CASE FILE #2026-AI-001

The Case of the Thinking Machine

How Dead Mathematicians Accidentally Built AI

Prof. Jörg Osterrieder

CLASSIFIED · 45 MINUTES · 10 SUSPECTS

The Crime Scene

BREAKING NEWS

Machine Scores Gold at Math Olympiad

The Crime Scene

BREAKING NEWS

Machine Scores Gold at Math Olympiad

IMO 2025: Google DeepMind's AlphaProof scores **35/42**

Nobel 2024: Physics prize to Hopfield & Hinton for neural networks

ChatGPT: 900 million users worldwide

35/42

IMO Gold Score

900M

ChatGPT Users

2024

Nobel for AI

The Crime Scene

BREAKING NEWS

Machine Scores Gold at Math Olympiad

IMO 2025: Google DeepMind's AlphaProof scores **35/42**

Nobel 2024: Physics prize to Hopfield & Hinton for neural networks

ChatGPT: 900 million users worldwide

How did we get here?

35/42

IMO Gold Score

900M

ChatGPT Users

2024

Nobel for AI

The Suspects Board

UNKNOWN CONNECTION TO AI

The Suspects Board

UNKNOWN CONNECTION TO AI

~100 BCE Chinese Mathematicians — counting rods

1654 Pascal & Fermat — gambling letters

1666 Newton — plague year calculus

1684 Leibniz — published calculus

1763 Bayes — published posthumously

The Suspects Board

UNKNOWN CONNECTION TO AI

~100 BCE Chinese Mathematicians — counting rods

1654 Pascal & Fermat — gambling letters

1666 Newton — plague year calculus

1684 Leibniz — published calculus

1763 Bayes — published posthumously

1844 Grassmann — vectors nobody read

1847 Cauchy — gradient descent

1858 Cayley — matrix algebra

1933 Kolmogorov — probability axioms

1948 Shannon — information theory

The Suspects Board

UNKNOWN CONNECTION TO AI

~100 BCE Chinese Mathematicians — counting rods

1654 Pascal & Fermat — gambling letters

1666 Newton — plague year calculus

1684 Leibniz — published calculus

1763 Bayes — published posthumously

1844 Grassmann — vectors nobody read

1847 Cauchy — gradient descent

1858 Cayley — matrix algebra

1933 Kolmogorov — probability axioms

1948 Shannon — information theory

None of them knew they were building AI.

The Victim’s Testimony

EXHIBIT A: AI FAILURES

The Victim’s Testimony

EXHIBIT A: AI FAILURES

Q: Is 9.11 greater than 9.9?
ChatGPT: “Yes, 9.11 > 9.9” WRONG

Q: How many R’s in “strawberry”?
ChatGPT: “Two” WRONG (it’s three)

The Victim's Testimony

EXHIBIT A: AI FAILURES

Q: Is 9.11 greater than 9.9?
ChatGPT: "Yes, 9.11 > 9.9" WRONG

Q: How many R's in "strawberry"?
ChatGPT: "Two" WRONG (it's three)

Something deeper is going on...

These aren't bugs. They're clues about how AI actually works. Understanding the math reveals *why* it fails — and why it succeeds.

The Central Question

> *“The same machine that wins gold at the Math Olympiad cannot count to three.”*
>
> — The Paradox of Modern AI

The Central Question

> *“The same machine that wins gold at the Math Olympiad cannot count to three.”*
>
> — The Paradox of Modern AI

THREAD 1

Where did the **math** come from?

THREAD 2

How does it all **fit together**?

THREAD 3

Why does it **fail**?

The Investigation Plan

TABLE OF CONTENTS

Act 1: The Crime Scene
Headlines, suspects, the paradox

Act 2: The Cold Cases
Five mathematical investigations:
Linear Algebra · Probability · Calculus
Information Theory · Optimization

Act 3: The Evidence Room
How all five connect in the transformer

Act 4: The Revelation
The verdict, the race, what “thinking” means

Act 5: The Case Continues
Open investigations, your turn, resources

ACT 2

The Cold Cases

Five Mathematical Investigations

CLUE #1

Chinese Counting Rods

CLUE #1

Chinese Counting Rods

~100 BCE Jiuzhang Suanshu — “Nine Chapters”

Chinese mathematicians laid bamboo rods in rows and columns to solve systems of equations. This *is* Gaussian elimination — 2,000 years before Gauss.

Arrange rods in rows and columns:

$$\begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{pmatrix}$$

Eliminate row by row…

CLUE #1

Chinese Counting Rods

~100 BCE Jiuzhang Suanshu — “Nine Chapters”

Chinese mathematicians laid bamboo rods in rows and columns to solve systems of equations. This *is* Gaussian elimination — 2,000 years before Gauss.

Han Dynasty context: The Silk Road is opening. Paper has just been invented. Rome is at its peak. And in China, someone is doing matrix operations with sticks on a counting board.

