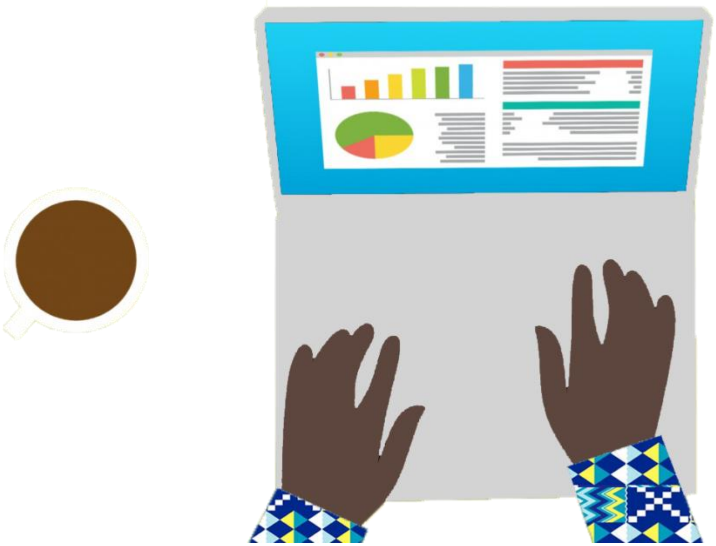
What the History of LLMs Teaches Us

The Graph Courses



Why learn the history of LLMs?

Forecasting requires understanding the past

Pitfalls can be better avoided

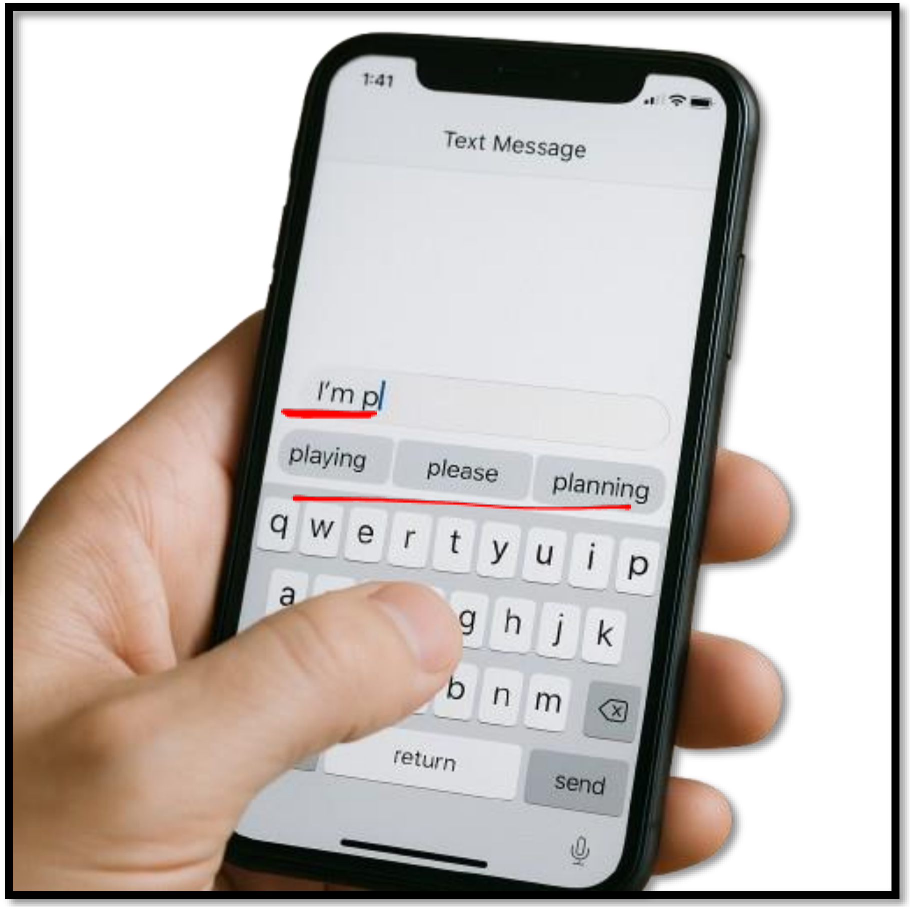
Question we'll use throughout

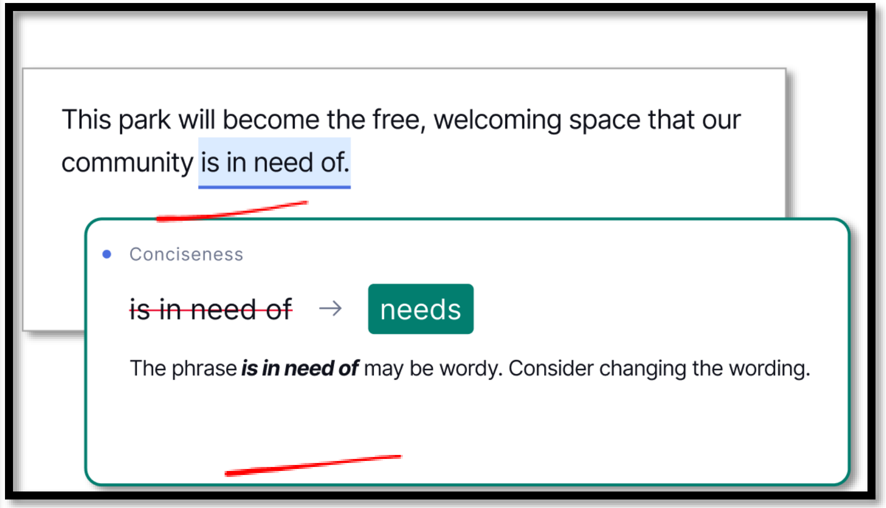
"Can you tell me who the current pope is"

What is a Language Model?

A computer system that **learns patterns** from text data to **predict** and generate **language**.

e.g. Google Translate, texting autocomplete, grammar checkers, LLMs





A Very Simple Language Model: Fourgram

- A fourgram estimates the probability of a word based on the three words that come right before it. Bigrams, trigrams etc. also common
- N-grams used in search boxes, autocomplete.

Toy example of fourgram model:

Training Corpus:

```
how to bake bread
weather today london uk
what is the weather like today
what is the capital of spain
what is the capital of france
best coffee shops nearby
what is the best way to learn French
```

Language Model Output:

what is the	
what is the capital	50%
what is the weather	25%
what is the best	25%

Model sees "what is the," and checks how this phrase was completed in the corpus.
"Capital" came next 2 times, "weather" 1 time, and "best" 1 time. So it thinks "capital" is more likely to follow.

