Build Carolina LLM Workshop

Dr. Hampton Smith

Who am I?

- Doctorate, Clemson University, 2013
- Alphabet née Google, Data Center Software Division
- CTO, Bandwagon

Fun Fact

Fun Fact

- Working with a new personal trainer

Fun Fact

- Working with a new personal trainer
- He is the devil

Fun Fact

- Working with a new personal trainer
- He is the devil
- I am in tremendous pain

Am I an AI expert?

Am I an AI expert?
Nope!

Current AI Boom

- ChatGPT - November 30, 2022
- 19 months ago

Current AI Boom

- ChatGPT - November 30, 2022
- 19 months ago
- “Attention Is All You Need” Vaswani et al. @ Google Brain (2017)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-

Then how would you hire an “AI Expert”?

- Neural networks
- Machine learning
- “Classical” AI
- Linear Algebra
- GPUs
- But…
  - At this point just familiarity with the tools, platforms, and lingo will let you get a lot done

How to use this workshop

- Code along!
- Or don’t, I’m a slide deck not a cop
  - Slides, code, and lecture notes are available in the workshop’s GitHub repo

Goals

- Learn how to stand up and integrate with various LLM models
- Acquire a toolkit of libraries and services for integrating with LLMs
- Gain confidence with general AI terminology and techniques
- Explore the primary LLM programming model: prompt engineering
- Establish starting points for a variety of future exploration, including training, fine-tuning, and retrieval-augmented generation

Less talk more rock

gradio

Inference API

🤗 Spaces
(or Heroku, AWS, GKE, whatever)

gradio

🤗 Inference API

ChatGPT

🤗 Spaces
(or Heroku, AWS, GKE, whatever)

gradio

🤗 Spaces
(or Heroku, AWS, GKE, whatever)

🤗 Inference API

ChatGPT

Claude

gradio

🤗 Spaces
(or Heroku, AWS, GKE, whatever)

🤗 Inference API

ChatGPT

Claude

gradio

LangChain

🤗 Spaces
(or Heroku, AWS, GKE, whatever)