Arrange rods in rows and columns:

$$\begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{pmatrix}$$

Eliminate row by row…

SUSPECT

Hermann Grassmann

Hermann Grassmann

1809–1877

Occupation: Schoolteacher in Stettin, Prussia

SUSPECT

Hermann Grassmann

Hermann Grassmann

1809–1877

Occupation: Schoolteacher in Stettin, Prussia

1844: Publishes *Ausdehnungslehre* (“Theory of Extension”) — invents vector spaces, linear independence, dimension. Nobody reads it.

SUSPECT

Hermann Grassmann

Hermann Grassmann

1809–1877

Occupation: Schoolteacher in Stettin, Prussia

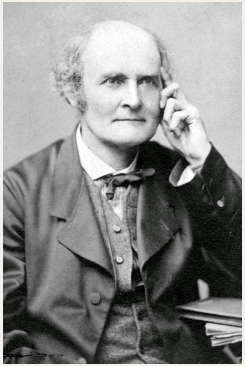
1844: Publishes *Ausdehnungslehre* (“Theory of Extension”) — invents vector spaces, linear independence, dimension. Nobody reads it.

Gives up math, becomes a Sanskrit scholar (wins a prize for it).

World in 1844: Marx writes the Communist Manifesto draft. Morse sends first telegraph. Grassmann invents the math of AI — and nobody notices.

SUSPECT

Arthur Cayley



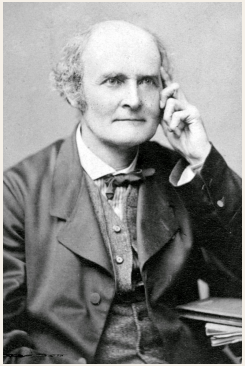
Arthur Cayley

1821–1895

Occupation: Lawyer (practiced law for 14 years to fund his math)

SUSPECT

Arthur Cayley



Arthur Cayley

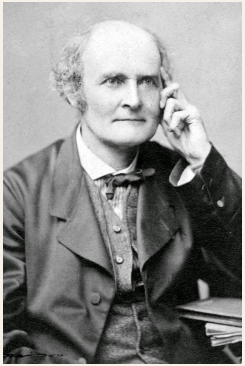
1821–1895

Occupation: Lawyer (practiced law for 14 years to fund his math)

1858: Invents matrix algebra *in his spare time*. Defines multiplication, inverses, determinants.

SUSPECT

Arthur Cayley



Arthur Cayley

1821–1895

Occupation: Lawyer (practiced law for 14 years to fund his math)

1858: Invents matrix algebra *in his spare time*. Defines multiplication, inverses, determinants.

Publishes **900+ papers** across his career.

World in 1858: Darwin will publish *Origin of Species* next year. Lincoln-Douglas debates. Cayley gives us the notation that every neural network uses.

EVIDENCE

Words Become Geometry

EVIDENCE

Words Become Geometry

2013: MIKOLOV

Word2Vec: train a neural network, and words arrange themselves in geometric space.

$$\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$$

EVIDENCE

Words Become Geometry

2013: MIKOLOV

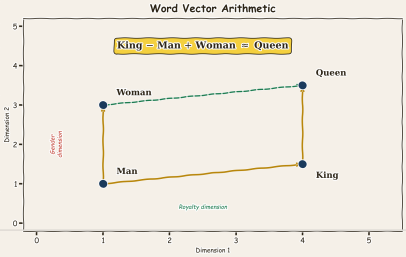
Word2Vec: train a neural network, and words arrange themselves in geometric space.

$$\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$$

GPT-4 today:
Every word → a vector in 12,288 dimensions.

Meaning *is* direction.
Analogy *is* vector arithmetic.

169 years from Grassmann to Word2Vec.



EVIDENCE

The Engine of AI

$$\mathbf{output} = W\mathbf{x} + \mathbf{b}$$

Matrix $\times$ vector + bias
The fundamental operation of every neural network

EVIDENCE

The Engine of AI

$$\mathbf{output} = W\mathbf{x} + \mathbf{b}$$

Matrix × vector + bias
The fundamental operation of every neural network

1.8 Trillion

MULTIPLICATIONS PER TOKEN

EVIDENCE

The Engine of AI

$$\mathbf{output} = W\mathbf{x} + \mathbf{b}$$

Matrix × vector + bias
The fundamental operation of every neural network

1.8 Trillion

MULTIPLICATIONS PER TOKEN

Every layer, every attention head, every feed-forward block: it is all matrix multiplication.

Cayley's 1858 notation, executed **1.8 trillion times** for every single word GPT-4 produces.

166 years from Cayley to ChatGPT.

