AI and You

What is AI, and how did we get here?

How can it help you with Amateur Radio?

This image generated by ChatGPT

Bruce Wampler, WA7EWC — January 18, 2025

Introduction

Overview of Presentation

- Why this presentation?
- What is AI?
- Supercomputer History
- AI History
- Overview of Generative AI
- How can it help you?
- Using AI to enhance Amateur Radio

An AI generated image to go with this presentation, including weird hallucinations.

What is AI?

"Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke

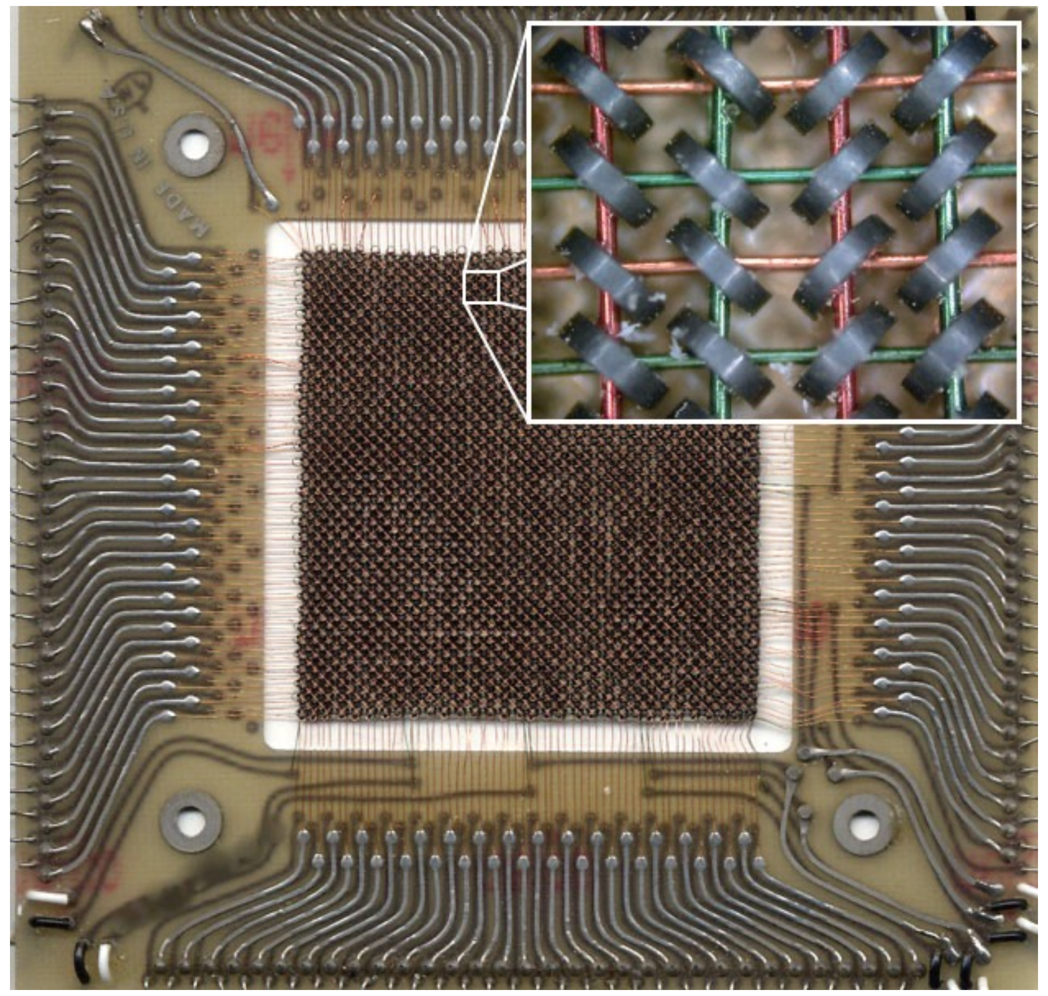
- Artificial Intelligence, in its broadest definition, is the effort to create systems that can simulate human-like intelligence.
- In just the past couple of years, it seems we may have reached the decades old goal of "real" AI with the introduction of *Generative AI,* which emphasizes our ability to engage in dialogue. Or, perhaps, to believe in magic.
- This presentation will attempt to explain how we got to this point, and maybe remove some of the magic. We got here through the advancement in computing power, as well as advancements in AI research, ultimately leading to modern Generative AI and the Large Language Models that are the core of chatbots.

