I Tried Explaining LLMs to a Friend...

It turned into a 10-year history lesson.

The foundation began with a simple idea: predicting the probability of the next word.

Today, we are onto collaborative teams, complex agents, and the frontier of world models.

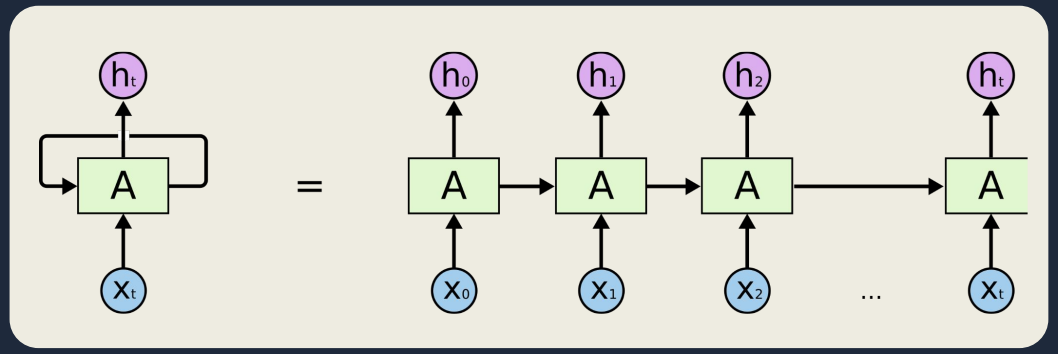
Let’s dive into the key moments that defined this journey, from the Scaling Wall to the ChatGPT shock.

For all the BBT-heads - Trying to be “less” Sheldon here so will spare the 2500 years history 😁

Pre-2016: The Path That Led To Nowhere

Sequential Processing

RNNs, LSTMs, and variants were the standard. They processed language word by word, like a human reading a line. This made them powerful but impossible to parallelize across massive GPU clusters.



The Scaling Wall

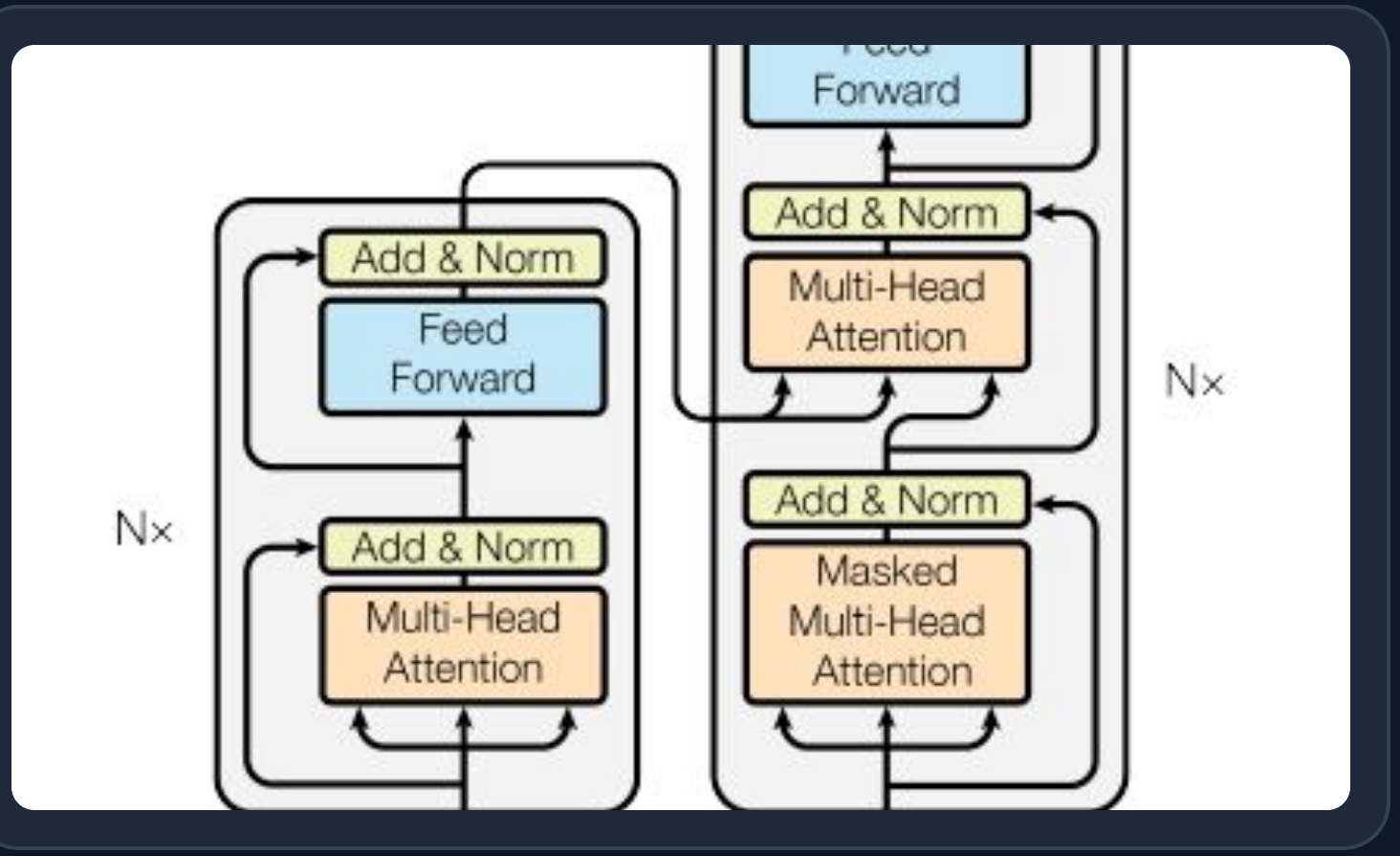
While deep learning was exploding in computer vision (ResNets), language models remained "stuck." They couldn't absorb the sheer scale of compute that was becoming available.

