How LLMs Work

Jacob Matthews

Good artists copy, great artists _____

Insanity is doing the same thing over and over again and expecting different _____

I just want you to know that, when we talk about war, we're really talking about _____

My favorite _____ is orange.

Your favorite color is _____

```
def mean(a: list[float]) -> float:\n\ttotal = 0\n\tfor num
in a:\n\t\ttotal _______
```

A language model is _____

A language model is _____

a. “a type of artificial intelligence that learns patterns in language so it can understand, generate, and respond to text” (ChatGPT)
b. “a type of AI system trained to understand and generate human language” (Claude)
c. “a model of the human brain’s ability to produce natural language” (Wikipedia)

A language model is _____

a. “a type of artificial intelligence that learns patterns in language so it can understand, generate, and respond to text” (ChatGPT)
b. “a type of AI system trained to understand and generate human language” (Claude)
c. “a model of the human brain’s ability to produce natural language” (Wikipedia)
d. a probability distribution over sequences of symbols

Einstein (GPT-2)

Insanity is doing the same thing over and over again and expecting different _____

“results”: 90.2%

“things”: 2.6%

“outcomes”: 2.2%

….

London is the capital of _____ (GPT-2)

1. London
2. p(“is” | “London”) = 1.2%
3. p(“the” | “London is”) = 5.6%
4. p(“capital” | “London is the”) = 1.4%
5. p(“of” | “London is the capital”) = 79.0%
    a. p(“the” | “London is the capital of”) = 29.0%
    b. p(“a” | “London is the capital of”) = 7.8%
    c. p(“London” | “London is the capital of”) = 3.4%
    d. p(“Britain” | “London is the capital of”) = 3.0%
    e. p(“England” | “London is the capital of”) = 2.8%

- Language model != chatbot
- The term “LLM” now encompasses many different modeling approaches, technologies, and use cases.
- Regardless of how you interact with them, LLMs are fundamentally about probability distributions over sequences.

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2024

Today Date: 21 January 2026

You are a helpful assistant<|eot_id|><|start_header_id|>user<|end_header_id|>

Can you make a seaborn heatmap for a similarity matrix where you only keep the upper triangular and put the y tick labels neatly along the diagonal?
<|eot_id|><|start_header_id|>assistant<|end_header_id|> _____

What this class is not about

- NLP in general
  - Our focus is more narrow and less formal. There's more to NLP than LLMs!
- LLM deployment, optimization, industry applications, application development
  - We work with small models and only briefly touch on some optimization approaches
- SOTA, niche, or emerging architectures/techniques
  - You'll be more prepared to approach these topics after you take this course, though

What is this course about?

Syllabus overview