🤗 Inference API

ChatGPT

Claude

gradio

88k stars
666 watching
13.8k forks

LangChain

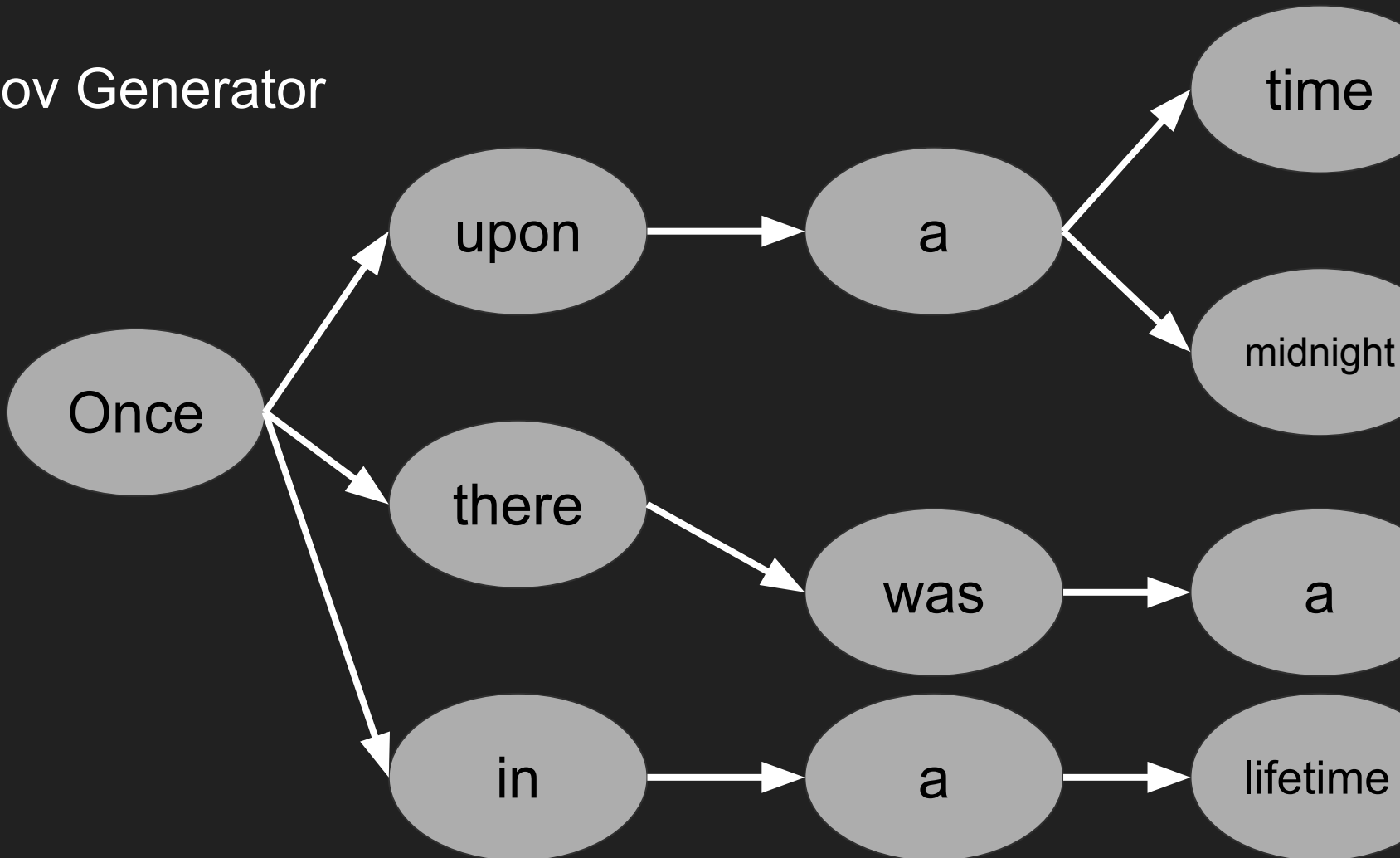
🤗 Spaces
(or Heroku, AWS, GKE, whatever)

🤗 Inference API

ChatGPT

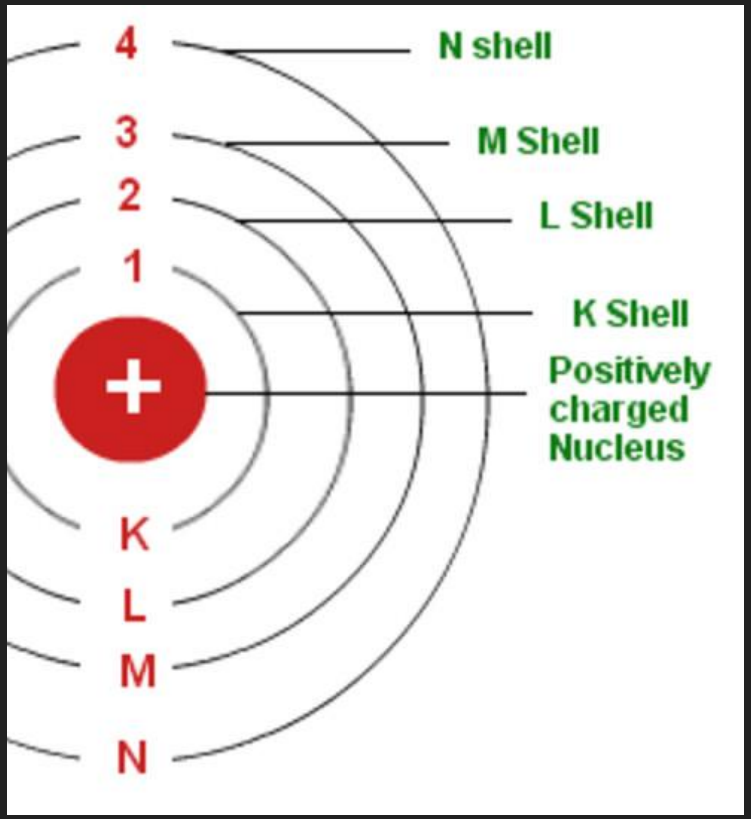
Claude

Markov Generator



Less. Talk. More. Rock.