- **Slow Training Speeds**
- **Vanishing Gradients**

2017: Attention Changed Everything

Transformers replaced sequence with **Parallel Attention**.

- GPUs were finally fully utilized, unlocking massive scaling.
- "Bigger models = better performance" became a law, not a guess.
- The era of the "Next Token Predictor" had officially begun.



Check out “Attention is All You Need” - The research paper with a simple idea that set in motion the AI revolution we are witnessing today!

2022: The ChatGPT Product Shock

Same underlying tech, better alignment.

RLHF

Reinforcement Learning from Human Feedback

Bridged the gap between a raw model and a helpful assistant.

From lab research to 100M users in weeks.

The Gold Rush Phase

Innovation exploded. So did the cracks.

RAG

- LLMs + Real Data
- Reduced hallucinations

Prompt Engineering

- “Talking to AI” now a skill
- New frameworks

Chain-of-thought

- Step-by-step thinking
- Better yet fragile output

Images and Video

- Midjourney, DALL·E
- AI becomes creative

Bias/Hallucinations

- Bard failure moment
- Reality ≠ alignment

Copyright Wars

- Lawsuits begin
- “Who owns the data?”

2024: AI Became an Efficiency Race

Metric	OpenAI GPT-4/4o	DeepSeek V3	Efficiency Gain
Training Cost	~$100M	$6M	**16x Cheaper**
Inference Cost (per 1M)	~$10.00	$0.42	**95% Reduction**
Strategy	Closed General-Purpose	MoE / Mixed Precision	Optimized

98%

PROCESSING COST SAVINGS

The narrative shifted from "best possible model" to "best model per dollar, watt, and latency budget"

Open-weight ecosystems (LLaMA) changed access forever.

2025: From Chatbots to Agents

Collaborative Teams

Multiple AI agents acting like a team—exchanging tasks and validating each other's outputs.