Supercomputer History

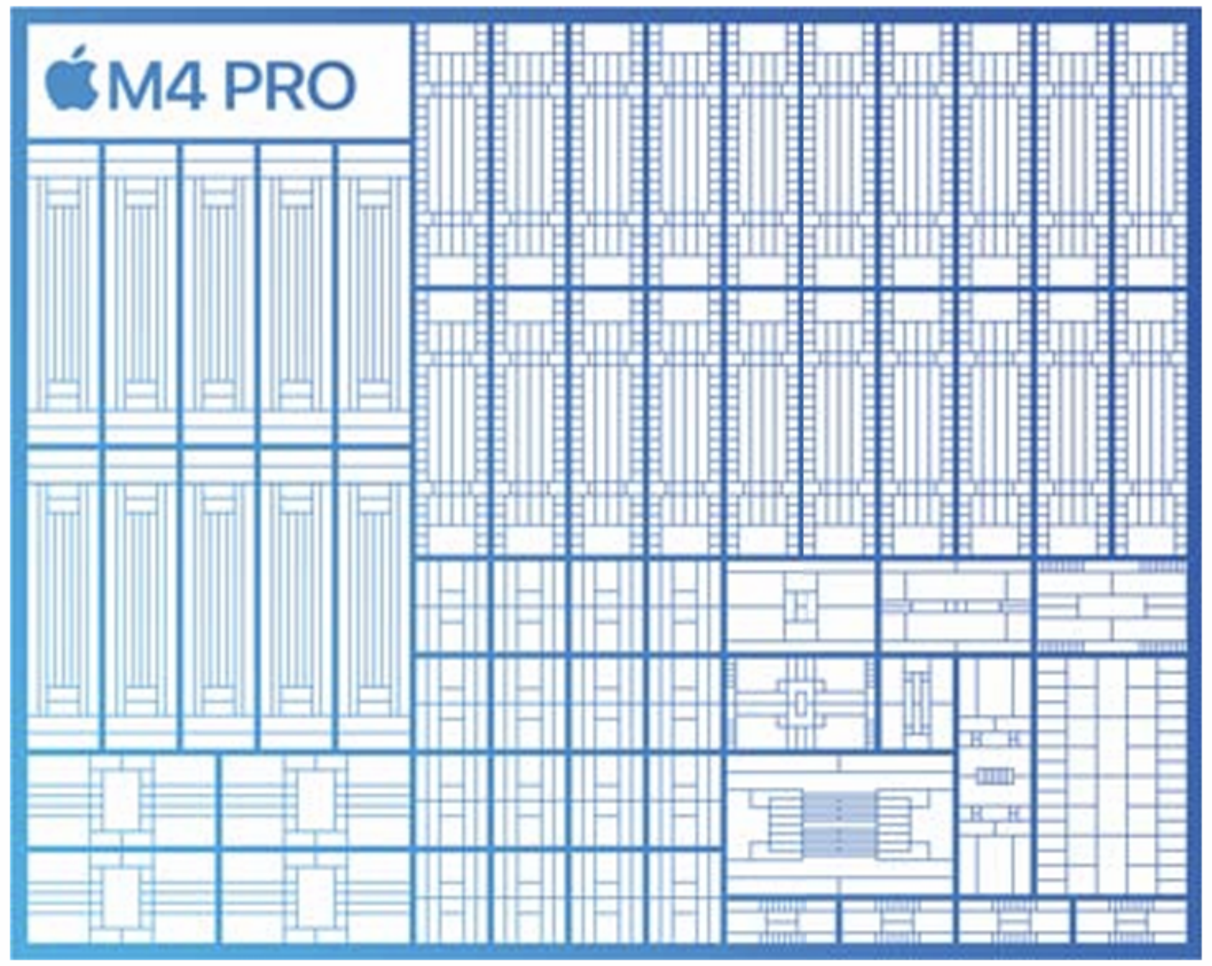
Evolution of Supercomputers: From Cray-1 to Frontier

