How LLMs Work?

High level look at internal workings for the general audience

Mohahmmad "Kiyarash" Fazeli

May 15, 2025

The Fundamental Surprise

- ChatGPT generates human-like text ...

The Fundamental Surprise

- ChatGPT generates human-like text ...
- ... through predicting next word in a text, one word at a time

The Fundamental Surprise

- ChatGPT generates human-like text ...
- ... through predicting next word in a text, one word at a time
- No explicit understanding - pure mathematical prediction

The Fundamental Surprise

- ChatGPT generates human-like text ...
- ... through predicting next word in a text, one word at a time
- No explicit understanding - pure mathematical prediction
- Emergent capabilities

The Core Mechanism

- Primary objective: Produce "reasonable continuation" of text

The Core Mechanism

- Primary objective: Produce "reasonable continuation" of text
- Analogous to human conversation prediction

The Core Mechanism

- Primary objective: Produce "reasonable continuation" of text
- Analogous to human conversation prediction
- Built from >1 trillion word patterns

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

What Makes Text "Reasonable"?

- Statistical likelihood derived from:

What Makes Text "Reasonable"?

- Statistical likelihood derived from:
  - Billions of web pages

What Makes Text "Reasonable"?

- Statistical likelihood derived from:
  - Billions of web pages
  - Digitized books

What Makes Text "Reasonable"?

- Statistical likelihood derived from:
  - Billions of web pages
  - Digitized books
  - Online conversations

What Makes Text "Reasonable"?

- Statistical likelihood derived from:
  - Billions of web pages
  - Digitized books
  - Online conversations
- Not rule-based - pure probability

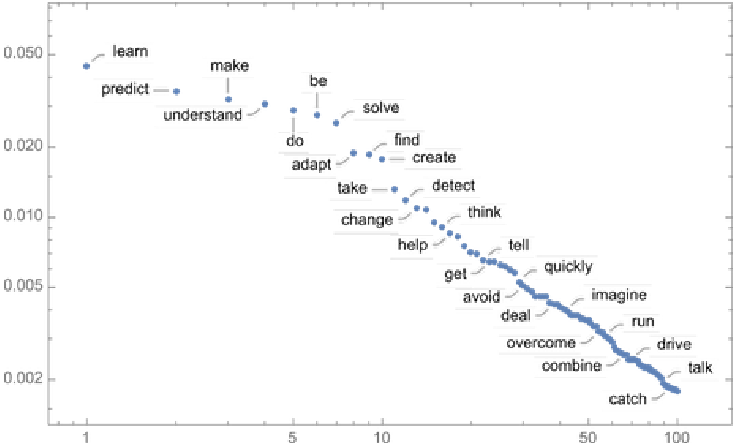
Example: Predicting the Next Word

- Input: "The best thing about AI is its ability to..."

The best thing about AI is its ability to

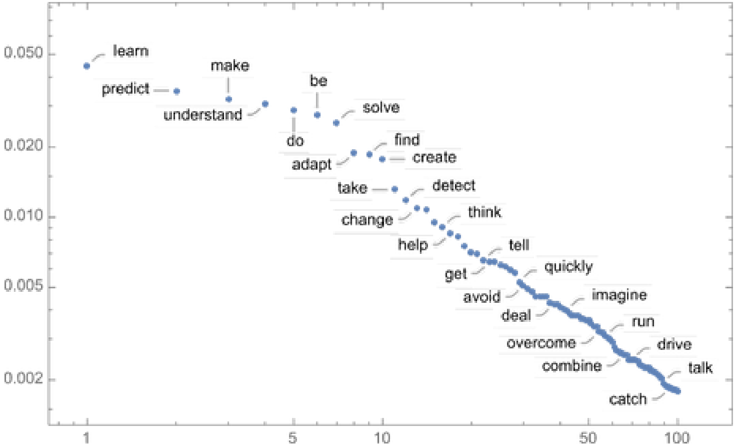
learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

