

# 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  
lukaszkaizer@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-to-German translation task, improving over the existing best results, including

# 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

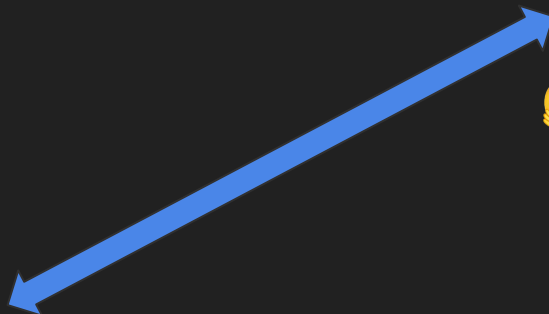


gradio



Spaces

(or Heroku, AWS, GKE, whatever)



Inference API



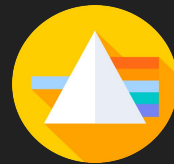
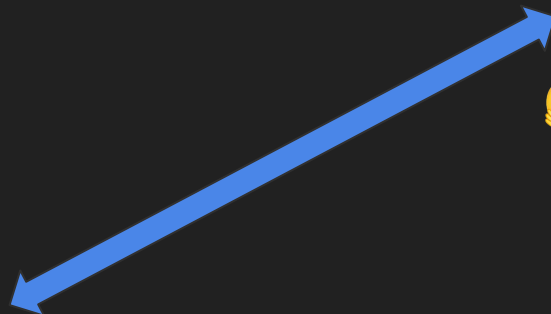


gradio



Spaces

(or Heroku, AWS, GKE, whatever)



Inference API



ChatGPT

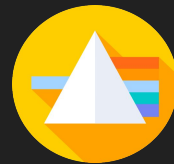
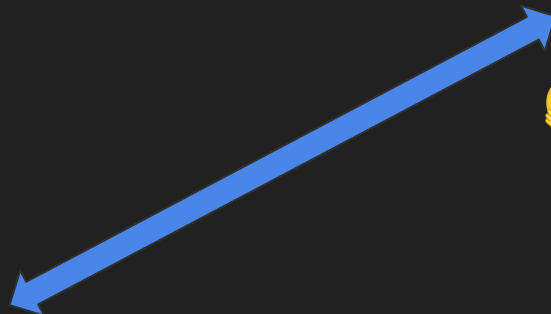


gradio



Spaces

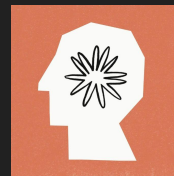
(or Heroku, AWS, GKE, whatever)



Inference API



ChatGPT



Claude

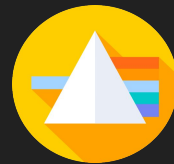
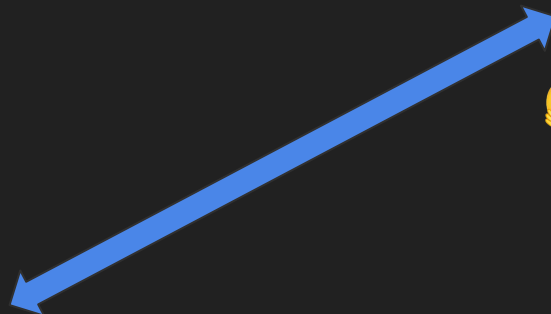


gradio



Spaces

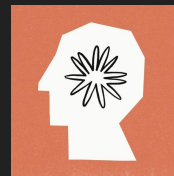
(or Heroku, AWS, GKE, whatever)



Inference API



ChatGPT



Claude





gradio

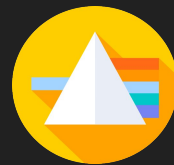


LangChain

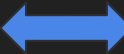


Spaces

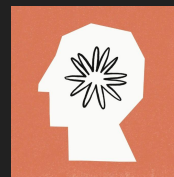
(or Heroku, AWS, GKE, whatever)



Inference API



ChatGPT



Claude



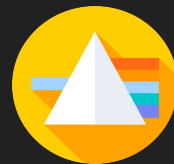


☆ 88k stars  
👁 666 watching  
🔗 13.8k forks

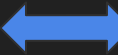


Spaces

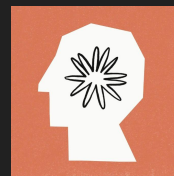
(or Heroku, AWS, GKE, whatever)