A supercomputer is a high-performance computing system that delivers significantly greater processing power, memory, and speed than general-purpose computers. It achieves this through parallel processing using thousands or millions of interconnected processors working simultaneously. Modern supercomputers can perform quintillions of calculations per second and are primarily used for complex scientific calculations, weather forecasting, molecular modeling, and other computationally intensive tasks.

- **Architecture** - functional units and organization
- **Processors** - How many CPUs and GPUs
- **Memory** - memory available
- **Cache** - memory caching levels
- **Size** - physical size
- **Power** - how much power used
- **Cooling** - how the system is cooled



Early Core Memory



Apple M4 CPU
28 Billion transistors

Speed and Scale

In the following slides, we will be discussing the speed of different computers using FLOPS (Floating Operations Per Second) measurement.

The number of FLOPS typically uses one of the following prefixes such as GFLOPS. Each level is 1,000 times larger than the previous.

Kilo (k): 10^3, Thousand

Mega (M): 10^6, Million

Giga (G): 10^9, Billion

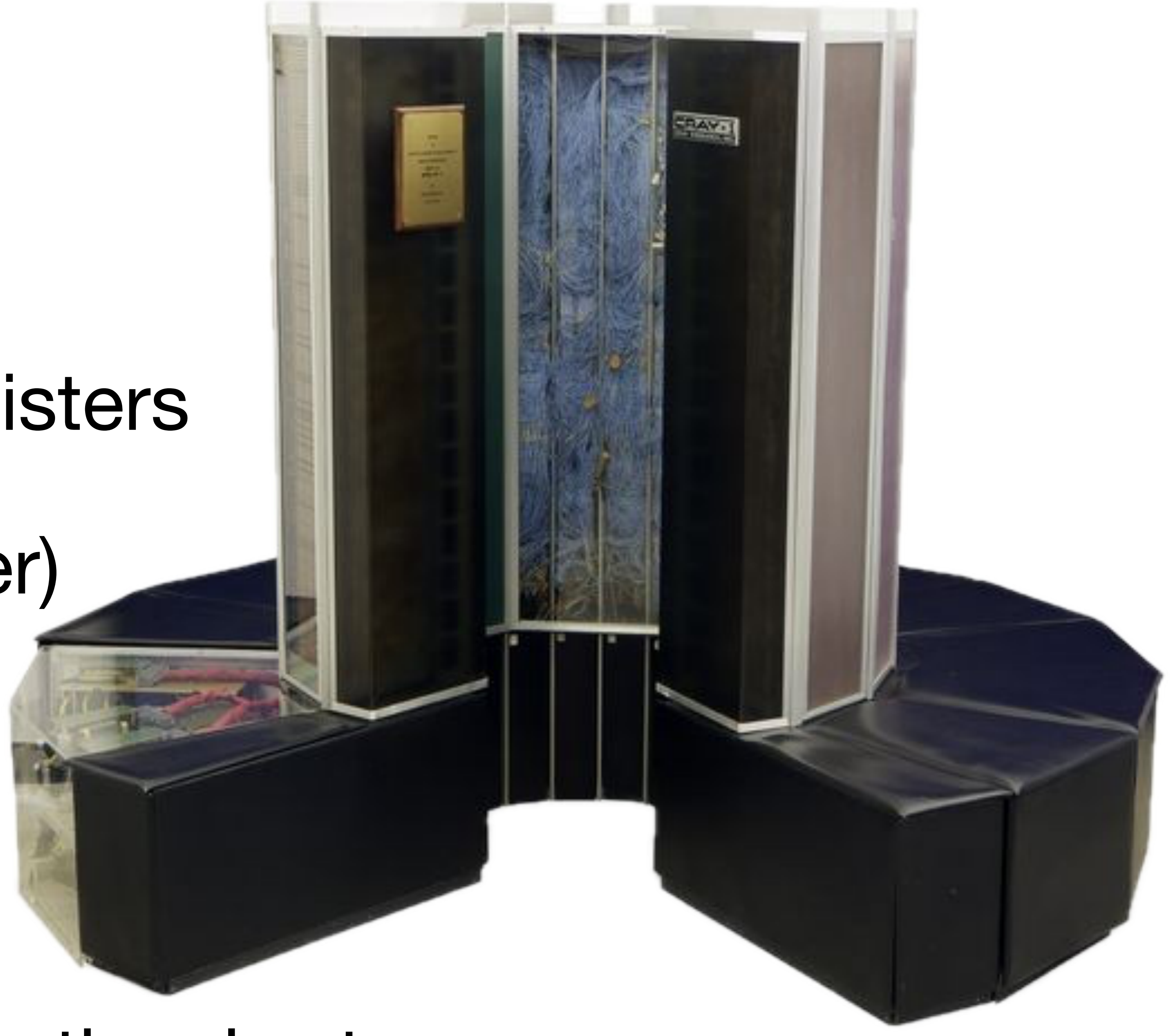
Tera (T): 10^12, Trillion

Peta (P): 10^15, Quadrillion

Exa (E): 10^18, Quintillion

Cray-1 (1976) - First Supercomputer to reach 160 MFLOPS

- Architecture: Vector processor with 12 independent functional units
- Single CPU running at 80 MHz
- Memory: 8 MB main memory (64-bit words)
- Cache: No typical memory cache, but had 16 vector registers
- Physical size: C-shaped cabinet (8.5 ft high , 9ft diameter)
- Power consumption: 115 kW
- Cooling: Freon cooling system with heat exchanger
- Notable for its iconic C-shaped design to keep cable lengths short