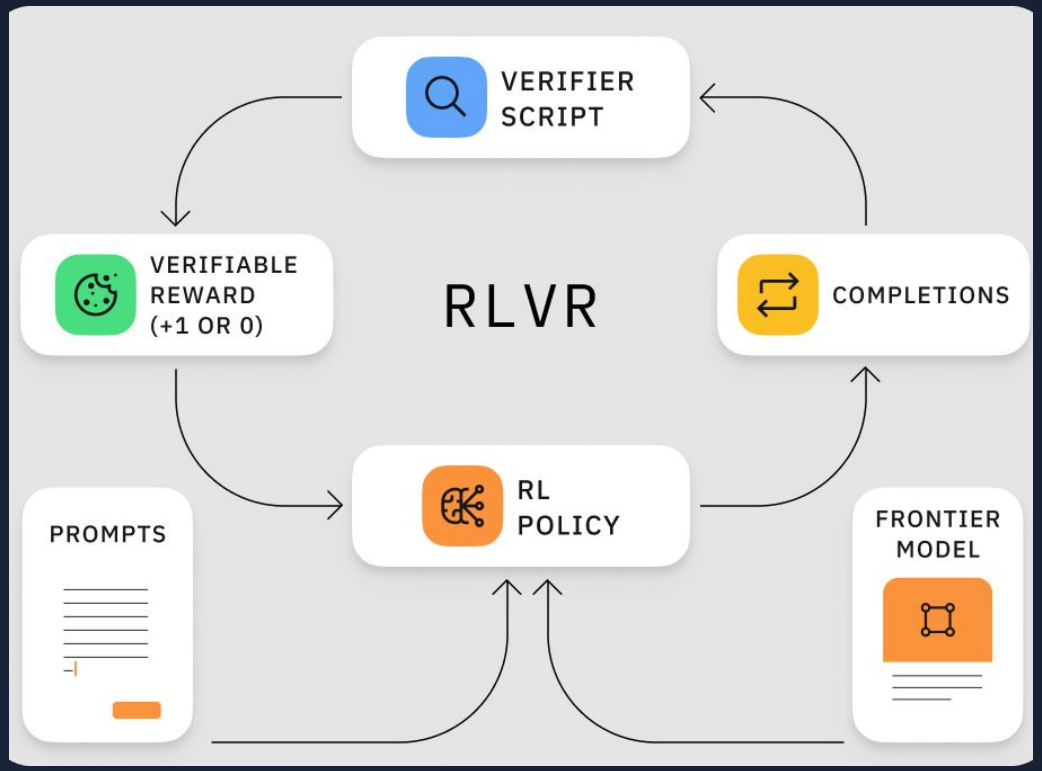
Interoperability

Protocols like MCP & A2A enable agents to communicate across different platforms and tools.

Structured Workflow

Systems now plan, act, and reason through repeatable workflows rather than simple responses.

System 2: Moving Beyond Prediction



Talking vs. Thinking

System 1 (Intuitive)