Live demo: fivegram trained on children's storybooks

```python

# Load the merged fairy tales data and build the model
with open("merged_fairy_tales.txt", "r", encoding="utf-8") as f:
    text = f.read()

# Build the model
words = preprocess_text(text)
model = build_model(words)

# Example usage:
generate_text(model, "Once upon a time")
# Each time the generation is slightly different
# Because the model samples among the common next words

# So you can write your own fairy tale about
# But the stories quickly loose coherence
generate_text(model, "The old man said")

# And the model only pays attention to the last 4 words
generate_text(model, "While falling into the volcano, the prince and princess")

# Another limitation: the input text MUST appear somewhere in the original data
# Otherwise there is no probability to sample from
generate_text(model, "Nigeria, the country situated")
⌘L to chat, ⌘K to generate
```

Summary of demo

- n-gram uses **naïve training/inference** (just counting; no understanding)
- Extremely **limited context**
- But still **proof of concept**: **statistical learning** from text can produce (partly) **coherent predictions**

Check your understanding

Question: An n-gram model, like the fourgram example, primarily makes predictions based on:

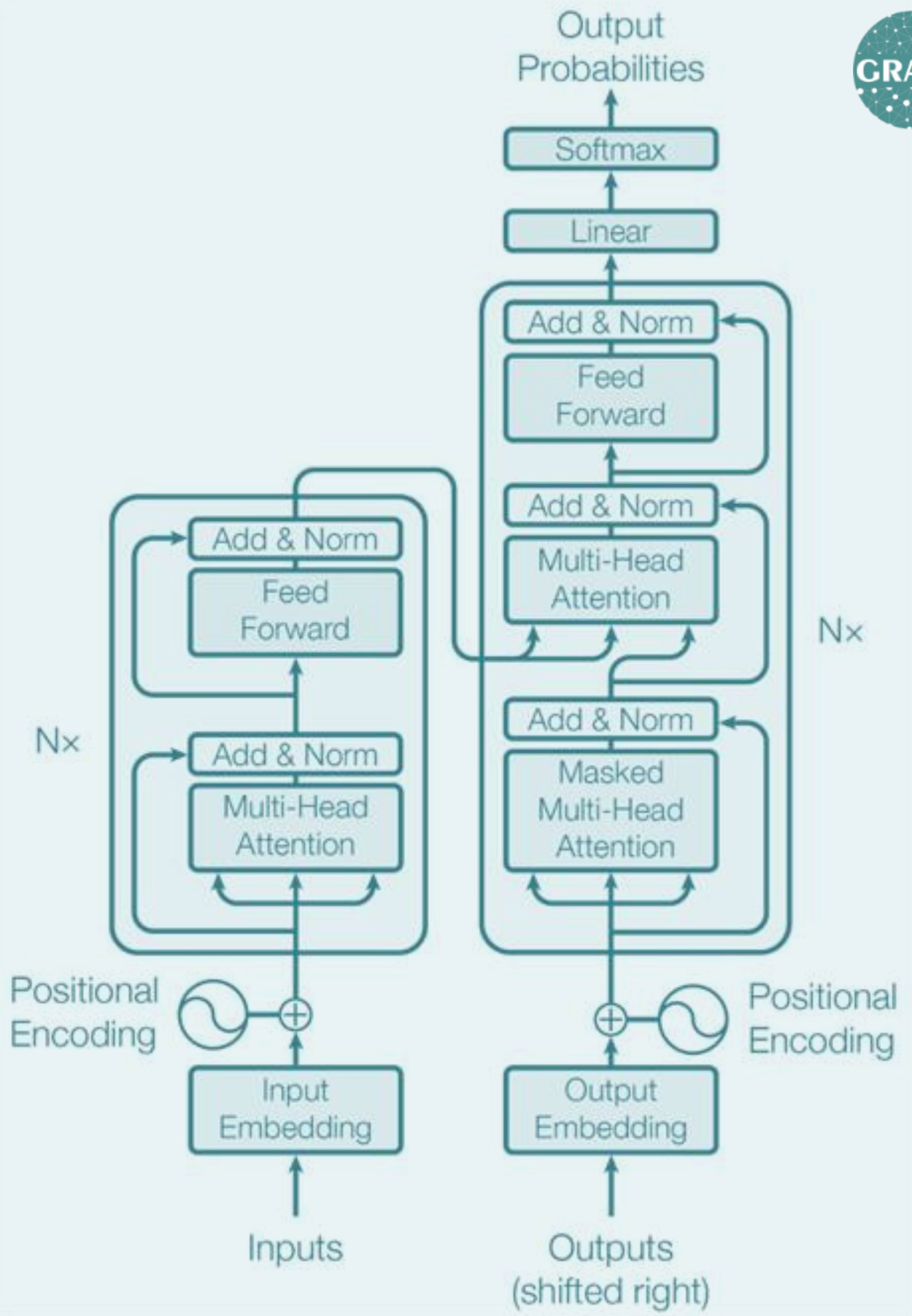
A. The grammatical structure of a sentence.