Inference API



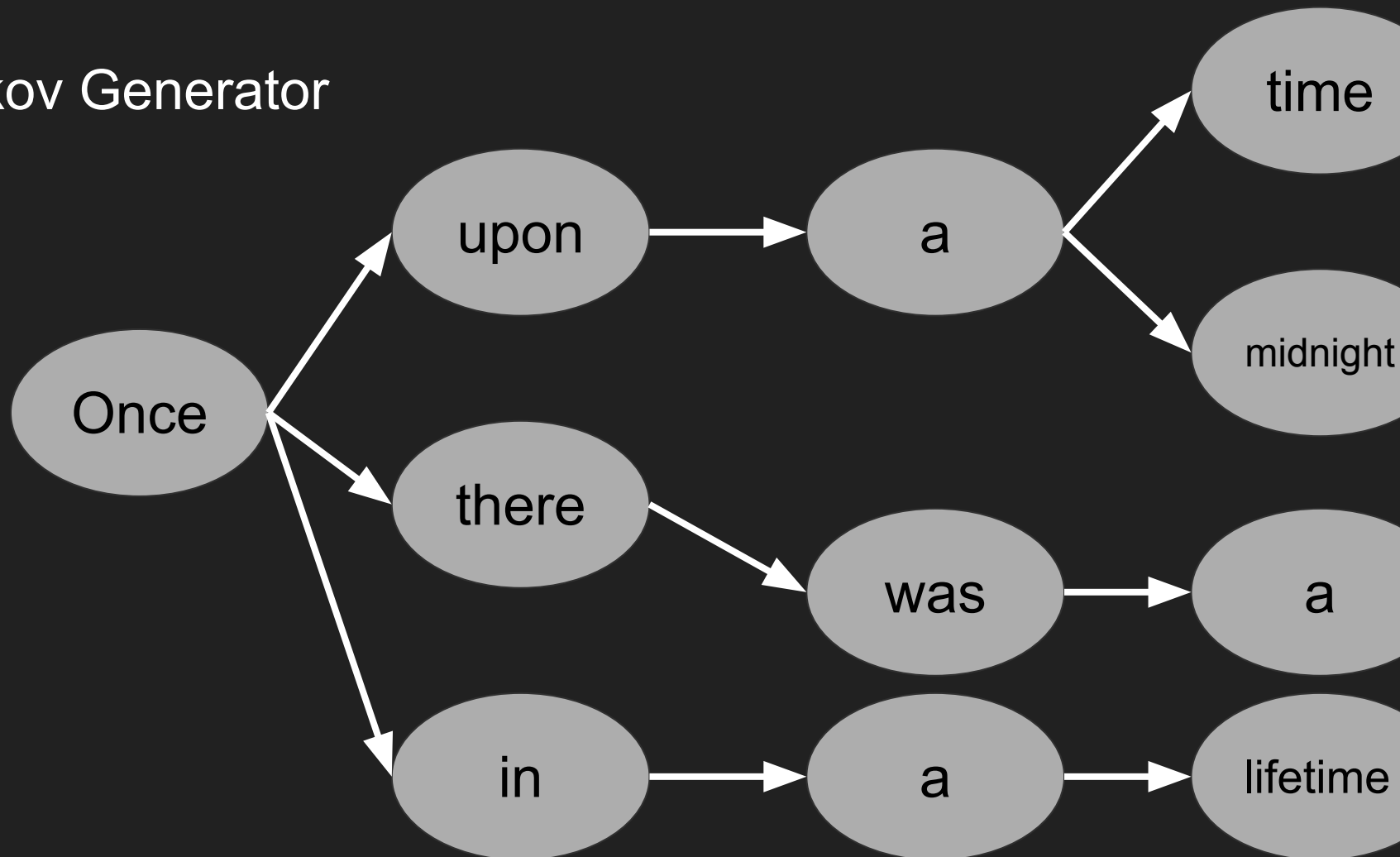
ChatGPT



Claude

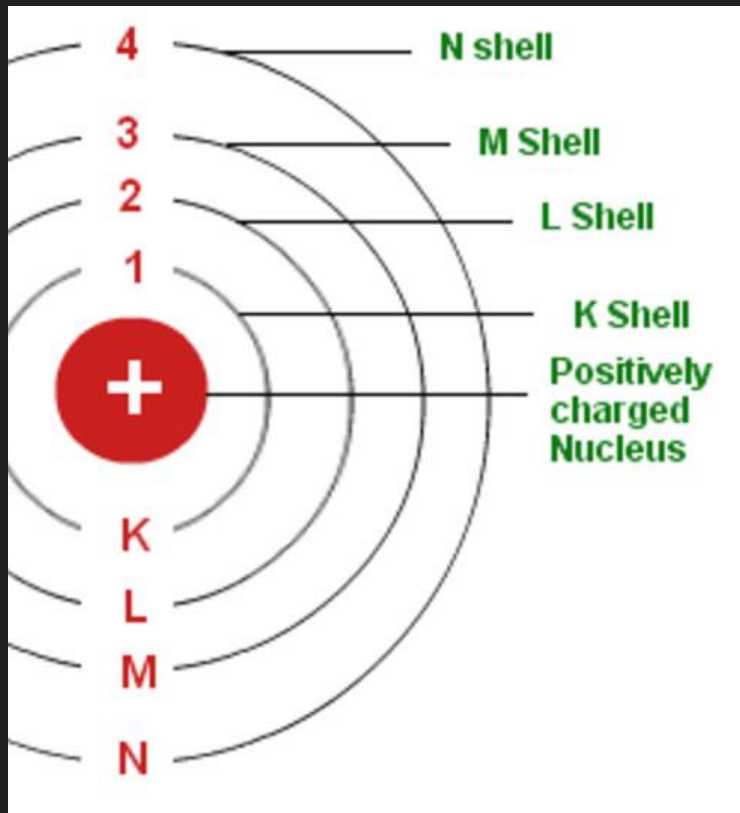


# Markov Generator



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

# Parameters

	Once	Upon	A	Time	Darcy	House	...
Once	.005	.1	.05	.01	.06	.01	
Upon	.1	.005	.1	.05	.1	.01	
A	.05	.001	.001	.1	.02	.05	
Time	.01	.005	.005	.01	.05	.02	
Darcy	.05	.001	.001	.1	.02	.05	
House	.05	.05	.05	.01	.01	.001	
...							

# 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

What movie features a "Heavy Petting Zoo" in which a man kisses a goat person?

11:08 PM ✓✓

**Hampton Smith**

What movie features a "Heavy Petting Zoo" in which a man kiss...

The movie you're referring to is "Freddy Got Fingered," a 2001 surreal comedy film directed by and starring Tom Green. In one of the scenes, Tom Green's character, Gord Brody, visits a "Heavy Petting Zoo" where he kisses a goat. The film is known for its absurd and controversial humor, and this particular scene is one of many that contributed to its notoriety.

11:08 PM

# An unhinged example

What movie features a "Heavy Petting Zoo" in which a man kisses a goat person? I'm not thinking of Freddie Got Fingered

11:09 PM ✓✓

**Hampton Smith**

What movie features a "Heavy Petting Zoo" in which a man kiss...

The movie you're thinking of is "Kung Pow! Enter the Fist." In this comedy film, there's a scene where the protagonist, The Chosen One, visits a place called the "Heavy Petting Zoo" and ends up kissing a goat person. The movie is known for its absurd and over-the-top humor, as it parodies martial arts films.