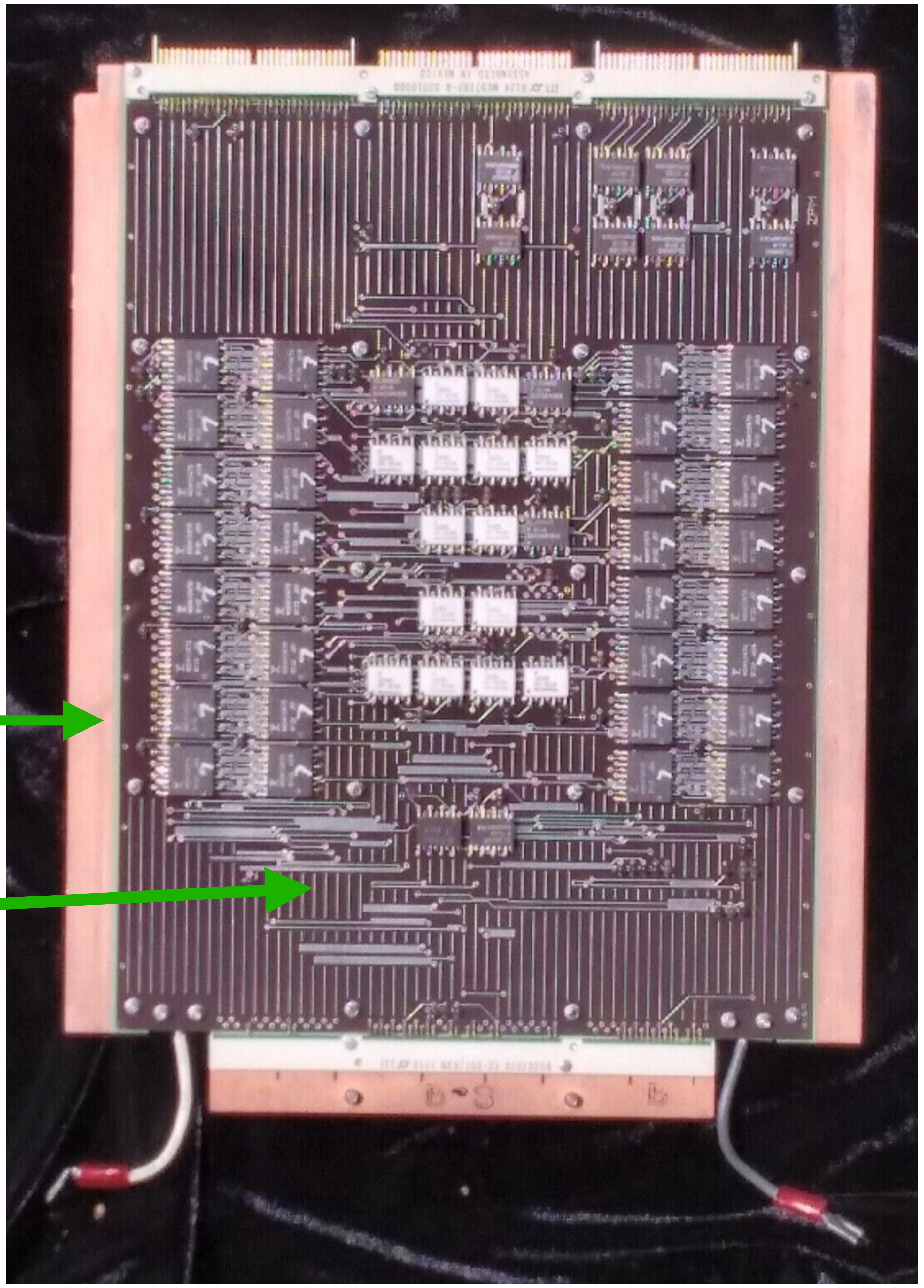


Supercomputers - Main Random Access Memory

Cray-1 used Freon-based cooling