B. The statistical likelihood of word sequences observed in training data.

C. A deep understanding of the meaning or intent behind the words.

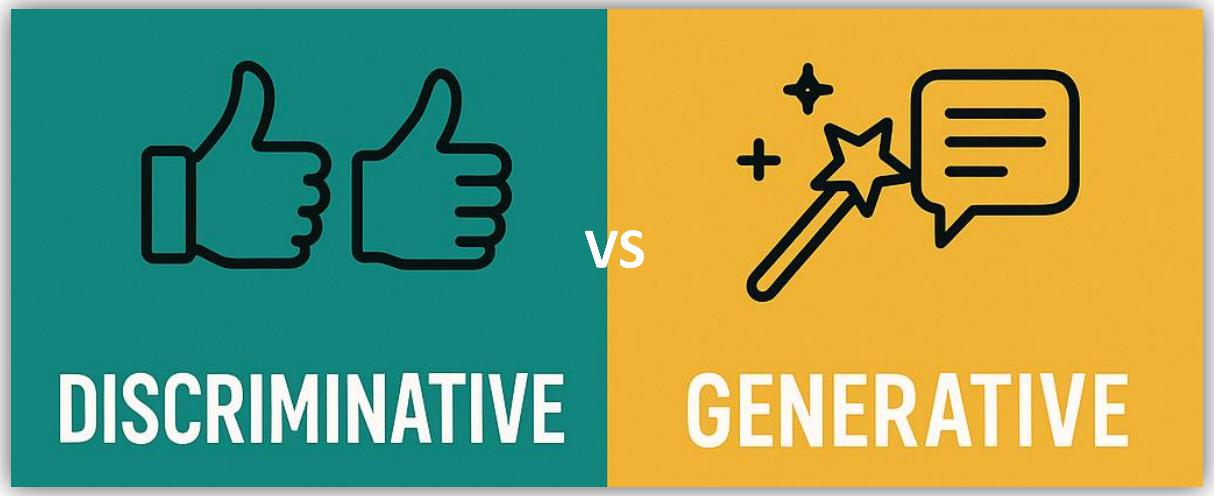
The rise of the GPTs

Generative
Pretrained
Transformers



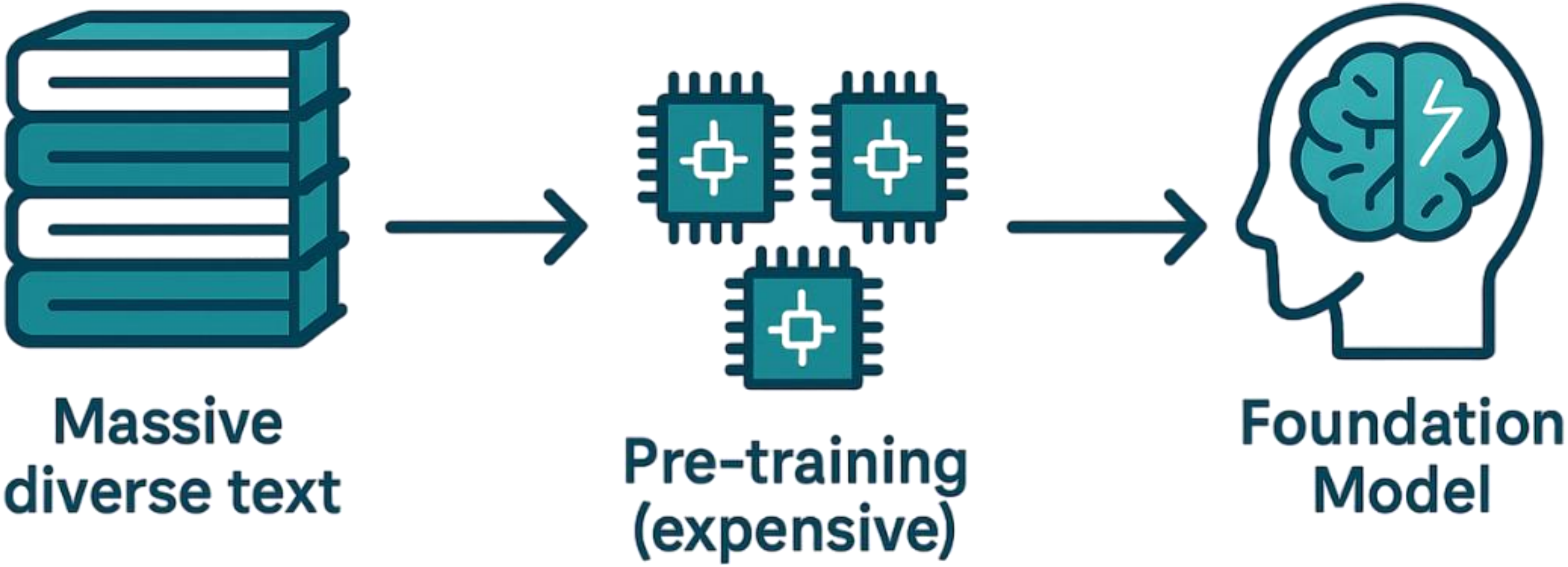
“Generative”?

- In the past, most language models were used for classification e.g. spam detection, sentiment analysis
- Generative models **generate** outputs based on prompts



“Pre-trained”?

- Our n-gram model (and many ML models) are small/simple enough to be trained quickly on new datasets or tasks.
- But GPTs need to be first pre-trained on **massive** datasets using lots of power and computation
- This creates a **foundation model** with important knowledge (grammar, facts, reasoning patterns)



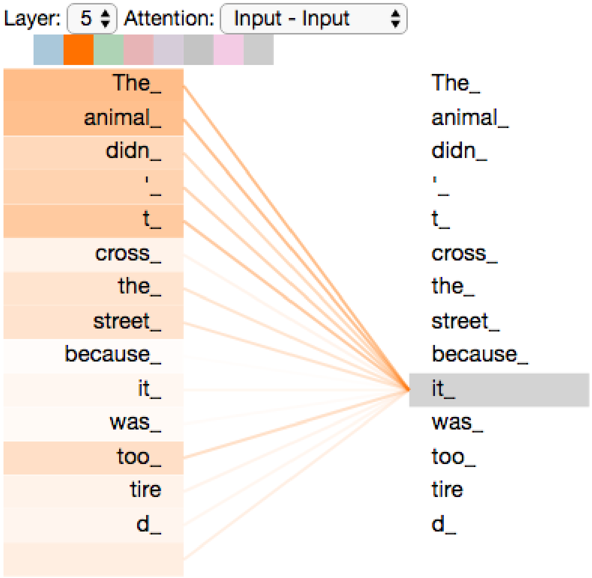
"Transformer"?

A special type of neural network (machine learning technique) that is great for language modeling.

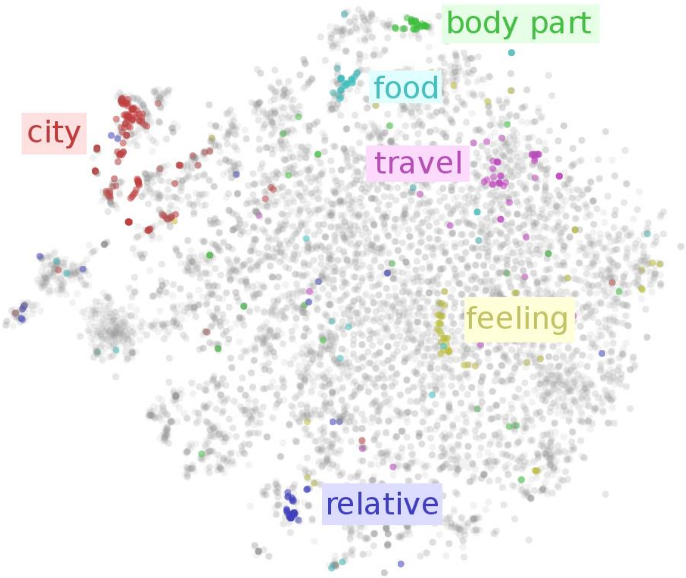
Technique published in 2017 by a Google team.

1. Sees the bigger picture

- Much longer context
- Decides which words matter for prediction (smart highlighter)



2. Rich internal representation of word meanings



3. Trains really well on modern GPUs

Check your understanding

In contrast to earlier classification models, “generative” LLMs can…

A. Assign a label (e.g. spam/not spam) to an input

B. Produce new text based on a prompt

What does “pre-trained” mean in GPT?

