ChatGPT: The Task

SAM SCOTT, MOHAWK COLLEGE, MAY 2023

The Core Task of ChatGPT

Given a text **prompt**, predict the natural language **token** (word) that comes next.

ChatGPT is a **Large Language Model** powered by a deep **Artificial Neural Network** architecture called a **Transformer**.

What is a Language Model?

It’s a probability distribution for sequences of tokens (words)

“it’s my day in the sun” vs “my it’s sun day the in”

p = 90% p = 0.1%

It can do conditional probabilities

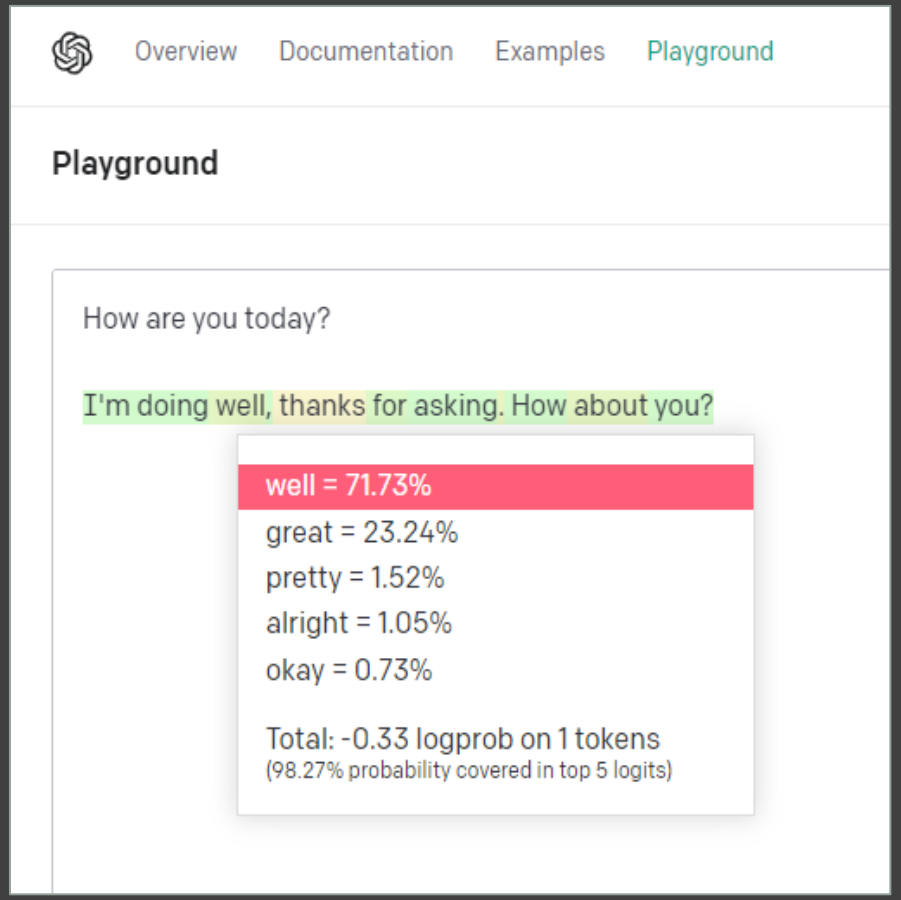
“it’s my day in the…” → “sun” p = 53%

“would you like to purchase the…” → “sun” p = 5%

ChatGPT: One Token at a Time

	PROMPT	OUTPUT
1	How are you today?	I
2	How are you today? I	'm
3	How are you today? I’m	doing
4	How are you today? I’m doing	well
5	How are you today? I’m doing well	,
6	How are you today? I’m doing well,	thanks
7	How are you today? I’m doing well, thanks	for
…	…	…

Try it Yourself!



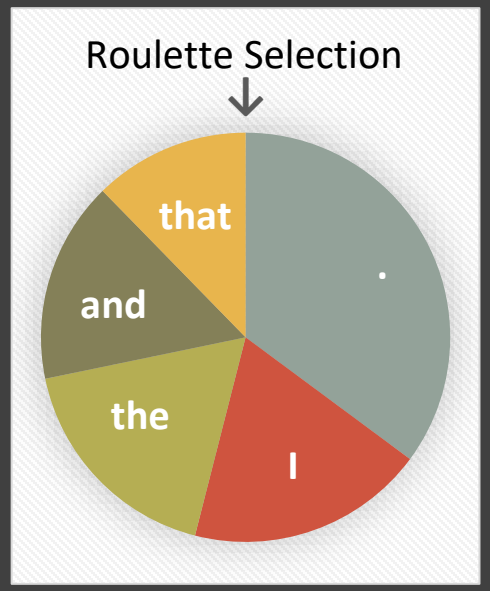
platform.openai.com/playground

- **Mode:** Complete
- **Show Probabilities:** Full Spectrum
- Enter a prompt
- Press **Submit**.
- Click a word…

A Tiny Language Model (TLM)

The TLM counts tokens (words) in a small dataset (62 MB of text)

Token	Count	Frequency (%)
.	135 319	6.7%
I	72 506	3.7%
the	68 432	3.4%
and	61 352	3.1%
that	47 409	2.4%



TLM 1: Roulette Selection, No Context

Those optimism I it doubt getting would we and? Are but you and so the. We have that cheering the yeah administrators find that thorn change there is even all're agree what.