That’s neat, I guess?



That’s neat, I guess?

Step 1

Analyze word frequency and Connections

Step 2

???

Step 3

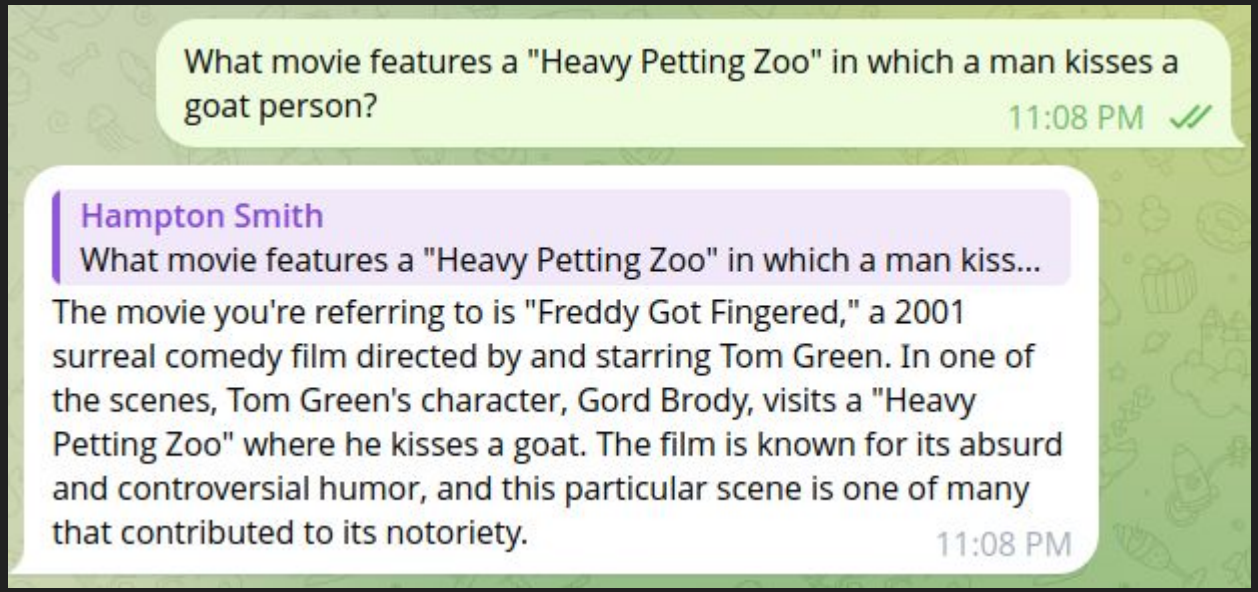
Profit! …by getting a lists of words by probability

