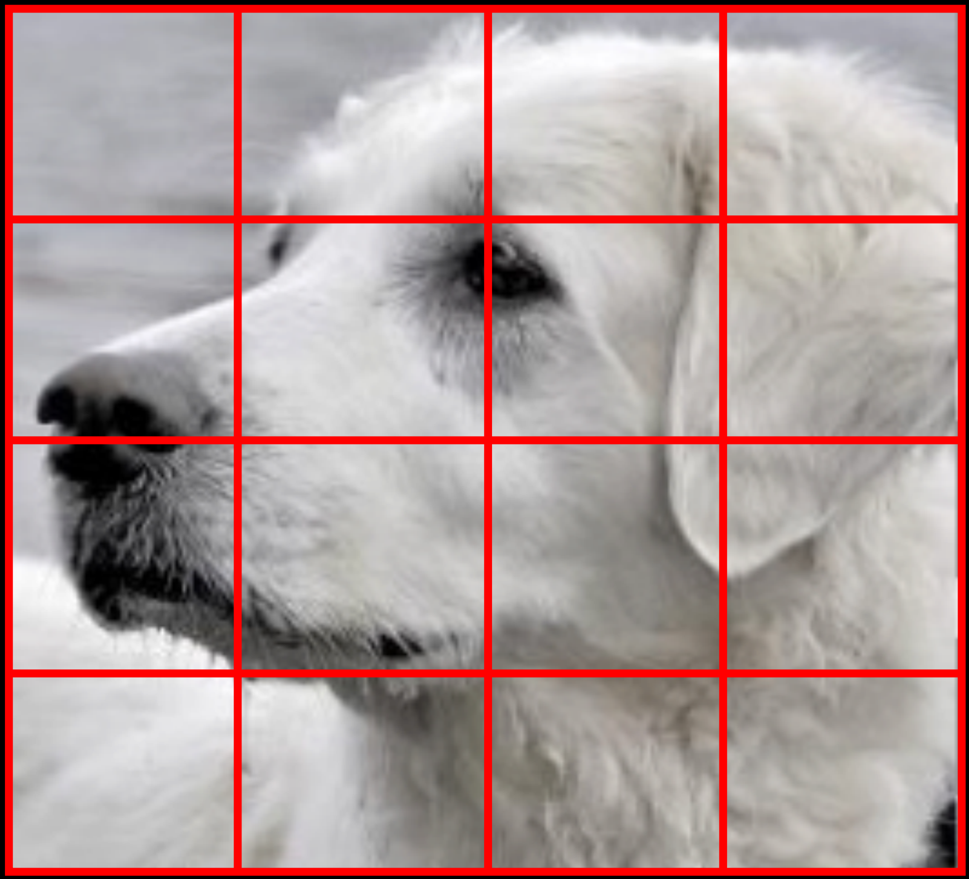
$$\text{loss } \mathcal{L} = \|x - D(E(x))\|^2$$




After training
loss $\mathcal{L} = \|x - D(E(x))\|^2$
$x \sim p_x$
Encoder
Decoder
784-D input
2-D ”latent” representation
784-D output


Variational Autoencoder
(adds sampling)
loss $\mathcal{L} = \|x - D(E(x))\|^2$
$x \sim p_x$
Encoder
$\mu(X)$
$\mathrm{diag}(X)$
Sample from the Gaussian
"Latent" encoding
Decoder
784-D input
784-D output