Probability Distribution in Action



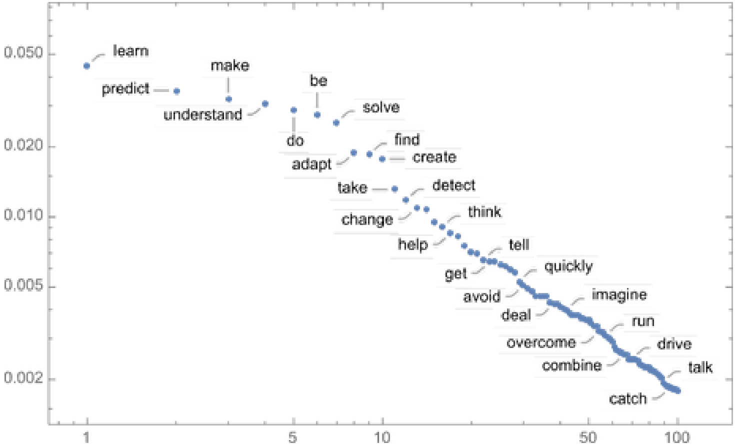
- High-probability choices: "learn", "adapt"

Probability Distribution in Action



- High-probability choices: "learn", "adapt"
- Creative possibilities: "dream", "collaborate"

Probability Distribution in Action



- High-probability choices: "learn", "adapt"
- Creative possibilities: "dream", "collaborate"
- Long-tail distribution pattern

It's Just Adding One Word at a Time

- LLMs(Deepseek, Chatgpt and etcs)

It's Just Adding One Word at a Time

- LLMs(Deepseek, Chatgpt and etcs)
- LLMs creates text through iterative prediction:

It's Just Adding One Word at a Time

- LLMs(Deepseek, Chatgpt and etcs)
- LLMs creates text through iterative prediction:

- Fundamental operation: "What's the next word given previous text?"

It’s Just Adding One Word at a Time

- LLMs(Deepseek, Chatgpt and etcs)
- LLMs creates text through iterative prediction:

- Fundamental operation: ”What’s the next word given previous text?”
- Trained on >1 trillion words from web pages/books

Controlled Randomness

- Pure max-probability leads to flat text:

  The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed,

Controlled Randomness

- Pure max-probability leads to flat text:

  The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed,

- Solution: Introduce temperature

Controlled Randomness

- Pure max-probability leads to flat text:

  The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed,

- Solution: Introduce temperature
- 0.8 optimal. Why 0.8? Empirical finding, not theory

Controlled Randomness

- Pure max-probability leads to flat text:

  > The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed,

- Solution: Introduce temperature
- 0.8 optimal. Why 0.8? Empirical finding, not theory
- Multiple runs create different outputs

Controlled Randomness

- Pure max-probability leads to flat text:
  The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed,
- Solution: Introduce temperature
- 0.8 optimal. Why 0.8? Empirical finding, not theory
- Multiple runs create different outputs
- Beam Search

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in
- Hallucination?

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit ”understanding” built in
- Hallucination?
- No difference, since no built in facts

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in
- Hallucination?
- No difference, since no built in facts
- What to do?

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in
- Hallucination?
- No difference, since no built in facts
- What to do?
- Verifying Manually

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit ”understanding” built in
- Hallucination?
- No difference, since no built in facts
- What to do?
- Verifying Manually
- Grounding

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in
- Hallucination?
- No difference, since no built in facts
- What to do?
- Verifying Manually
- Grounding
- RAG

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit ”understanding” built in
- Hallucination?
- No difference, since no built in facts
- What to do?
- Verifying Manually
- Grounding
- RAG
- Chain of thought

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in
- Hallucination?
- No difference, since no built in facts
- What to do?
- Verifying Manually
- Grounding
- RAG
- Chain of thought
- MCP

The Fundamental Surprise

- Human-like text emerges from simple next-word prediction
- No explicit "understanding" built in
- Hallucination?
- No difference, since no built in facts
- What to do?
- Verifying Manually
- Grounding
- RAG
- Chain of thought
- MCP
- Code Eval

The Fundamental Surprise

- Discovery vs Invention

The Fundamental Surprise

- Discovery vs Invention
- Information disclosure(prompts,personal, commercial)

The Fundamental Surprise

- Discovery vs Invention
- Information disclosure(prompts,personal, commercial)
- Bad data(MS twitter bot, reddit and google suggests)

The Fundamental Surprise

- Discovery vs Invention
- Information disclosure(prompts,personal, commercial)
- Bad data(MS twitter bot, reddit and google suggests)
- Undermining programmers?

The Fundamental Surprise

- Discovery vs Invention
- Information disclosure(prompts,personal, commercial)
- Bad data(MS twitter bot, reddit and google suggests)
- Undermining programmers?
- Geoffrey Hinton, Acemoglu

The Fundamental Surprise

- Discovery vs Invention
- Information disclosure(prompts,personal, commercial)
- Bad data(MS twitter bot, reddit and google suggests)
- Undermining programmers?
- Geoffrey Hinton, Acemoglu
- Retrieval vs. Generation

Opensource?