11:09 PM

# 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

f SHARE X 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

VIEW

1. Click "View"
2. Install Firefox Ex
3. Enjoy EasyView

EasyV

Most Popular

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

How do we add history?

How do we “fake” a chat?

# Prompt Engineering



Syntax, grammar, semantics,  
logic

Model

Tasks,  
Domains

Fine tuning

Prompt

Talk like a pirate?

Answer in rhyme?

Only wrong answers?

Refuse to answer certain questions?

Pretend to be a character?

Reason about a made-up domain?

# This goes beyond silly stuff...

- Research suggests that planting the prompt, “Let’s reason step-by-step...” or “Answer carefully...” can improve answer quality
- We can plant useful information in the prompt about the user, the session, the account, or our service
- Reduce hallucination by instructing AI not to make stuff up
- We can demand a certain type of output

# Prompt Engineering has limitations

- It's always piggybacking on the model and the fine-tuning (it can't "learn new things")
- Context!



# Retrieval Augmented Generation (RAG)

```
const vectorStore = await HNSWLib.fromTexts(  
  ["mitochondria is the powerhouse of the cell"],  
  ...  
);  
  
const prompt =  
  PromptTemplate.fromTemplate(  
    `Answer the question based only on the following context:  
    {context}  
  
    Question: {question}`  
  );  
  
const chain = ...; // Combine vector with prompt to fill in {context}  
  
const result = await chain.invoke("What is the powerhouse of the cell?");
```