Encode big pictures patch by patch



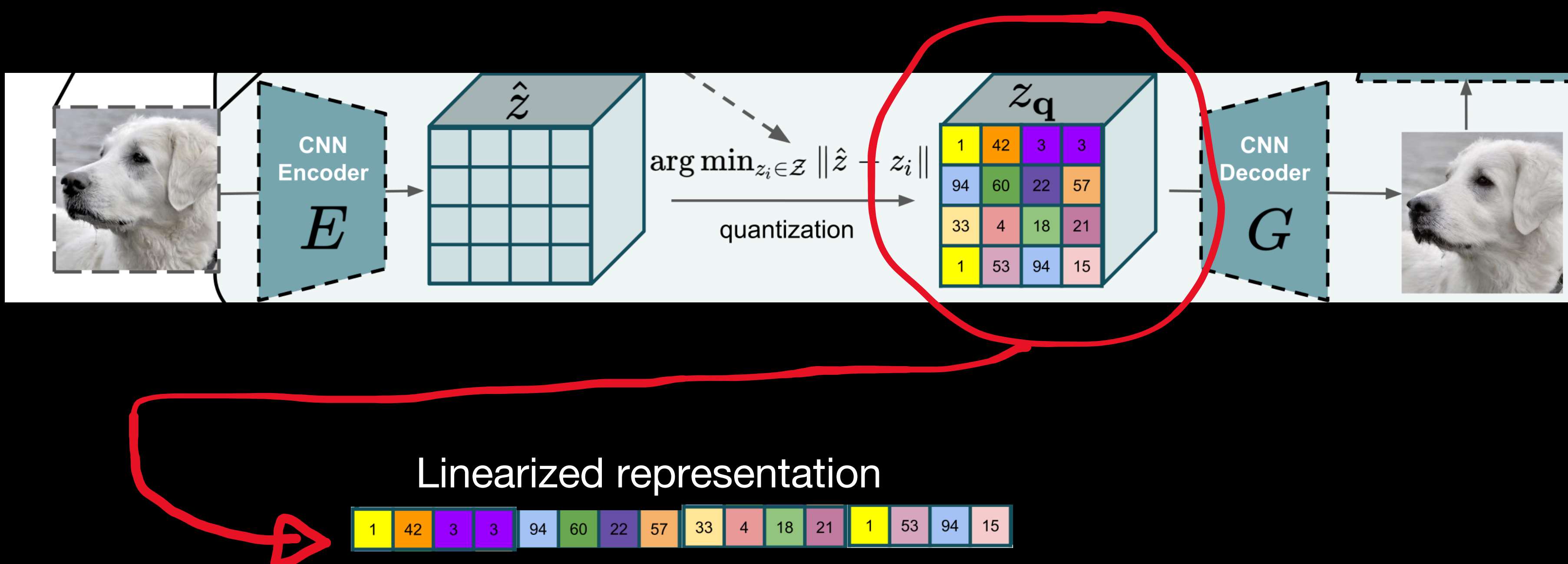
Can we use the latent representation as tokens?