CLUE #2

A Gambler’s Complaint

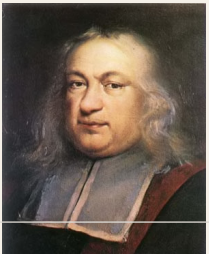
CLUE #2

A Gambler’s Complaint

1654 Chevalier de Méré loses money

A French nobleman complains to Blaise Pascal: “The math says I should win, but I keep losing!” Pascal writes to Fermat. Their letters invent probability theory.

Pascal



CLUE #2

A Gambler's Complaint

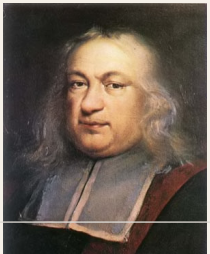
1654 Chevalier de Méré loses money

A French nobleman complains to Blaise Pascal: "The math says I should win, but I keep losing!" Pascal writes to Fermat. Their letters invent probability theory.

The Problem of Points: Two players, interrupted game. How to split the pot *fairly* based on who was winning?

World in 1654: Louis XIV rules France. Cromwell rules England. And two mathematicians invent probability over a gambling complaint.

Pascal



SUSPECT

Thomas Bayes

Thomas Bayes

1701–1761

Occupation: Presbyterian minister, amateur mathematician

SUSPECT

Thomas Bayes

Thomas Bayes

1701–1761

Occupation: Presbyterian minister, amateur mathematician

Shy, published almost nothing in his lifetime. His theorem found after death, published **1763** by friend Richard Price.

SUSPECT

Thomas Bayes

Thomas Bayes

1701–1761

Occupation: Presbyterian minister, amateur mathematician

Shy, published almost nothing in his lifetime. His theorem found after death, published **1763** by friend Richard Price.

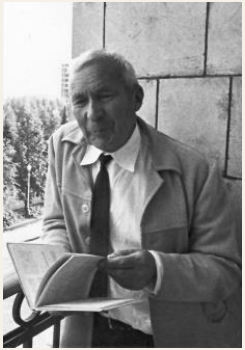
$$P(H \mid E) = \frac{P(E|H)\,P(H)}{P(E)}$$

Update your belief when new evidence arrives.

Fun fact: The famous portrait may not even be him — it might be someone else entirely.

SUSPECT

Andrey Kolmogorov



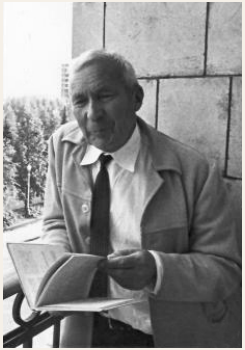
Andrey Kolmogorov

1903–1987

Occupation: Soviet mathematician, Moscow State University

SUSPECT

Andrey Kolmogorov



Andrey Kolmogorov

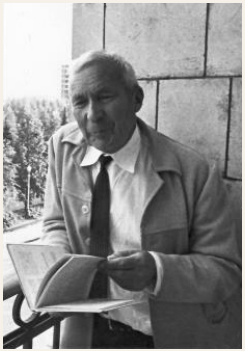
1903–1987

Occupation: Soviet mathematician, Moscow State University

1933: Publishes axioms of probability *in German* — puts the entire field on rigorous footing. Three axioms. Everything follows.

SUSPECT

Andrey Kolmogorov



Andrey Kolmogorov

1903–1987

Occupation: Soviet mathematician, Moscow State University

1933: Publishes axioms of probability *in German* — puts the entire field on rigorous footing. Three axioms. Everything follows.

Key connection: The softmax function in every LLM satisfies Kolmogorov's axioms exactly — it produces a valid probability distribution.

World in 1933: Hitler takes power. FDR inaugurated. Dust Bowl begins. Kolmogorov axiomatizes probability.

EVIDENCE

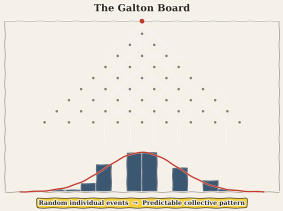
The Galton Board

EVIDENCE

The Galton Board

1889: GALTON'S QUINCUNX

Drop balls through a grid of pegs. Each ball bounces randomly left or right. At the bottom: a perfect bell curve.

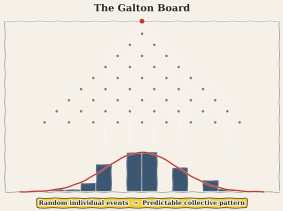


EVIDENCE

The Galton Board

1889: GALTON'S QUINCUNX

Drop balls through a grid of pegs. Each ball bounces randomly left or right. At the bottom: a perfect bell curve.



The LLM parallel:

Random tokens → coherent text

Each token is a ball bouncing through probability distributions. Individually random, collectively meaningful.

Randomness, constrained by structure, produces order.