Copper plate attached to cooling system

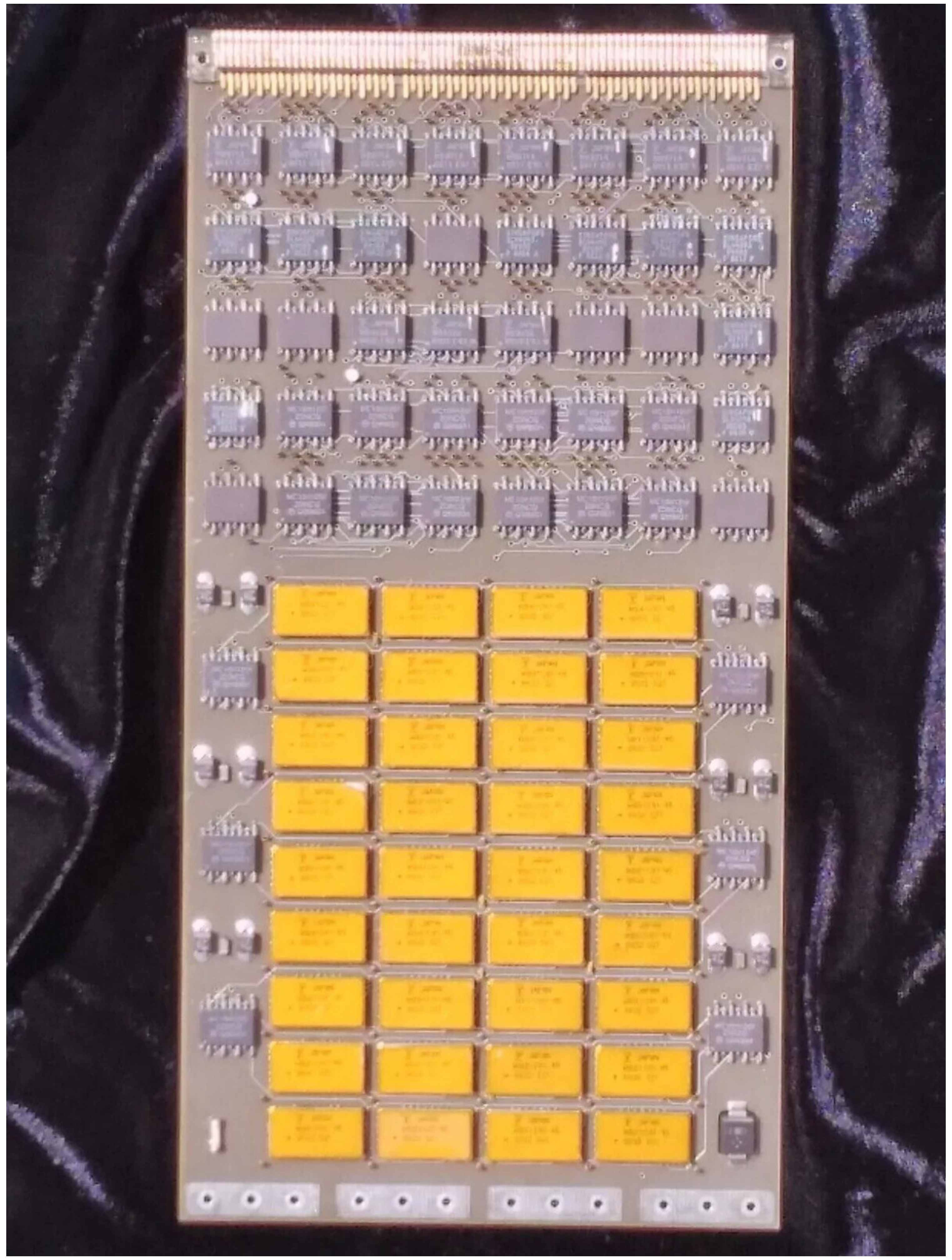
Notice long traces for memory timing signals