A. The model is manually trained with grammar rules

B. It’s first trained on massive unlabeled text, then fine-tuned on tasks

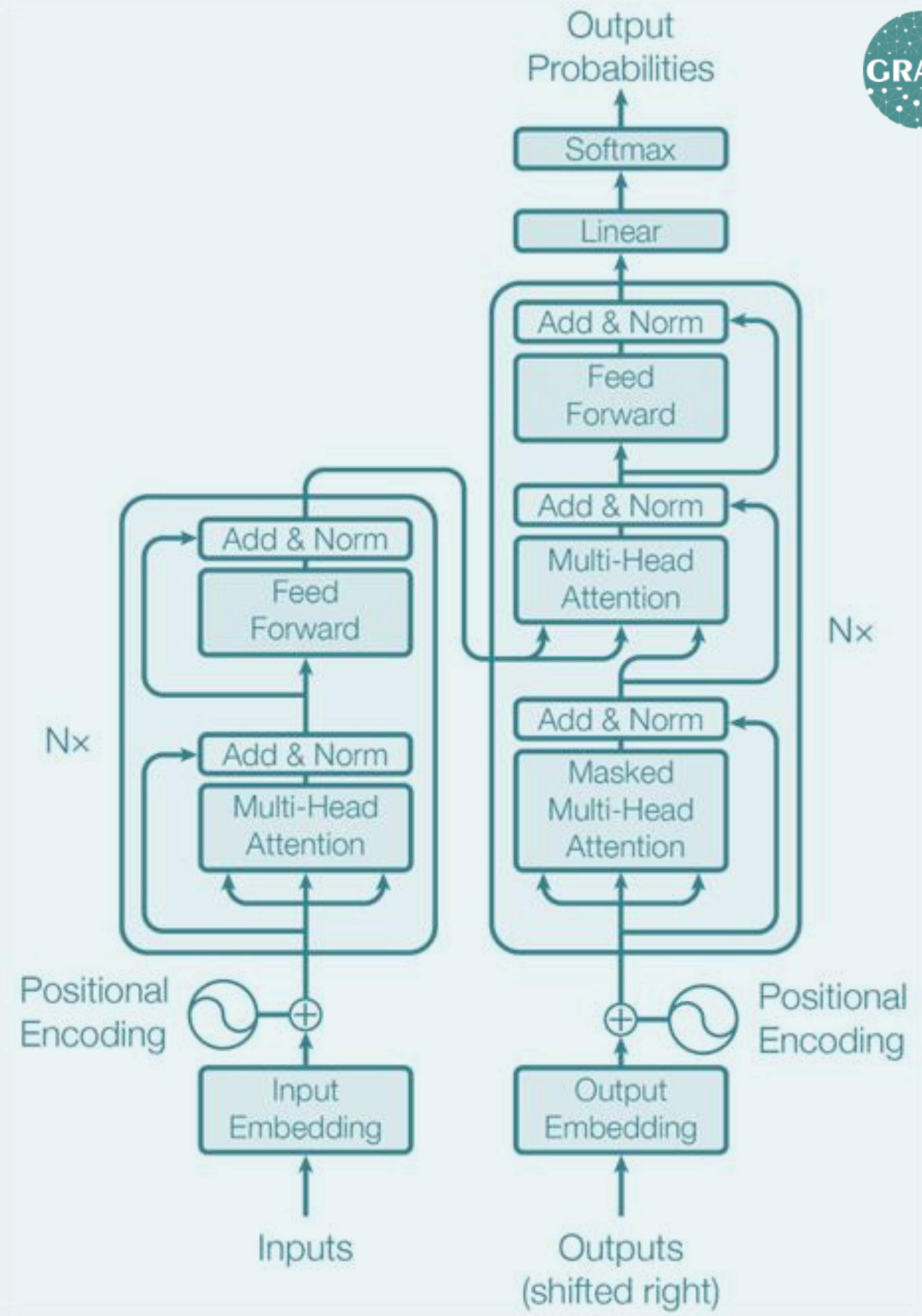
C. Comes pre-loaded with a fixed set of responses to common questions.

What kind of data does a language model use during its next-token pre-training phase?

A. Large amounts of raw, unlabeled text from the web

B. Human-annotated examples of question–answer pairs

The Evolution of the GPTs



GPT-N series: Evolution of Models

Model (year)	Model Innovation	"Can you tell me who the current pope is"	Training Corpus Size (Estimate)
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**	**"Because I'm dying for that information"** Autocomplete of typical trained text. Ignores the question's intent.	~1 B → 300 B tokens (5 GB → 45 TB raw)
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF	**"Pope Francis is the current pope"** Model follows instruction but is stale.	GPT-3 base 300 B + ≲ 1 B RLHF tokens
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt	**"The Church is currently without a pope"** Knows how to use Google to get up-to-date information.	~ 10–13 T tokens (undisclosed)
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation	*(THINKING)…quietly reasons, runs a quick search…* **"The Church is currently without a pope"** More capable with using Google and other tools to arrive at answers.	~ 10-20 T pre-training tokens + large-scale RL & tool-use fine-tuning (exact size undisclosed)

Next-token prediction. How a language model is pre-trained

Language model
(millions of parameters/weights
encoding mathematical functions)

Training data:
Billions of sequences with known, masked completions

Jack and Jill went up the _____

Each word is represented by high-dimensional vector (series of numbers)

Model "guesses", and learns from mistake

Bok: 3%
Tap: 2%
Cat: 1%
…
Hill: 0.05%

Learning algorithm (gradient descent) calculates how far the guessed answers are from the correct answer and nudges the weight "dials" in an intelligent way improve probability of the correct answer

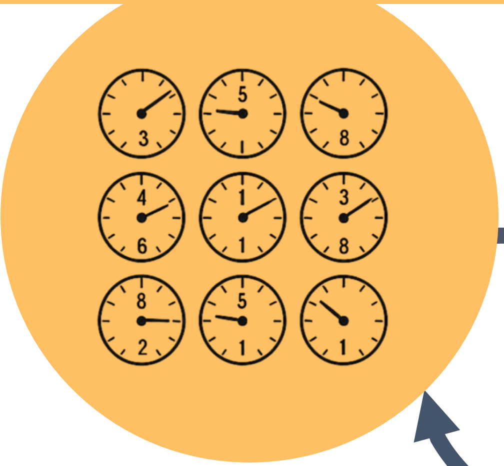
Next-token prediction. How a language model is pre-trained

Training data:
Billions of sequences with known, masked completions

Jack and Jill went up the _____

Each word is represented by high-dimensional vector (series of numbers)

Language model
(millions of parameters/weights encoding mathematical functions)