- Opensource?

Opensource?

- Opensource?
- Open weights

Opensource?

- Opensource?
- Open weights
- Finetunning

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations
- GPU Nvidia market segmentation

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations
- GPU Nvidia market segmentation
- Usage -> still limited

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations
- GPU Nvidia market segmentation
- Usage -> still limited
- Quantization -> slow feasible

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations
- GPU Nvidia market segmentation
- Usage -> still limited
- Quantization -> slow feasible
- Smaller models

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations
- GPU Nvidia market segmentation
- Usage -> still limited
- Quantization -> slow feasible
- Smaller models
- Distillation

Opensource?

- Opensource?
- Open weights
- Finetunning
- Training -> closed source, 5M$+
- GPU back of napkin calculations
- GPU Nvidia market segmentation
- Usage -> still limited
- Quantization -> slow feasible
- Smaller models
- Distillation
- Polling resources?

Eval

- Perplexity?

Eval

- Perplexity?
- MMLU

Eval

- Perplexity?
- MMLU
- MMLU like in Farsi

Eval

- Perplexity?
- MMLU
- MMLU like in Farsi
- Livecoding Benchmark

Eval

- Perplexity?
- MMLU
- MMLU like in Farsi
- Livecoding Benchmark
- Chatbot Arena Explanation

Token Generation Process

- Text is chunked into tokens (words/subwords)
- Model estimates probability distribution over possible next tokens
- Selection involves randomness (temperature parameter)

Example of tokens: Deepseek

antidisestablishmentarianism

noun

The doctrine or political position that opposes the withdrawal of state recognition of an established church; -- used especially concerning the Anglican Church in England. Opposed to disestablishmentarianism.

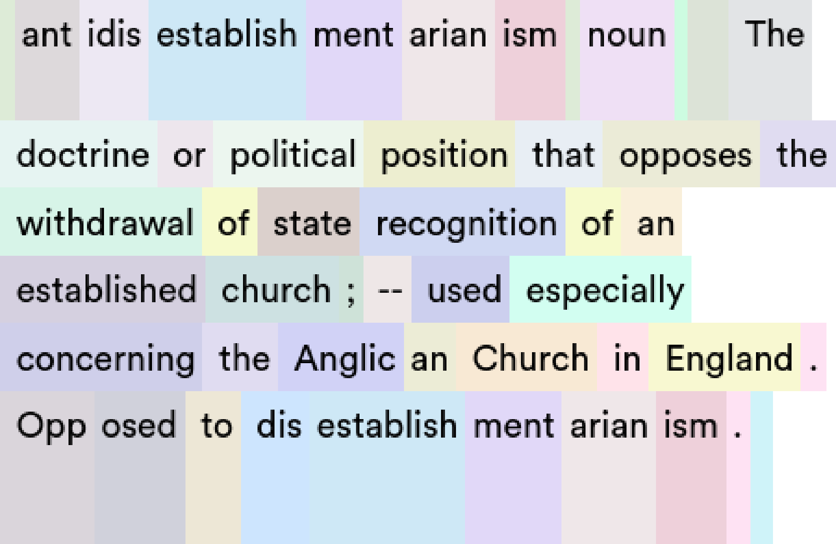
Example of tokens: Deepseek

Deepseek

ant idis establish ment arianism noun The doctrine or political position that opposes the withdrawal of state recognition of an established church ; -- used especially concerning the Anglican Church in England . Opp osed to dis establish ment arianism .

Example of tokens: ChatGPT

GPT-4(o, o-mini, o1-preview, o1-mini)



2nd Example of tokens

Kiyarash Fazeli

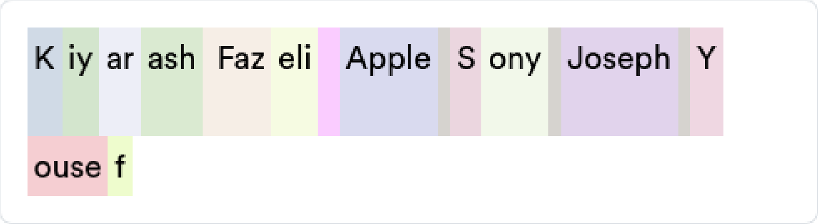
Apple

Sony

Joseph

Yousef

2nd Example of tokens



Deepseek

GPT

Probability in Action

At Each Step:

- Calculates probabilities for all possible next tokens
- Doesn't just pick highest probability
- Maintains creativity through probabilistic sampling

The Iterative Nature

- Each step feeds output back as input
- Context window grows with each iteration
- Surprisingly maintains coherence over long sequences

Why Does This Work?