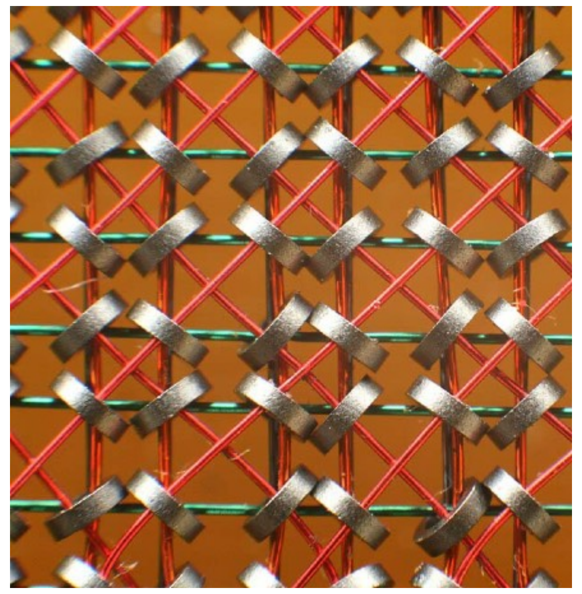
Cray-1 4K x 32 bit memory, each side

Board Vertical Height: 8 inches

50 nanosecond access time SRAM

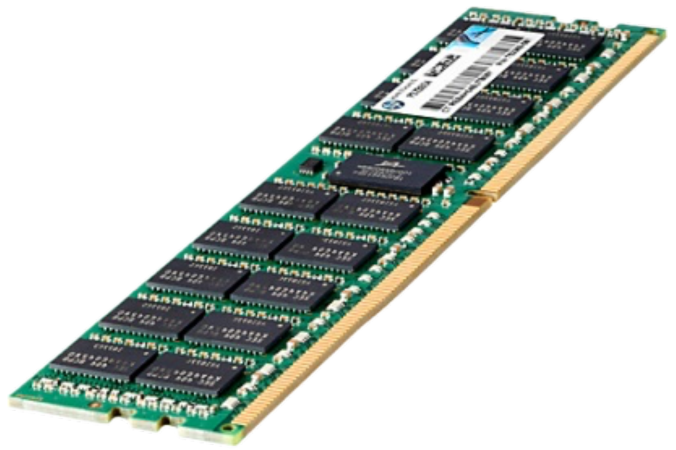


Cray-2 18MB stacked into 8 board unit



Magnetic Core
1 microsecond access time