EVIDENCE

50,000 Words, One Probability

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}}$$

Turn any list of numbers into a
valid probability distribution

EVIDENCE

50,000 Words, One Probability

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}}$$

Turn any list of numbers into a
valid probability distribution

Kolmogorov's axioms satisfied:
$\forall i : P(w_i) \geq 0 \quad \sum_i P(w_i) = 1$

EVIDENCE

50,000 Words, One Probability

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}}$$

Turn any list of numbers into a **valid probability distribution**

Kolmogorov's axioms satisfied:

$\forall i : P(w_i) \geq 0 \quad \sum_i P(w_i) = 1$

Every time GPT picks a word, softmax converts 50,000+ raw scores into probabilities.

The highest probability wins.
Temperature controls randomness.

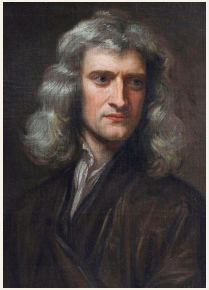
A gambler's complaint → axioms → every word your AI writes.

CLUE #3

The Nastiest Fight in Math

CLUE #3

The Nastiest Fight in Math



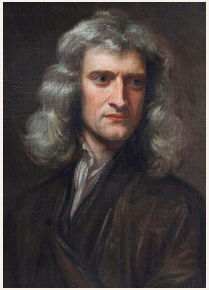
Newton, 1666

VS

Leibniz, 1684

CLUE #3

The Nastiest Fight in Math



Newton, 1666

VS

Leibniz, 1684

Newton: Invented calculus during the plague (1666), locked it in a drawer.

Leibniz: Published first (1684), better notation.

The scandal: Newton rigged the Royal Society investigation.

World: Great Plague, Great Fire of London. Newton hiding on a farm, inventing physics.

The Royal Society "investigated" — with Newton as president. Verdict: Newton first. But we use Leibniz's notation: $\frac{dy}{dx}$

SUSPECT

Augustin-Louis Cauchy

Augustin-Louis Cauchy

1789–1857

Occupation: French mathematician, notoriously prolific

SUSPECT

Augustin-Louis Cauchy

Augustin-Louis Cauchy

1789–1857

Occupation: French mathematician, notoriously prolific

1847: Describes gradient descent — solve equations by repeatedly stepping in the direction that reduces error.

"Like being blindfolded on a hill: feel which way is downhill, take a step, repeat."

SUSPECT

Augustin-Louis Cauchy

Augustin-Louis Cauchy

1789–1857

Occupation: French mathematician, notoriously prolific

1847: Describes gradient descent — solve equations by repeatedly stepping in the direction that reduces error.

"Like being blindfolded on a hill: feel which way is downhill, take a step, repeat."

World in 1847: Irish Famine kills a million. Mexican-American War. Cauchy describes the algorithm that will train every neural network.

EVIDENCE

How AI Learns

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

Parameters = current − **learning rate** × **gradient**

EVIDENCE

How AI Learns

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

Parameters = current − **learning rate** × **gradient**

Derivative = direction (which way is down?)

Learning rate η = step size (how far to go?)

Repeat billions of times = trained model

EVIDENCE

How AI Learns

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

Parameters = current − **learning rate** × **gradient**

Derivative = direction (which way is down?)

Learning rate η = step size (how far to go?)

Repeat billions of times = trained model

The chain rule:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial w}$$

Leibniz's 1684 notation powers backpropagation: compute how each weight affects the loss, layer by layer, backwards.

SUSPECT

Geoffrey Hinton

Geoffrey Hinton

Born 1947

Occupation: Computer scientist, “Godfather of Deep Learning”

SUSPECT

Geoffrey Hinton

Geoffrey Hinton

Born 1947

Occupation: Computer scientist, “Godfather of Deep Learning”

1986: Publishes backpropagation with Rumelhart & Williams in *Nature*. The chain rule + gradient descent, applied to neural networks.

SUSPECT

Geoffrey Hinton

Geoffrey Hinton

Born 1947

Occupation: Computer scientist, “Godfather of Deep Learning”

1986: Publishes backpropagation with Rumelhart & Williams in *Nature*. The chain rule + gradient descent, applied to neural networks.

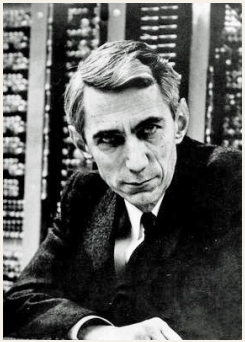
Decades of rejection. “Neural nets are dead.” Persists anyway.

2024: Nobel Prize in Physics for neural network foundations.

Leibniz’s chain rule $\rightarrow$ 338 years $\rightarrow$ Nobel Prize.