- Emergent complexity from simple rules
- Massive neural network
- Captures linguistic patterns at multiple scales

Key Limitations

Important Caveats

- No true "understanding" of content
- Can't perform logical deductions
- Errors compound through iterations

Summary

- Statistical next-word prediction at scale
- Emergent complexity from simple iteration
- Combination of pattern recognition + controlled randomness

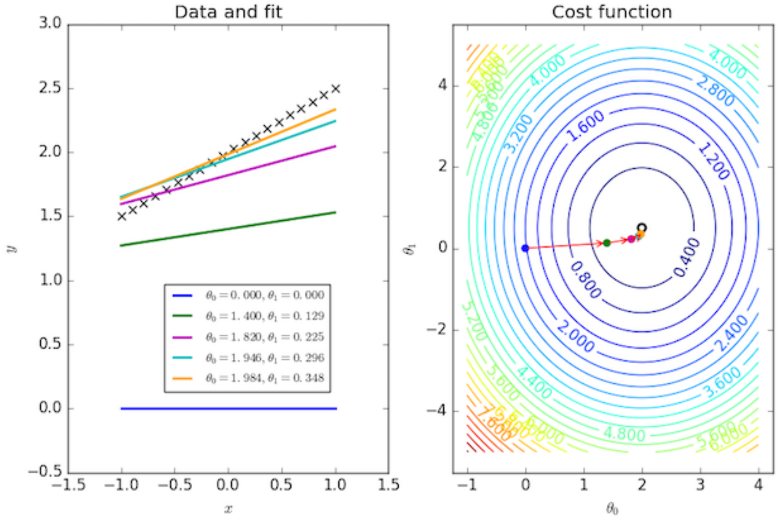
Goldberg IPython Notebook

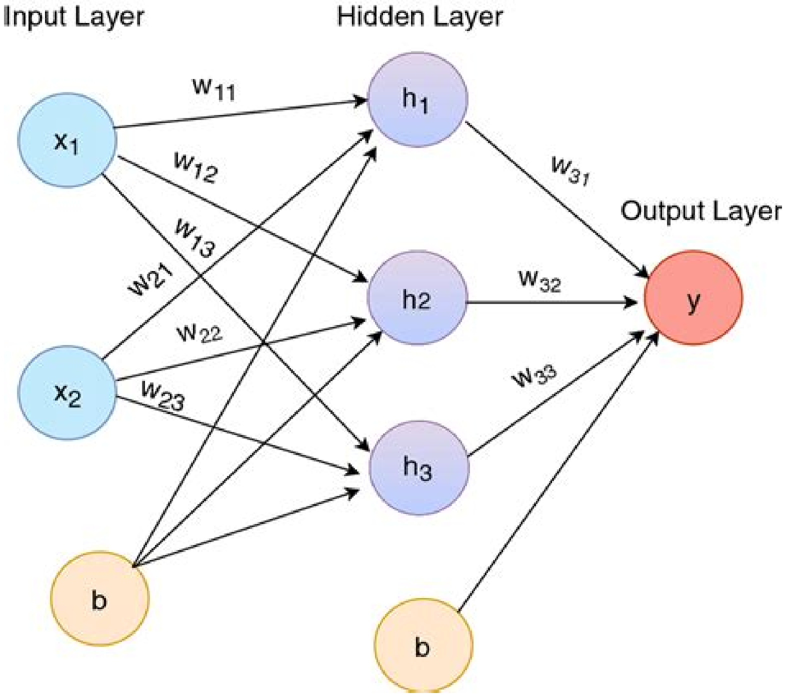
Differentiable Functions

Input→Output Pairs
Cost Minimisation

Random Start $\rightarrow$ $-$*Gradient* Steps

Example of Linear Regression




Input Layer
Hidden Layer
Output Layer
x_1
x_2
b
h_1
h2
h_3
b
y
w_{11}
w_{12}
w_{13}
w_{21}
w_{22}
w_{23}
w_{31}
w_{32}
w_{33}

3Blue1Brown Visual Intuition Series

Emdedding

My Thesis Examples

Attention as Enriching Embedding

...

Result of Attention: Context Length

...

MCP