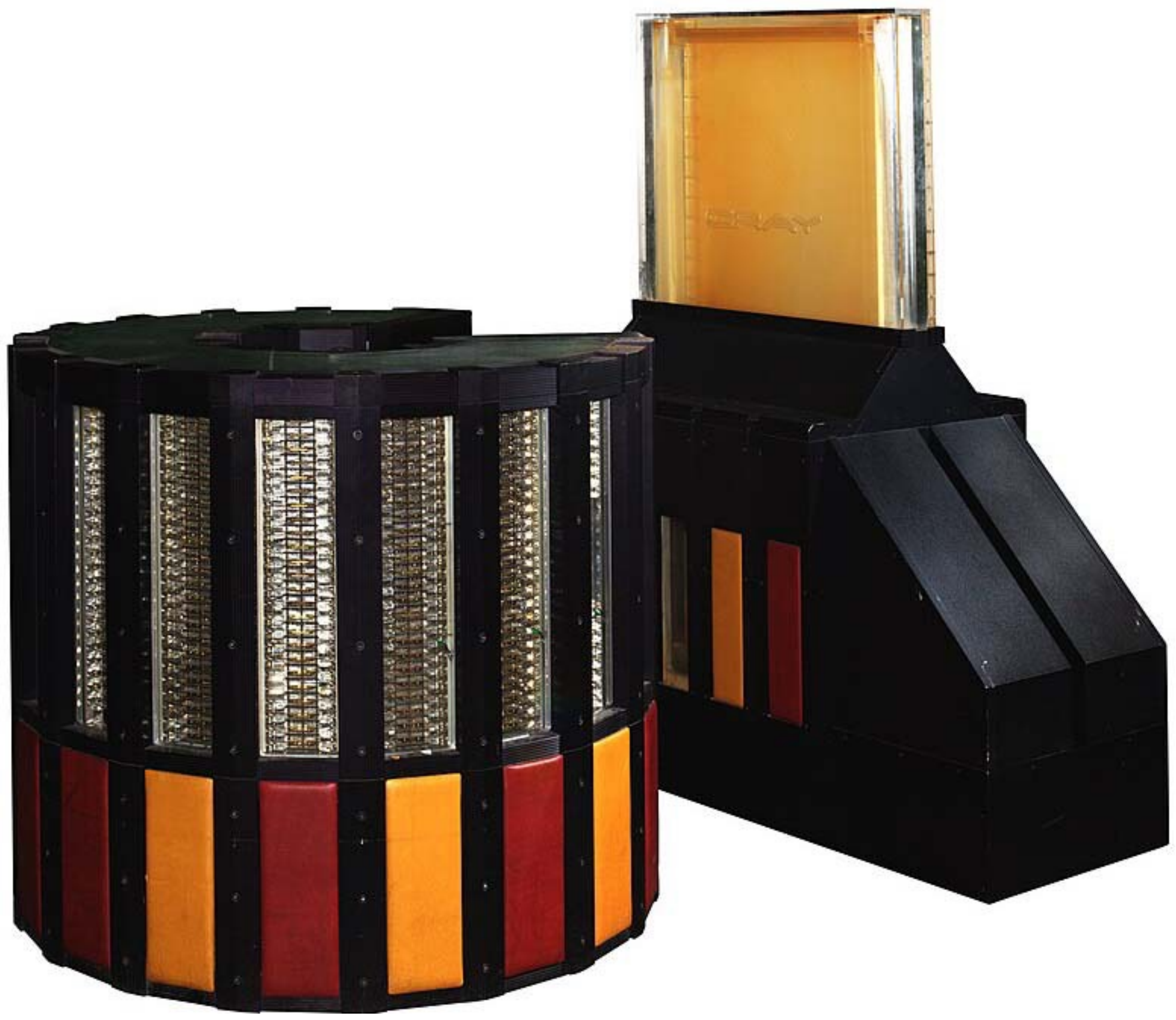
Cray-2 used Fluorinet for immersion cooling



- Modern desktop computer: 50 ns DDR5 DRAM access
- L1 Cache: 1 to 4 ns
- Modern mainframe: 100+ ns

Current memory systems use multiple levels of cache.