Model “guesses”, and learns from mistake

Hill: 14%
Mountain: 11%
Cat: 9%
…

Learning algorithm (gradient descent) calculates how far the guessed answers are from the correct answer and nudges the weight “dials” in an intelligent way improve probability of the correct answer

Data GPT-3 trained on

570 GB of text

Wikipedia
The Free Encyclopedia

Common Crawl

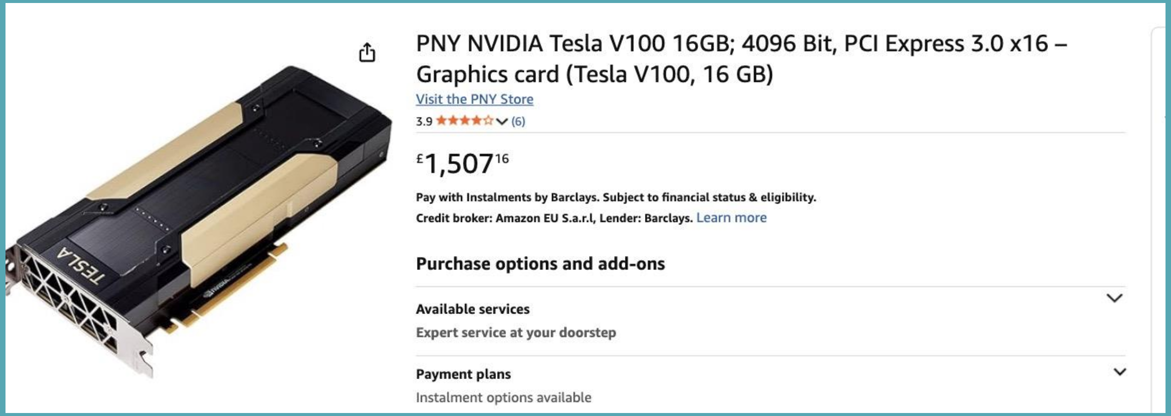
GitHub

etc.

= ~100,000 bibles worth of text

Compute GPT-3 trained with

~ 1000 **NVIDIA v100 GPUs** running non-stop for ~1 month



Check your understanding

What is the core training objective in next-token prediction?

A. Maximize the probability of each next token given its preceding context

B. Classify each input sequence into predefined categories

GPT-N series: Evolution of Models

Model (year)	Model Innovation	"Can you tell me who the current pope is"	Training Corpus Size (Estimate)
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**	**"Because I'm dying for that information"** Autocomplete of typical trained text. Ignores the question's intent.	~1 B → 300 B tokens (5 GB → 45 TB raw)
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF	**"Pope Francis is the current pope"** Model follows instruction but is stale.	GPT-3 base 300 B + ≲ 1 B RLHF tokens
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt	**"The Church is currently without a pope"** Knows how to use Google to get up-to-date information.	~ 10–13 T tokens (undisclosed)
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation	*(THINKING)…quietly reasons, runs a quick search…* **"The Church is currently without a pope"** More capable with using Google and other tools to arrive at answers.	~ 10-20 T pre-training tokens + large-scale RL & tool-use fine-tuning (exact size undisclosed)

Live demo: Next-token predictors are not that helpful

```
# Example prompts
# Great storyteller!
talk_to_gpt2("Once upon a time,")

# But not so great at answering questions
talk_to_gpt2("Can you tell me who the current pope is?")

# Few-shot example
few_shot_prefix = """
QUESTIONS & ANSWERS

Question: What is the capital of France?
Answer: Paris

Question: Can you tell me who wrote Romeo and Juliet?
Answer: William Shakespeare

Question:"""

talk_to_gpt2(few_shot_prefix + "Can you tell me who the current pope is?")
```

Summary of Demo

- **Basic Text Completion** - GPT-2 generates creative continuations to prompts but not great at question answering.
- **Example Prompting**- Few-shot prompts improved GPT-2's ability to follow instructions and answer questions.
- **Structure Control** - Adding special tokens like "<EOM>" helps control output

Check your understanding

Why might a purely pretrained language model (like GPT-2) struggle with direct questions?

A. It’s trained to continue text, not to give accurate answers

B. It lacks enough memory for simple factual questions

C. It cannot recognize language patterns at all

GPT-N series: Evolution of Models

Model (year)	Model Innovation	"Can you tell me who the current pope is"	Training Corpus Size (Estimate)
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**	**"Because I'm dying for that information"** Autocomplete of typical trained text. Ignores the question's intent.	~1 B → 300 B tokens (5 GB → 45 TB raw)
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF	**"Pope Francis is the current pope"** Model follows instruction but is stale.	GPT-3 base 300 B + ≲ 1 B RLHF tokens
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt	**"The Church is currently without a pope"** Knows how to use Google to get up-to-date information.	~ 10–13 T tokens (undisclosed)
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation	*(THINKING)…quietly reasons, runs a quick search…* **"The Church is currently without a pope"** More capable with using Google and other tools to arrive at answers.	~ 10-20 T pre-training tokens + large-scale RL & tool-use fine-tuning (exact size undisclosed)

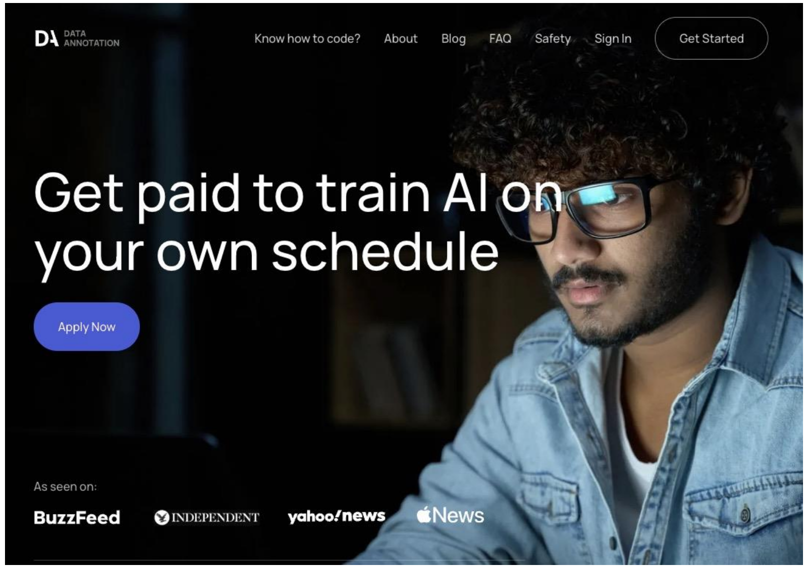
Instruction fine-tuning: From autocomplete to helpful assistant

Training the base model on demonstrations of desired assistant behaviors using human-written instructions and responses.