CLUE #4

The Unicycling Genius



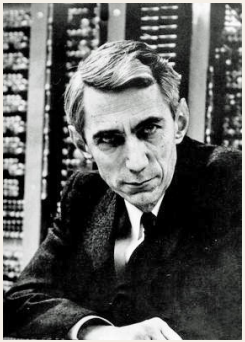
Claude Shannon

1916–2001

Occupation: Bell Labs engineer, MIT professor

CLUE #4

The Unicycling Genius



Claude Shannon

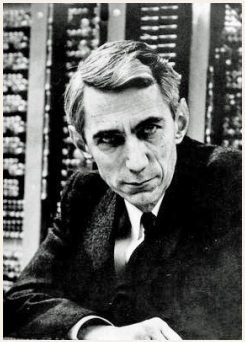
1916–2001

Occupation: Bell Labs engineer, MIT professor

1948: “A Mathematical Theory of Communication” — invents the bit, information entropy, channel capacity. One paper creates an entire field.

CLUE #4

The Unicycling Genius



Claude Shannon

1916–2001

Occupation: Bell Labs engineer, MIT professor

1948: "A Mathematical Theory of Communication" — invents the bit, information entropy, channel capacity. One paper creates an entire field.

Also: juggling unicyclist, flame-throwing trumpet player, built a machine whose only purpose was to turn itself off.

World in 1948: Post-WWII, Berlin Airlift, first transistor at Bell Labs. Shannon gives us the math of information itself.

EVIDENCE

Cross-Entropy: The Loss Function of AI

$$H(P, Q) = -\sum_{x} P(x) \log Q(x)$$

How **surprised** is the model?

EVIDENCE

Cross-Entropy: The Loss Function of AI

$$H(P, Q) = -\sum_{x} P(x) \log Q(x)$$

How **surprised** is the model?

P = true distribution (actual next word)

Q = model's prediction

Minimize surprise = learn language

EVIDENCE

Cross-Entropy: The Loss Function of AI

$$H(P, Q) = -\sum_{x} P(x) \log Q(x)$$

How **surprised** is the model?

P = true distribution (actual next word)

Q = model's prediction

Minimize surprise = learn language

Shannon designed this for telephone networks.

"How many bits do I need to encode this message?"

Same formula, 76 years later, is the training objective of every large language model.

A formula for telephones trains every AI.

EVIDENCE

Shannon’s Blueprint → ChatGPT

EVIDENCE

Shannon’s Blueprint → ChatGPT

SHANNON 1948

Source → Encoder → Channel → Decoder → Destination

EVIDENCE

Shannon’s Blueprint → ChatGPT

SHANNON 1948

Source → Encoder → Channel → Decoder → Destination

CHATGPT 2024

User → Tokenizer → Transformer → Output Layer → Response

EVIDENCE

Shannon's Blueprint → ChatGPT

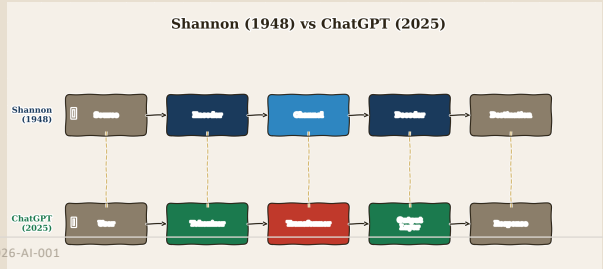
SHANNON 1948

Source → Encoder → Channel → Decoder → Destination

CHATGPT 2024

User → Tokenizer → Transformer → Output Layer → Response

The architecture of every LLM is Shannon's communication model with neural networks replacing the components. The math is identical. The application changed from sending telegrams to generating text.



CLUE #5

Making the Impossible Possible

CLUE #5

Making the Impossible Possible

1951 Robbins & Monro — **Stochastic Gradient Descent**

Don’t use all the data. Use random samples. It’s noisy but fast enough.

CLUE #5

Making the Impossible Possible

1951 Robbins & Monro — **Stochastic Gradient Descent**

Don’t use all the data. Use random samples. It’s noisy but fast enough.

1964 Polyak — **Momentum**

Add a “memory” of past gradients. Like a ball rolling downhill with inertia.

CLUE #5

Making the Impossible Possible

1951 Robbins & Monro — **Stochastic Gradient Descent**

Don’t use all the data. Use random samples. It’s noisy but fast enough.

1964 Polyak — **Momentum**

Add a “memory” of past gradients. Like a ball rolling downhill with inertia.

2014 Kingma & Ba — **Adam**

Adaptive learning rates per parameter. 200,000+ citations. The default optimizer of the deep learning era.

200K+

CITATIONS FOR ADAM

Without these optimizers, training a trillion-parameter model would take centuries.

With them: weeks.

EVIDENCE

The Scaling Law

$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}$$

Loss decreases as a power law of model size

EVIDENCE

The Scaling Law

$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}$$

Loss decreases as a power law of model size

Kaplan et al., 2020

More parameters → predictably better.
More data → predictably better.
More compute → predictably better.

Companies spend billions on this bet.