So now we can understand…

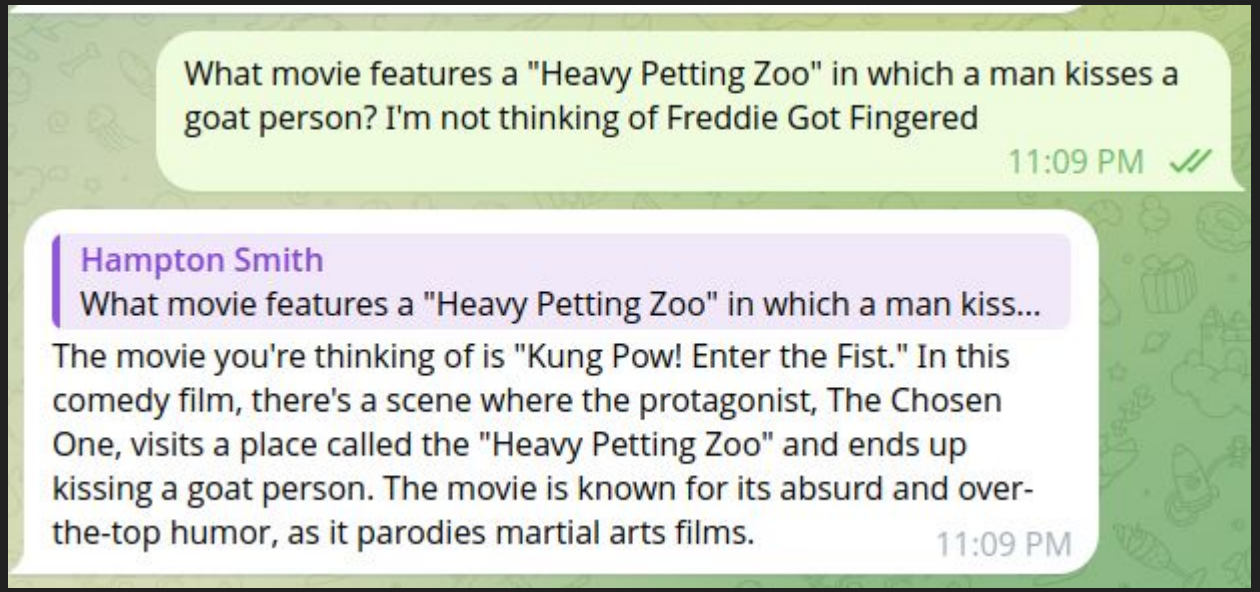
- top_p (and its cousin top_k)
- temperature
- …and we can also understand why LLMs *hallucinate*

An unhinged example

An unhinged example



An unhinged example

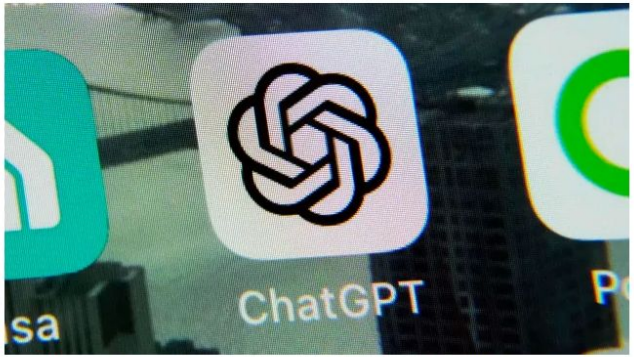


A less (more?) unhinged example

Lawyers blame ChatGPT for tricking them into citing bogus case law

BY LARRY NEUMEISTER, ASSOCIATED PRESS - 06/08/23 11:25 PM ET

SHARE POST



FILE – The ChatGPT app is displayed on an iPhone in New York, May 18, 2023.A judge is deciding whether to sanction two lawyers who blamed ChatGPT for tricking them into including fictitious legal research in a court filing. The lawyers apologized at a hearing Thursday, June 8, 2023, in Manhattan federal court for their roles in written submissions that seemed to leave Judge P. Kevin Castel both baffled and disturbed at what happened. (AP Photo/Richard Drew, File)

NEW YORK (AP) — Two apologetic lawyers responding to an angry judge in Manhattan federal court blamed ChatGPT Thursday for tricking them into including fictitious legal research in a court filing.

Attorneys Steven A. Schwartz and Peter LoDuca are facing possible punishment over a filing in a lawsuit against an airline that included references to past court cases that Schwartz thought were real, but were

We've also *already learned* these vocab words:

- Parameter
- Model
- Token
- Attention

Additional vocab

- AI - Distinguish “classical” vs “transformer-” or “attention-” based
- LLM - Neural network
  + huge input corpus (e.g. all of wikipedia)
  + (probably) attention
- Transformer - an architecture for implementing attention-based AI well suited to implementation on GPUs

What Changed?

1) The transformer architecture
2) The upfront work is now done

Questions?

Homework*

1) What is the effect of extreme settings for top_p and temperature?
2) Can you get your chatbot to hallucinate?
3) How do adding phrases like “Answer step-by-step,” or “Speculate wildly” affect the chat bot’s answers?
4) Our chatbot lacks context. How would we set this up to answer “What is the capital of Germany?” and then “How many people live there?” in a way that it would understand what “*there”* means?

* totally optional, the world is awful, go be with your families