Example training data

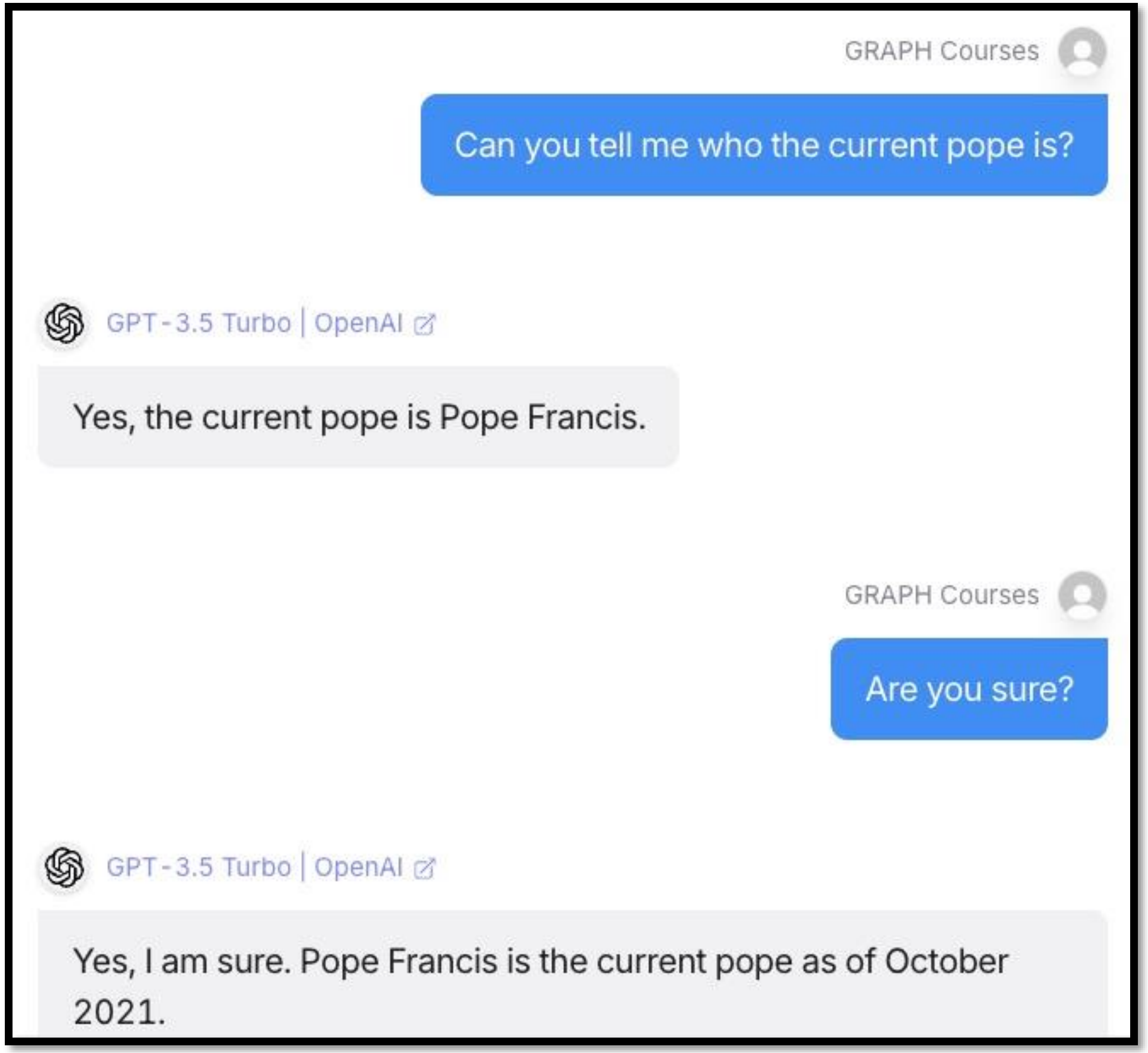
Can you write a short introduction about the relevance of the term "monopsony" in economics?…	prompter
"Monopsony" refers to a market structure where there is only one buyer for a particular good or…	assistant
How can one fight back when a monospony had been created?	prompter
Monopsony refers to a market structure where there is only one buyer of a good or service. In the…	assistant

Quite expensive. Need lots of humans



https://huggingface.co/datasets/OpenAssistant/oasst1/viewer/default/train?views%5B%5D=train

Live demo: Using ChatGPT 3.5



Summary of Demo

- Instruction-tuning has turned a powerful autocomplete into a helpful assistant.
- But training data is stale (stuck at the point when the data for pre-training was scraped)

Check your understanding

The key innovation that distinguished ChatGPT-3.5 from earlier GPT models (like GPT-3) was its significantly improved capability for:

A) Accessing and incorporating real-time information from the internet.