EVIDENCE

The Scaling Law

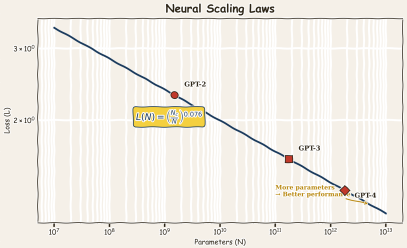
$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}$$

Loss decreases as a power law of model size

Kaplan et al., 2020

More parameters → predictably better.
More data → predictably better.
More compute → predictably better.

Companies spend billions on this bet.



Plot twist: DeepSeek R1 matched GPT-4 performance for $6 million instead of $100M+.

Scaling laws are real, but efficiency matters too.

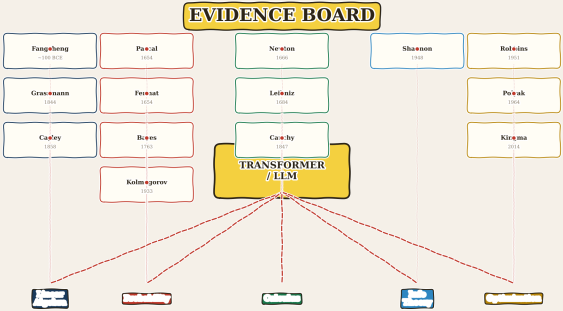
ACT 3

The Evidence Room

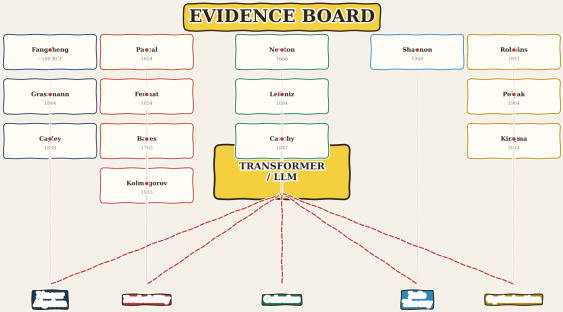
Connecting All the Evidence

The Evidence Board

The Evidence Board



The Evidence Board



Lin. Alg.

Vec-tors

Probab.

Soft-max

Cal-cu-lus

Back-prop

Info Thy

Loss

Op-tim.

Adam

The Attention Mechanism

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

The Attention Mechanism

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

$Q = XW_Q$	Cayley's matrices
$K = XW_K$	Grassmann's vectors
softmax	Kolmogorov's axioms
$\sqrt{d_k}$	Scaling for stability

The Attention Mechanism

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

$Q = XW_Q$	Cayley's matrices
$K = XW_K$	Grassmann's vectors
softmax	Kolmogorov's axioms
$\sqrt{d_k}$	Scaling for stability

Detective annotations:

Grassmann $\rightarrow$ words as vectors X

Cayley $\rightarrow$ weight matrices W_Q, W_K, W_V

Kolmogorov $\rightarrow$ softmax probabilities

Cauchy $\rightarrow$ gradient descent trains W

Shannon $\rightarrow$ cross-entropy loss guides training

One Forward-Backward Pass

FORWARD

BACKWARD

One Forward-Backward Pass

FORWARD BACKWARD

Forward pass (prediction):

1. Words → vectors (Grassmann)
2. Matrix multiply through layers (Cayley)
3. Attention scores via dot products
4. Softmax → probabilities (Kolmogorov)
5. Pick next word

Backward pass (learning):

6. Compare to truth: cross-entropy (Shannon)
7. Chain rule backward (Leibniz)
8. Gradient for every weight (Cauchy)
9. Adam updates parameters (Kingma)
10. Repeat trillions of times

One Forward-Backward Pass

FORWARD BACKWARD

Forward pass (prediction):

1. Words → vectors (Grassmann)
2. Matrix multiply through layers (Cayley)
3. Attention scores via dot products
4. Softmax → probabilities (Kolmogorov)
5. Pick next word

Backward pass (learning):

6. Compare to truth: cross-entropy (Shannon)
7. Chain rule backward (Leibniz)
8. Gradient for every weight (Cauchy)
9. Adam updates parameters (Kingma)
10. Repeat trillions of times

All five cold cases, running in a single cycle, billions of times per second.

The Numbers

1.7T
Parameters

13T
Training Tokens

$100M+
Training Cost

900M
Users

$6M
DeepSeek Cost

The Numbers

1.7T	13T	$100M+	900M	$6M
Parameters	Training Tokens	Training Cost	Users	DeepSeek Cost

GPT-4 is estimated at 1.7 trillion parameters trained on 13 trillion tokens at a cost exceeding $100 million. It serves 900 million users. Then DeepSeek built a competitor for roughly $6 million. The math works at every scale — the question is how cleverly you use it.