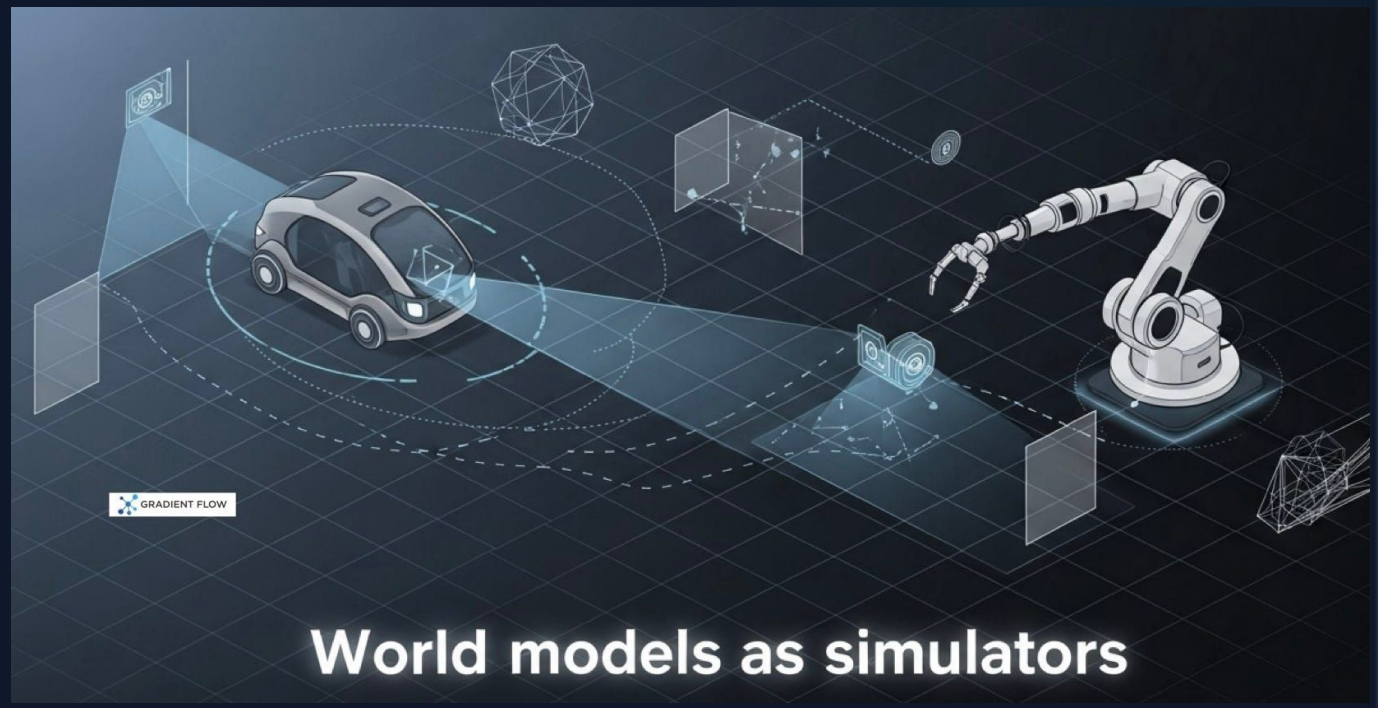
RLHF gave us models that are fluent and fast but prone to "hallucinations" and shallow patterns.

System 2 (Reasoning)

RL with verifiable rewards enables actual reasoning. Models now prove and verify their logic before they predict the next token.

The Frontier: World Models (e.g. JEPA)

- **Beyond Tokens:** Today we predict the next word. Tomorrow, we predict the next *state* of the world.
- **JEPA:** Joint Embedding Predictive Architecture learns like a human—by observation, not just massive data labeling.
- **Reality vs Description:** Systems that understand physics, occlusions, and cause-effect rather than just statistical text patterns.



The Real Story is Wilder

This was the **"non-PhD version"**. Not a Literature Review. Nor an exhaustive list. I just picked some key moments in the journey till today.

The transition from text prediction to **world understanding** is just beginning.

Additional Resources

- History of LLMs: Timeline & Evolution
- The Evolution of Language Models
- Transformers to Agentic AI Timeline
- The Agentic AI Handbook

Deep dive on evolution of deep learning: `sivampillai.com/essays/e002-deep-learning-2016`

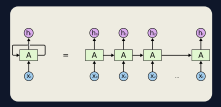
What is your bet on the next big moment for AI?

Feel free to share your comments below

Image Sources

[你认为节目中哪些角色的化学反 应最好，哪些最差？ : r/bigbangtheory]

Source: https://www.reddit.com/r/bigbangtheory/



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Source: https://colah.github.io

https://miro.medium.com/v2/resize:fit:736/1*wOAw-yM55afcI0ZUQ_73Lw.png

Source: mehmetozkaya.medium.com

https://labelbox.com/blog/how-to-create-data-for-reinforcement-learning-with-verifiable-rewards-rlvr/

Source: https://labelbox.com/

https://substackcdn.com/image/fetch/$s_!E39d!,f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2F091ba9a1-bae0-439f-ac3e-5d626a7f6ae1_1799×846.jpeg

Source: gradientflow.substack.com