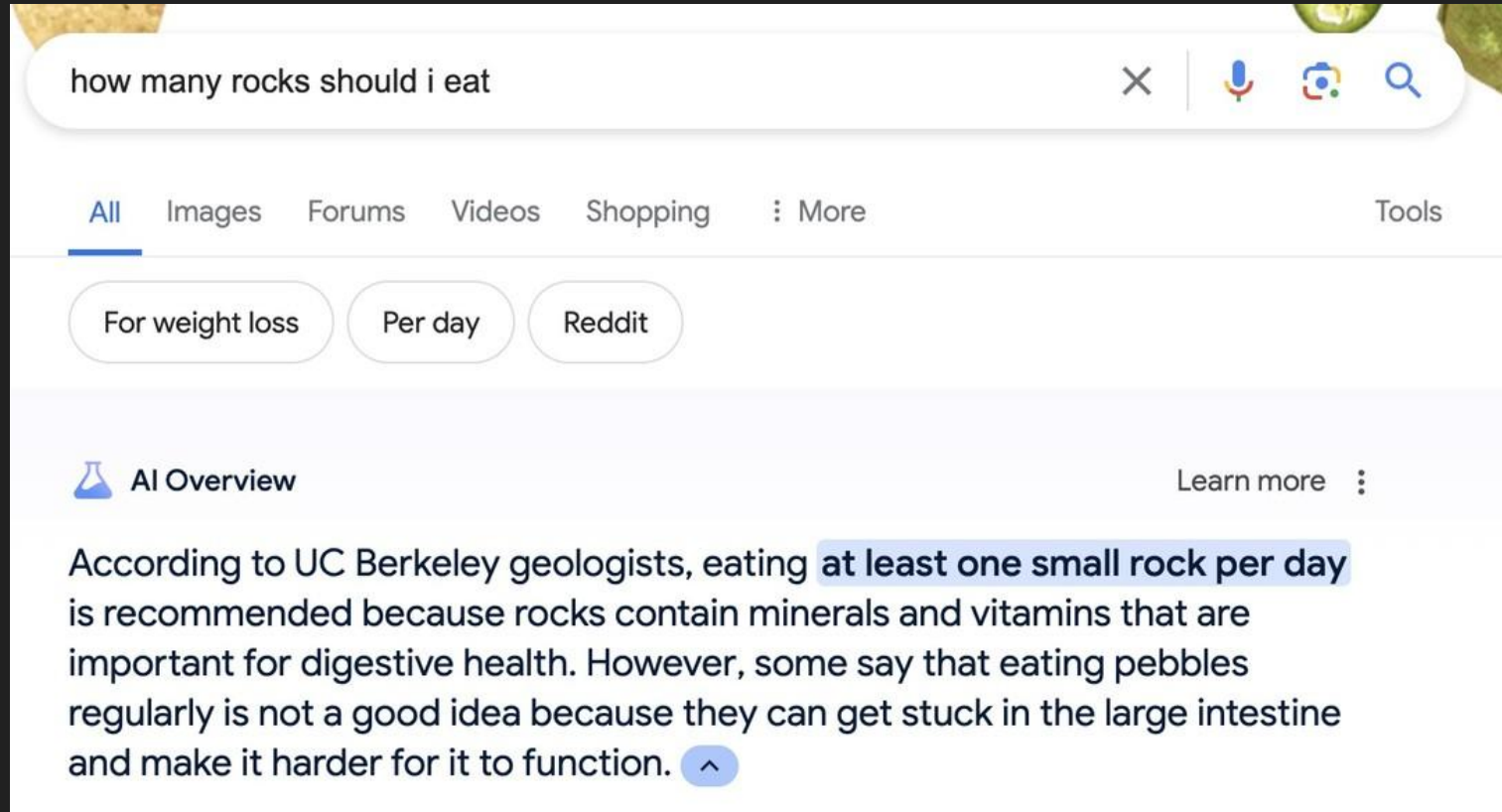
# Agents

```
api_wrapper = WikipediaAPIWrapper(  
    tool = WikipediaQueryRun(api_wrapper=api_wrapper)  
)  
agent = create_tool_calling_agent(llm, [tool], prompt)  
agent_executor = AgentExecutor(agent=agent, tools=tools, verbose=True)  
agent_executor.invoke("Give me facts about Greenville, SC!")
```

# Fine Tuning

# Professional Responsibility & Ethics


# Responsibility to disclose




how many rocks should i eat

All Images Forums Videos Shopping More Tools

For weight loss Per day Reddit

 AI Overview [Learn more](#)

According to UC Berkeley geologists, eating **at least one small rock per day** is recommended because rocks contain minerals and vitamins that are important for digestive health. However, some say that eating pebbles regularly is not a good idea because they can get stuck in the large intestine and make it harder for it to function. 

# Training Bias



Valentin Hofmann

@vjhofmann

...

🌟 New paper 🌟

We discover a form of covert racism in LLMs that is triggered by dialect features alone, with massive harms for affected groups.

For example, GPT-4 is more likely to suggest that defendants be sentenced to death when they speak African American English.



Dialect prejudice predicts AI decisions about people's character, employability, and criminality

Valentin Hofmann<sup>1-3\*</sup>, Pratyusha Ria Kalluri<sup>4</sup>, Dan Jurafsky<sup>4</sup>, Sharese King<sup>5\*</sup>

<sup>1</sup>Allen Institute for AI   <sup>2</sup>University of Oxford   <sup>3</sup>LMU Munich

<sup>4</sup>Stanford University   <sup>5</sup>The University of Chicago

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