Brilliant and Broken

BRILLIANT

BROKEN

Brilliant and Broken

BRILLIANT

- ✓ Gold medal at Math Olympiad
- ✓ Passes the bar exam
- ✓ Writes publishable code
- ✓ Translates 100+ languages
- ✓ Diagnoses rare diseases

BROKEN

Brilliant and Broken

BRILLIANT

- ✓ Gold medal at Math Olympiad
- ✓ Passes the bar exam
- ✓ Writes publishable code
- ✓ Translates 100+ languages
- ✓ Diagnoses rare diseases

BROKEN

- × “9.11 > 9.9”
- × Cannot count R’s in “strawberry”
- × Invents fake citations
- × Confidently wrong about facts
- × No persistent memory

Brilliant and Broken

BRILLIANT

- ✓ Gold medal at Math Olympiad
- ✓ Passes the bar exam
- ✓ Writes publishable code
- ✓ Translates 100+ languages
- ✓ Diagnoses rare diseases

BROKEN

- × “9.11 > 9.9”
- × Cannot count R’s in “strawberry”
- × Invents fake citations
- × Confidently wrong about facts
- × No persistent memory

Pattern completers, not fact databases. The math explains both the brilliance and the failures. LLMs are extraordinary at pattern matching and terrible at things that require counting, tracking, or grounding in external reality.

ACT 4

The Revelation

The Verdict

The Verdict

VERDICT

Who Built AI?

The Verdict

VERDICT

Who Built AI?

Chinese mathematicians — counting rods

Grassmann — vector spaces

Cayley — matrix algebra

Pascal & Fermat — probability

Bayes — updating beliefs

Kolmogorov — axioms

Newton & Leibniz — calculus

Cauchy — gradient descent

Shannon — information theory

Hinton — backpropagation

The Verdict

VERDICT

Who Built AI?

Chinese mathematicians — counting rods
Grassmann — vector spaces
Cayley — matrix algebra
Pascal & Fermat — probability
Bayes — updating beliefs

Kolmogorov — axioms
Newton & Leibniz — calculus
Cauchy — gradient descent
Shannon — information theory
Hinton — backpropagation

> *“None of them knew. All of them contributed. The case is closed.”*

The Unreasonable Effectiveness

> *“The unreasonable effectiveness of mathematics in the natural sciences is something bordering on the mysterious.”*
>
> — Eugene Wigner, 1960

The Unreasonable Effectiveness

> *"The unreasonable effectiveness of mathematics in the natural sciences is something bordering on the mysterious."*
>
> — Eugene Wigner, 1960

Invented for X…

Vectors → geometry
Matrices → solving equations
Probability → gambling
Calculus → planetary motion
Entropy → telegraph networks

…now powers Y

Vectors → word meaning
Matrices → neural computation
Probability → language generation
Calculus → backpropagation
Entropy → training objective

The Race

TIMELINE

2017 Vaswani et al. — “Attention Is All You Need”

2018 BERT (Google) — bidirectional understanding

2018 GPT-1 (OpenAI) — 117M parameters

2019 GPT-2 — “too dangerous to release”

2020 GPT-3 — 175B parameters, few-shot learning

2020 Scaling laws discovered (Kaplan)

2022 ChatGPT — 100M users in 2 months

2023 GPT-4 — multimodal, bar exam

2024 Nobel Prize to Hopfield & Hinton

2024 Claude 3, Gemini, Llama 3

2025 DeepSeek, IMO gold, open weights

2025 AI agents, reasoning models

The Race

TIMELINE

2017 Vaswani et al. — “Attention Is All You Need”

2018 BERT (Google) — bidirectional understanding

2018 GPT-1 (OpenAI) — 117M parameters

2019 GPT-2 — “too dangerous to release”

2020 GPT-3 — 175B parameters, few-shot learning

2020 Scaling laws discovered (Kaplan)

2022 ChatGPT — 100M users in 2 months

2023 GPT-4 — multimodal, bar exam

2024 Nobel Prize to Hopfield & Hinton

2024 Claude 3, Gemini, Llama 3

2025 DeepSeek, IMO gold, open weights

2025 AI agents, reasoning models

Eight years from one paper to reshaping civilization.

What Makes Machines “Think”?

What Makes Machines “Think”?

They don’t think.

What Makes Machines “Think”?

They don’t think.

They PREDICT.

What Makes Machines “Think”?

They don’t think.

They PREDICT.

Pattern matching $\neq$ understanding

An LLM finds the most probable next token given all previous tokens.
It has learned *patterns* from trillions of words.
Those patterns are so rich they *look* like understanding.

But: no world model. No persistent memory. No intent.
The strawberry test proves it: brilliant pattern completion, zero comprehension.

Five Suspects, Case Closed

LINEAR ALGEBRA

$$\mathbf{y} = W\mathbf{x} + \mathbf{b}$$

PROBABILITY

$$\text{softmax}(z_i)$$

CALCULUS

$$\frac{\partial L}{\partial w}$$

INFO THEORY

$$H(P, Q)$$

OPTIMIZATION

$$\theta_{t+1} = \theta_t - \eta \nabla L$$

Five Suspects, Case Closed

LINEAR ALGEBRA

$$\mathbf{y} = W\mathbf{x} + \mathbf{b}$$

PROBABILITY

$$\text{softmax}(z_i)$$

CALCULUS

$$\frac{\partial L}{\partial w}$$

INFO THEORY

$$H(P, Q)$$

OPTIMIZATION

$$\theta_{t+1} = \theta_t - \eta \nabla L$$

Five branches of mathematics. 2000+ years of history. Ten mathematicians who never met. One machine that writes, reasons, translates, codes, and sometimes cannot count to three.