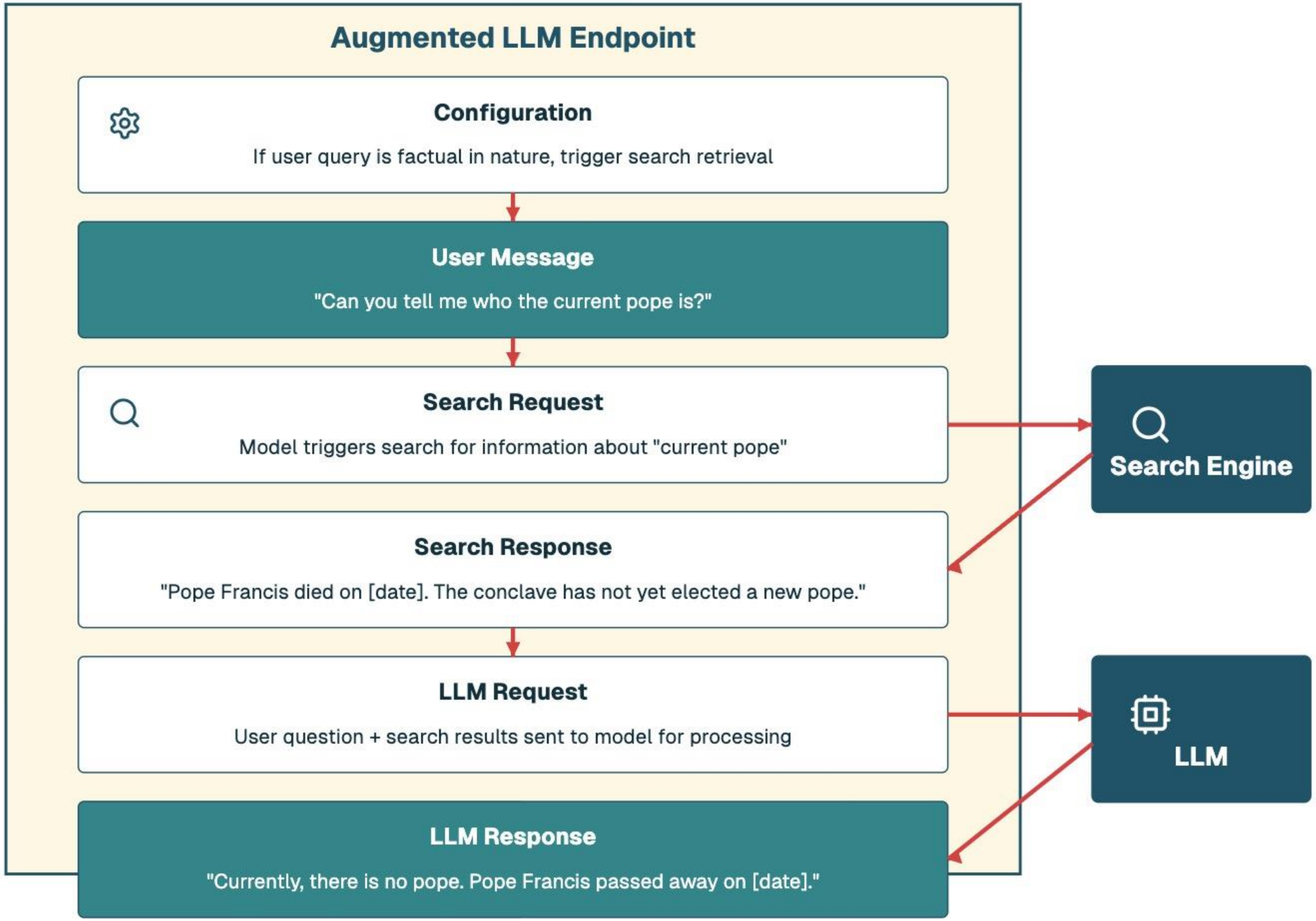
B) Generating much longer and more varied styles of text.

C) Understanding and following instructions to act as an assistant.

GPT-N series: Evolution of Models

Model (year)	Model Innovation	"Can you tell me who the current pope is"	Training Corpus Size (Estimate)
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**	**"Because I'm dying for that information"** Autocomplete of typical trained text. Ignores the question's intent.	~1 B → 300 B tokens (5 GB → 45 TB raw)
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF	**"Pope Francis is the current pope"** Model follows instruction but is stale.	GPT-3 base 300 B + ≲ 1 B RLHF tokens
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt	**"The Church is currently without a pope"** Knows how to use Google to get up-to-date information.	~ 10–13 T tokens (undisclosed)
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation	*(THINKING)…quietly reasons, runs a quick search…* **"The Church is currently without a pope"** More capable with using Google and other tools to arrive at answers.	~ 10-20 T pre-training tokens + large-scale RL & tool-use fine-tuning (exact size undisclosed)

Tool use (fine-tune/system prompt): Teaching model to use tools



Live demo: Using ChatGPT-4

Can you tell me who the current pope is?

W Searching the web

As of May 7, 2025, the papacy is currently vacant following the death of Pope Francis on April 21, 2025, at the age of 88 . The College of Cardinals convened in the Sistine Chapel today for the first round of voting to elect the next pope. However, the appearance of black smoke from the chapel chimney signaled that no candidate achieved the required two-thirds majority vote . Vatican News +1 Wikipedia +5 AP News +3

The conclave will continue with additional voting sessions in the coming days

Check your understanding

“Tool-use” fine-tuning (e.g. in GPT-4) teaches the model to…

A. Write its own training code

B. Call external APIs like Search or utilities like calculators when needed

C. Compress its parameters to fit on smaller/lighter devices when needed

GPT-N series: Evolution of Models

Model (year)	Model Innovations	“Can you tell me who the current pope is”	Training Corpus Size (Estimate)
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**	**“Because I'm dying for that information”** Autocomplete of typical trained text. Ignores the question’s intent.	~1 B → 300 B tokens (5 GB → 45 TB raw)
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF	**“Pope Francis is the current pope”** Model follows instruction but is stale.	GPT-3 base 300 B + ≲ 1 B RLHF tokens
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt	**“The Church is currently without a pope”** Knows how to use Google to get up-to-date information.	~ 10–13 T tokens (undisclosed)
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation	*(THINKING)…quietly reasons, runs a quick search…* **“The Church is currently without a pope”** More capable with using Google and other tools to arrive at answers.	~ 10-20 T pre-training tokens + large-scale RL & tool-use fine-tuning (exact size undisclosed)

Chain-of-Thought Reinforcement: Teaching models to think out loud

Sam had 20 balloons, gave 4 balloons to each of his 3 friends, and then found 5 more balloons. How many balloons does Sam have now? *(Show your work)*

LLM

Reinforce (reward) correct solution path

Model Output (Trace 1 – Correct)
• Sam begins with 20 balloons.
• He gives each of his 3 friends 4 balloons (total given: 4 × 3 = 12).
• Remaining after giving: 20 − 12 = 8.
• He then finds 5 more balloons.
• Total balloons now = 8 + 5 = 13.
13

Ground truth:
13

Model Output (Trace 2 – Wrong)
• Sam starts with 20 balloons.
• He gives each of his 3 friends 4 balloons (4 × 3 = 12).
• So we add: 20 + 12 = 32.
• He then finds 5 more balloons.
• Total balloons now = 32 + 5 = 37.

37

Live demo: Using o3

Thought for 32 seconds

- The user is asking about the pope. While I believe it's still Pope Francis as of May 2025, I should double-check just to make sure. Since this relates to a current event (and leadership roles can change), I plan to search for the latest updates on who the current pope is. I'll look up "current pope" or "Pope as of 2025" to confirm and provide the most accurate information!

Searched the web

current pope 2025 | who is pope now may 2025

New York Post | CBS News | The Guardian | 3 more

- The user asked about the pope, and while I initially thought it was still Pope Francis, I wanted to double-check since things can change. After searching, it turns out Pope Francis passed away on April 21, 2025, and there is currently a conclave in progress to elect a new pope. Looks like my assumption was off, but I'll make sure to keep the user updated with the latest information!

Done

As of today (7 May 2025) the Apostolic See is **vacant**.