No: They’re real valued

Transformers need a finite dictionary

Solution: Quantize

The whole quantization shebang



Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In CVPR, 2021

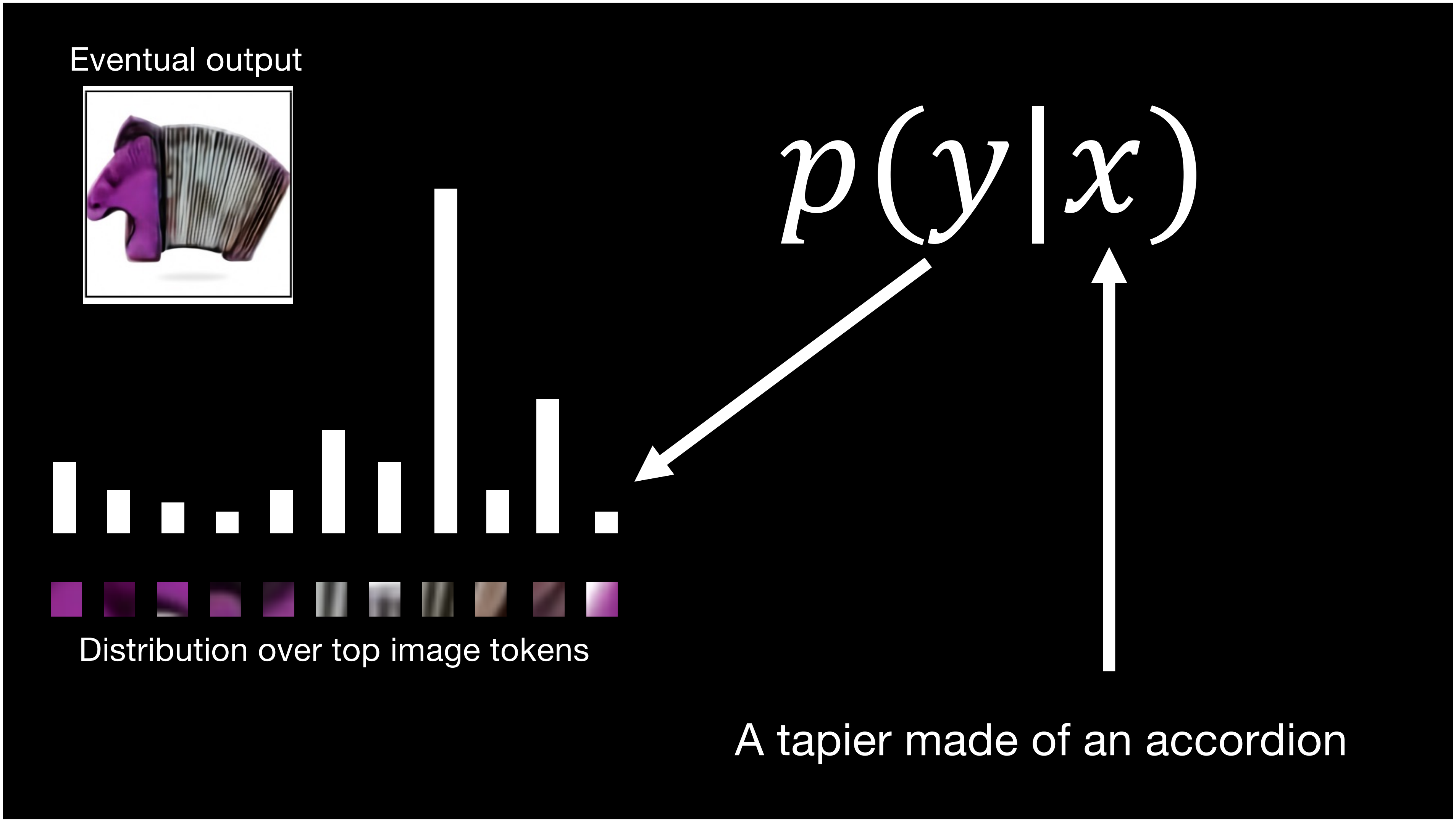
Now we’re ready…

Example: The original DALL-E

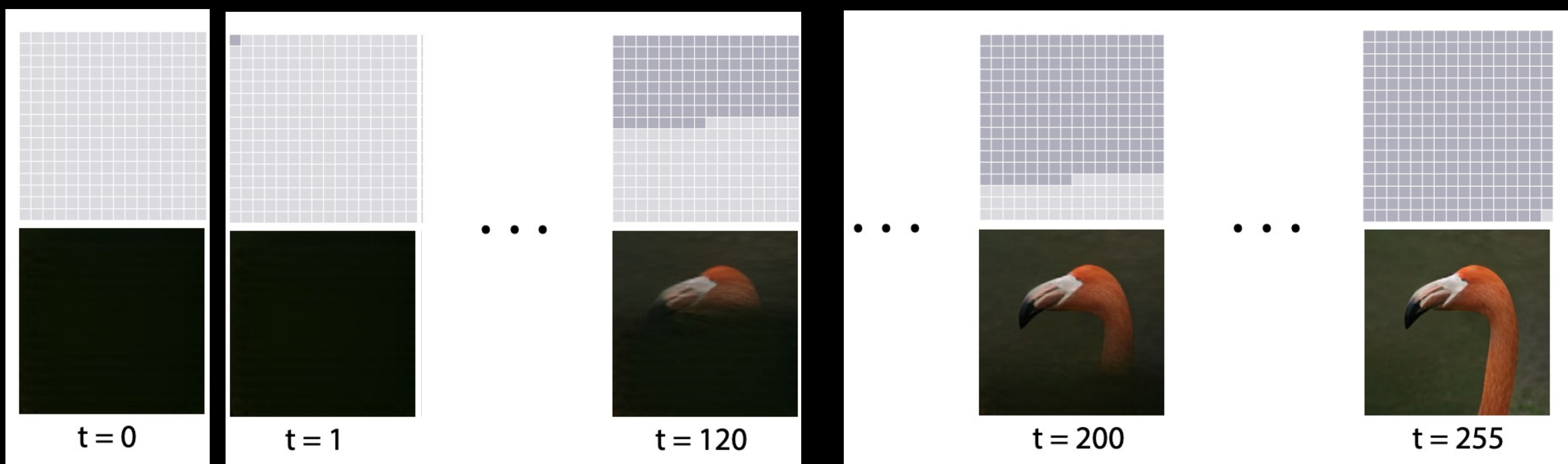
- Encodes a 256 by 256 image with a discrete Variational Auto Encoder (VAE)
- Each token from the VAE encodes a patch of pixels
- Image is now a 32 by 32 (ie 1024) token sequence

(a) a tapir made of accordion. a tapir with the texture of an accordion.

Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International Conference on Machine Learning. PMLR, 2021.

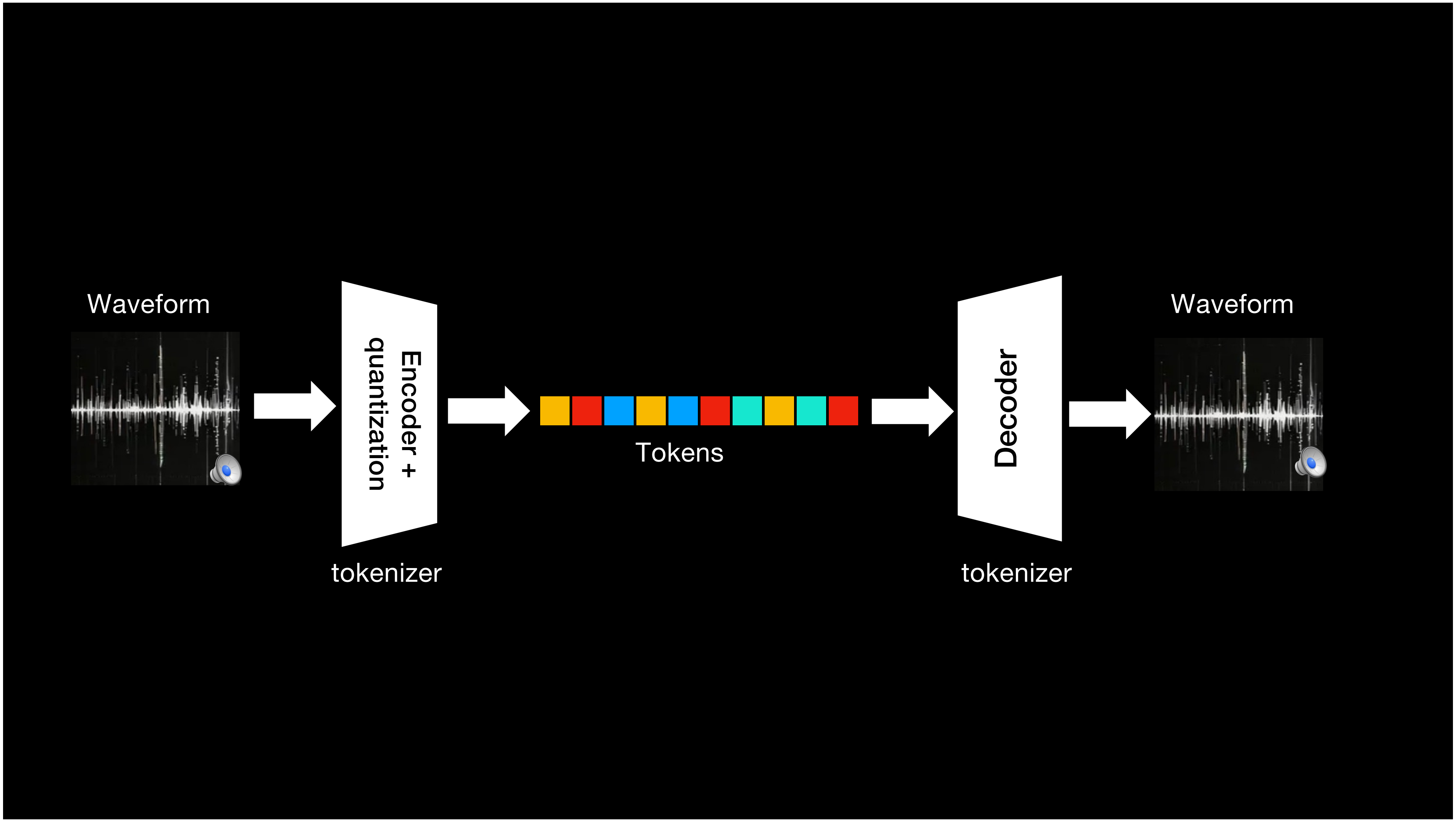

Eventual output
$p(y|x)$
Distribution over top image tokens
A tapier made of an accordion