CASE CLOSED

Five Suspects, Case Closed

LINEAR ALGEBRA

$$\mathbf{y} = W\mathbf{x} + \mathbf{b}$$

PROBABILITY

$$\text{softmax}(z_i)$$

CALCULUS

$$\frac{\partial L}{\partial w}$$

INFO THEORY

$$H(P, Q)$$

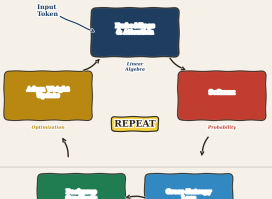
OPTIMIZATION

$$\theta_{t+1} = \theta_t - \eta \nabla L$$

Five branches of mathematics. 2000+ years of history. Ten mathematicians who never met. One machine that writes, reasons, translates, codes, and sometimes cannot count to three.

CASE CLOSED

One Training Step: All Five Pillars



ACT 5

The Case Continues

Open Investigations and Your Turn

Active Investigations

OPEN CASES

Sparse Attention
Full attention is $O(n^2)$. Can we make it $O(n \log n)$?
Longformer, BigBird, Mamba.

Calibration
Models say “I’m 90% sure” but are wrong 40% of the time. How do we fix confidence?

Second-Order Methods
Beyond first derivatives. Can curvature information speed up training?

Interpretability
We built it. We trained it. We still don’t fully understand *what it learned*.

New Pillars?
Topology, category theory, differential geometry — the next mathematical foundations of AI?

Your Turn

CAREER PATHS

ML Engineer
Build and deploy models
Median: $160K+

AI Safety Researcher
Ensure AI benefits humanity
Median: $170K+

Research Scientist
Invent new architectures
Median: $180K+

Quantitative Analyst
Math + AI in finance
Median: $200K+

Data Scientist
Extract insight from data
Median: $130K+

Your Turn

CAREER PATHS

ML Engineer
Build and deploy models
Median: $160K+

AI Safety Researcher
Ensure AI benefits humanity
Median: $170K+

Research Scientist
Invent new architectures
Median: $180K+

Quantitative Analyst
Math + AI in finance
Median: $200K+

Data Scientist
Extract insight from data
Median: $130K+

Every one of these careers uses the five mathematical pillars we covered today.

Your First Case

FREE RESOURCES

GitHub Student Developer Pack
Free Copilot, cloud credits, tools
`education.github.com`

Hugging Face
Open-source models, datasets, spaces
`huggingface.co`

Kaggle
Real datasets, competitions, free GPUs
`kaggle.com`

Competitions
Kaggle, USACO, Math Olympiad, Science Bowl
Start competing *now* — colleges notice.

Google Colab
Free Jupyter notebooks with GPU
`colab.research.google.com`

Your First Case

FREE RESOURCES

GitHub Student Developer Pack
Free Copilot, cloud credits, tools
`education.github.com`

Hugging Face
Open-source models, datasets, spaces
`huggingface.co`

Kaggle
Real datasets, competitions, free GPUs
`kaggle.com`

Competitions
Kaggle, USACO, Math Olympiad, Science Bowl
Start competing *now* — colleges notice.

Google Colab
Free Jupyter notebooks with GPU
`colab.research.google.com`

All free. All accessible to you today.

The Investigation Continues

> *“For over 2,000 years, mathematicians solved problems they thought were abstract and useless. They were wrong about the useless part. Every one of their ideas now lives inside a machine that writes poetry, proves theorems, and sometimes cannot count to three. The investigation is not over. The next breakthrough is waiting for someone in this room.”*
>
> — On the Shoulders of Centuries

CASE FILE #2026-AI-001

Case Dismissed

Thank You

Questions?

Prof. Jörg Osterrieder

https://digital-ai-finance.github.io/mathematics-for-ai/

Appendix: Evidence Inventory

FORMULA REFERENCE

LINEAR ALGEBRA

$\mathbf{y} = W\mathbf{x} + \mathbf{b}$ (neural network layer)

$\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$

PROBABILITY

$P(H \mid E) = \frac{P(E|H)\,P(H)}{P(E)}$ (Bayes)

$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$

CALCULUS

$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$ (gradient descent)

$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial w}$ (chain rule)

INFORMATION THEORY

$H(P, Q) = -\sum_x P(x) \log Q(x)$ (cross-entropy)

OPTIMIZATION

$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}$ (scaling law)

ATTENTION

$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$

$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$