Check your understanding

Encouraging "Chain-of-Thought" processes in LLMs, where the model "thinks out loud" or shows its work, is primarily aimed at:

A. Significantly reducing the computational resources needed to generate a response.

B. Improving the model's reasoning process and its ability to correctly solve more complex, multi-step problems.

GPT-N series: Evolution of Models

Model (year)	Model Innovations	"Can you tell me who the current pope is"	Training Corpus Size (Estimate)
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**	**"Because I'm dying for that information"** Autocomplete of typical trained text. Ignores the question's intent.	~1 B → 300 B tokens (5 GB → 45 TB raw)
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF	**"Pope Francis is the current pope"** Model follows instruction but is stale.	GPT-3 base 300 B + ≲ 1 B RLHF tokens
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt	**"The Church is currently without a pope"** Knows how to use Google to get up-to-date information.	~ 10–13 T tokens (undisclosed)
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation	*(THINKING)…quietly reasons, runs a quick search…* **"The Church is currently without a pope"** More capable with using Google and other tools to arrive at answers.	~ 10-20 T pre-training tokens + large-scale RL & tool-use fine-tuning (exact size undisclosed)

An analogy to human learning

Model (year)	Model Innovations
GPT-1 → GPT-3 (2018-2020)	**Next-token prediction (pretraining)**
ChatGPT-3.5 (2022)	**Instruction-following** fine-tune + RLHF
ChatGPT-4 (2023-2024)	**Tool-use** fine-tune or tool-use system prompt
o-series (o1, o3) (2024-2025)	**Chain-of-thought** deliberation

Pretraining ↔ Exposition
(background knowledge)

Instruction Fine-tune ↔ Worked Examples
(problem + solution for imitation)

Tool Fine-tune ↔ Calculator Instructions
Showing how to use tools

Chain of Thought RL ↔ Practice Exercises
Prompt to practice, trial & error till you reach the correct answer

If you want to calculate the actual size of a specimen seen with a microscope, you need to know the diameter of the microscope's field of vision. This can be calculated with a special micrometre, or on a light microscope with a simple ruler. The size of the specimen can then be worked out. Drawings or photographs of specimens are often enlarged. To calculate the magnification of a drawing or photograph, a simple formula is used:

magnification = size of image/by size of specimen.

Worked example

You are walking outside with a friend who is wearing a red and white shirt. Explain why the shirt appears to be red and white.

Solution

Sunlight is a mixture of all of the wavelengths (colours) of visible light. When sunlight strikes the red pigments in the shirt, the blue and the green wavelengths of light are absorbed, but the red wavelengths are reflected. Thus, our eyes see red. When sunlight strikes the white areas of the shirt, all the wavelengths of light are reflected and our eyes and brain interpret the mixture as white.

Calculator and Google use are allowed

2 Exercises

These questions are found throughout the text. They allow you to apply your knowledge and test your understanding of what you have just been reading.

The answers to these are given in the eBook at the end of each chapter.

Exercises

25 Explain why a blue object appears to be blue to the human eye.
26 Explain why black surfaces (like tarmacadam and asphalt) get much hotter in sunlight than lighter surfaces (like stone and concrete).
27 Plants produce sugars by photosynthesis. What do plants do with the sugars after that?
28 Why do most plants produce an excess of sugars in some months of the year?

Caveats of our “history”

• Left out some important innovations:
- Multimodality (4 to 4o)
- Longer context length

• Only focused on one company, but similar sequence in others (Anthropic, Google, Meta)

• Simplifications for approachability

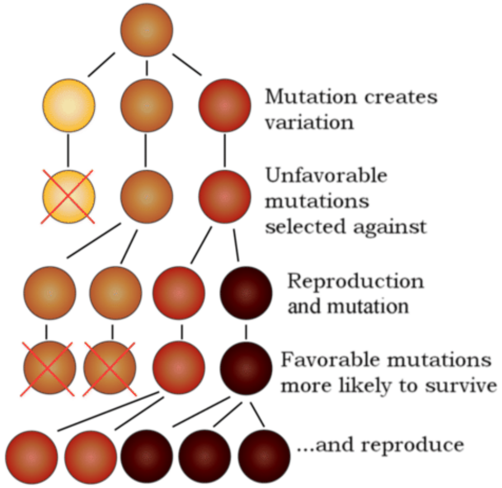
But how can such simple training procedures give rise to intelligence?! Two responses

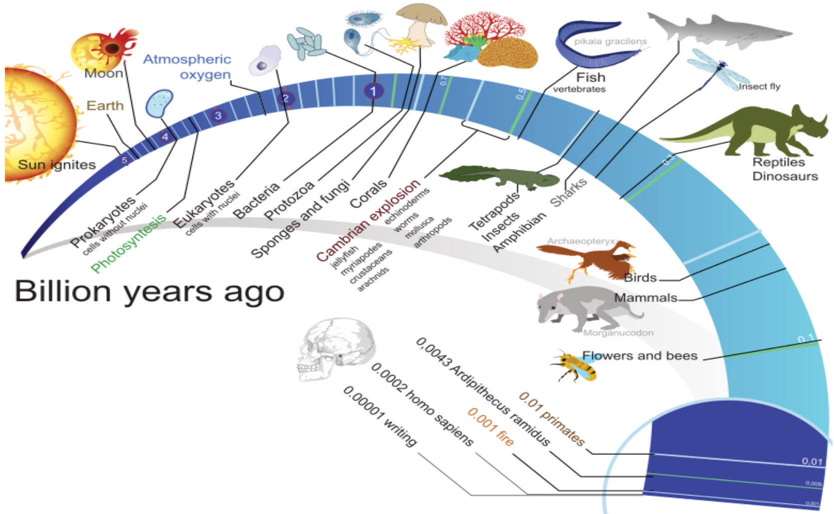
1: You're in good company!

Anthropic CEO Admits We Have No Idea How AI Works

"This lack of understanding is essentially unprecedented in the history of technology."

2: Consider the other simple procedure that gave rise to intelligence when scaled over trillions of iterations: natural selection/evolution





Why learn the history of LLMs?

Forecasting requires understanding the past

Pitfalls can be better avoided

Future trends

Improving Intelligence

The rapid pace from GPT-1 (2018) to o3 (2025) suggests increasingly improving capabilities ahead.

Expanding Context

Models will handle increasingly larger documents and conversations natively.

Better Multimodality

Integration of text, image, audio, and video in unified models.

Agentic Systems

Reasoning models are glimpse into a future of agents that can plan, execute tasks, and learn from their experiences.

Persistent Limitations

Hallucination

Models are still autocompleters and storytellers (though very intelligent ones). Hallucination problems persist and are built into how the models are trained.

Models have little self-knowledge

Models are trained on internet text. Have little knowledge of their own limitations. Often overconfident.

Knowledge Staleness

Base models only know information up to their training cutoff date.

Compute Costs

Training and running models is expensive in compute and therefore in dollars. The main players are very rich companies.

Privacy Concerns

Because of the compute needs, it's hard to train or run your own models locally. Privacy will continue to be an issue.

Bias Issues

Models reflect biases present in their training data