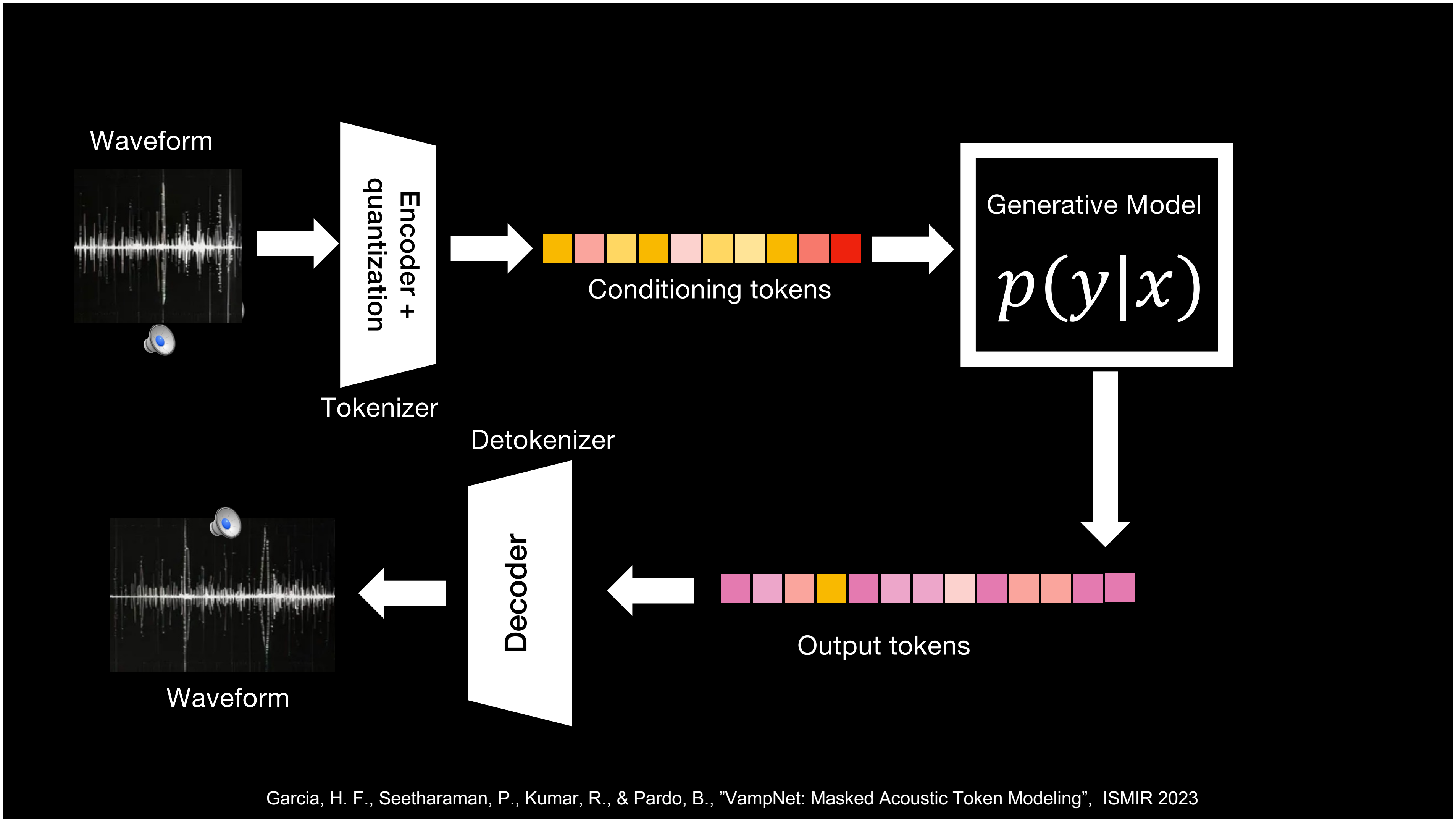
Making a flamingo



Chang, Huiwen, et al. "Maskgit: Masked generative image transformer." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

You can condition on any kind of data


Waveform
Encoder +
quantization
tokenizer
Tokens
Decoder
tokenizer
Waveform


Waveform
Encoder +
quantization
Tokenizer
Conditioning tokens
Generative Model
$p(y|x)$
Detokenizer
Decoder
Output tokens
Waveform
Garcia, H. F., Seetharaman, P., Kumar, R., & Pardo, B., "VampNet: Masked Acoustic Token Modeling", ISMIR 2023

The possibilities are endless

Video to sound FX

Video to music

Speech to music

Movies from still images

And so on….

https://imagen.research.google/video/

Our lab: The Rhythm In Anything (prompting just on rhythm)

But can we do away with quantizing?

(We’ll need latent diffusion for that)