TLM 2: One Word of Context

TLM 2 Counts bigrams (pairs of tokens)

bigram	Count	Frequency*
the same	1102	1.7%
the other	1031	1.6%
the time	854	1.3%
the way	811	1.2%
the people	709	1.1%

* percentage of all “the ___” bigrams

TLM 2: One Word Context

Prompt: I

Context: "I" …

have been to be an interesting. The one of the only the same time with them. They don't get up the time I'm going to be some of people.

TLM 2: One Word Context

Prompt: hello!

Context: ‘!’ …

The whole lot. The first year. The same problem. In the last two years ago the same thing. The people who had a lot of a good.

TLM 3: Two Word Context (trigrams)

Prompt: Tell me something interesting.

Context: 'interesting .’ …

Oh I know that's a big change. There is a little bit more. It has been the way it should be a little bit of money.

TLM 4: Three Word Context (4-grams)

Prompt: Tell me something interesting.

Context: 'something interesting .' ...

One possible solution is a quite radical one. Say if each household in britain was only allowed 1 car or each registered voter was allowed the volume would be immediately privileged in such a situation exhibits unfailing resolve and determination.

From Tiny to Large Language Models

LLMs are massively scaled up

- GPT-3
  - 570 GB (499 Billion tokens) of training text
  - $12 million to train
  - Predictions based on 2048-grams
- GPT-4
  - Training data and costs undisclosed
  - Predictions based on 8192-grams or 32768-grams

From Tiny to Large Language Models

The TLM approach does **not** scale up

- Storage
  - 28 000 tokens, 350 000 bigrams, 880 000 trigrams, 1.2 million 4-grams….
  - How many 2048-grams?
- Sparsity
  - The longer the sequence, the rarer it is.
  - “The big green” → 1 occurrence in the TLM corpus
    - “The big red” and “the big blue” aren’t there at all…

Large Language Models must **generalize**!

Next Video: Journey Through an LLM

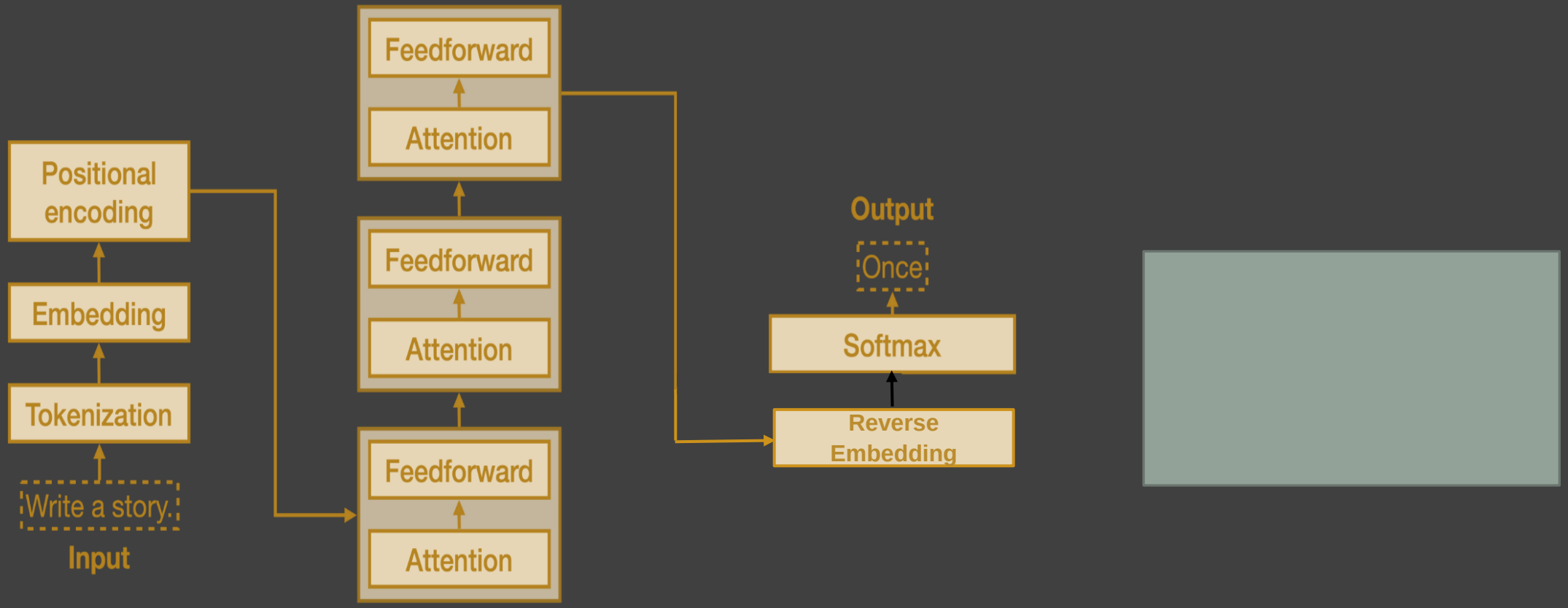


Diagram adapted from https://txt.cohere.com/what-are-transformer-models