Current memory cell size ~100–150 nm^2.



Cray-2 (1985) - First to reach 1.9 GFLOPS

- Architecture: Four vector processors
- Clock speed: 244 MHz
- Memory: 256 MB main memory
- Cache: 16KB instruction buffer per processor
- Physical size: Smaller than Cray-1, about 4 feet high cylinder
- Power: 195 kW
- Cooling: Unique liquid fluorinert immersion cooling system

Intel Red (1997) - First to reach 1 TFLOPS

- Architecture: Massively parallel, x86-based
- Processors: 9,298 Pentium Pro processors at 200 MHz
- Memory: 594 GB distributed memory
- Cache: 256KB L2 cache per processor
- Physical size: 104 cabinets covering 2,500 square feet
- Power: 850 kW
- Cooling: Air-cooled with chilled water heat exchangers
- First supercomputer to break the teraFLOPS barrier

IBM Roadrunner (2008) - First to reach 1 PFLOPS (1.105 petaFLOPS)

- Architecture: Hybrid CPU/Cell processor design

- Processors: 12,960 IBM PowerXCell 8i CPUs and 6,480 AMD Opteron dual-core processors

- Memory: 98 TB RAM + Cache memory

- Cache: L1, L2, and L3 cache hierarchy

- Physical size: 278 refrigerator-sized racks

- Power: 2.35 MW

- Cooling: Water-cooled system

- First hybrid supercomputer to use hardware accelerators extensively

Tianhe-1A (2010) - Notable Early GPU Implementation (China)

- Architecture: Breakthrough CPU/GPU hybrid design
- Processors: 14,336 Intel Xeon X5670 CPUs and 7,168 NVIDIA Tesla M2050 GPUs
- Memory: 262 TB RAM + GPU +Disk memory
- Cache: Multi-level with CPU and GPU memory hierarchies
- Physical size: 103 cabinets
- Power: 4.04 MW
- Cooling: Liquid cooling system

- Achieved 2.57 PFLOPS, pioneering large-scale GPU use in supercomputing

Frontier (2022) - First to reach 1 EFLOPS (1.102 exaFLOPS) (cost: ~$1+ billion)

- Architecture: AMD EPYC CPUs with AMD Instinct MI250X GPUs
- Processors: 9,400+ AMD EPYC CPUs and 37,000+ AMD Instinct GPUs
- Memory: 8,800+ TB total (RAM + GPU + Local SSD + Cluster) system memory
- Cache: Complex hierarchy with High Bandwidth Memory (HBM)
- Physical size: 74 cabinets covering 7,300 square feet
- Power: 21-29 MW at peak
- Cooling: Advanced liquid cooling system
- First supercomputer to break the exaFLOPS barrier
- Currently 2nd Fastest Super Computer to LLNL's El Capitan (1.7 EFLOPS)

nVidia DGX SuperPOD - A scalable GPU/TPU AI system - Up to 63 exaFLOPS

- Architecture: Scalable GPU-based system - DGX Nodes (70-140)
- Processors: AMD EPYC (64 Cores + H100 GPU/TPU)
- Memory: 70-280 TB
- Power: 1MW to 2 MW
- Physical size: 70-140 Racks, 2000 to 4100 sq ft
- Cooling: Liquid with advanced heat exchangers
- Cost: $50 million to $100 million (a bargain)
- First released 2021 - Latest in 2024
- Helped enable advanced AI Chat models

High End Tower Gaming PC (Just for comparison)

- CPU: Intel i9-1490 - 8 performance, 16 efficiency
- Memory: 64 GB main, 4TB SSD Disk
- Power: 800 Watts
- Cooling: Liquid
- Estimated Speed: (CPU + GPU) 90 TeraFLOPS
- Cost: $4,000+
- Use: AI workloads, Gaming, Video Editing

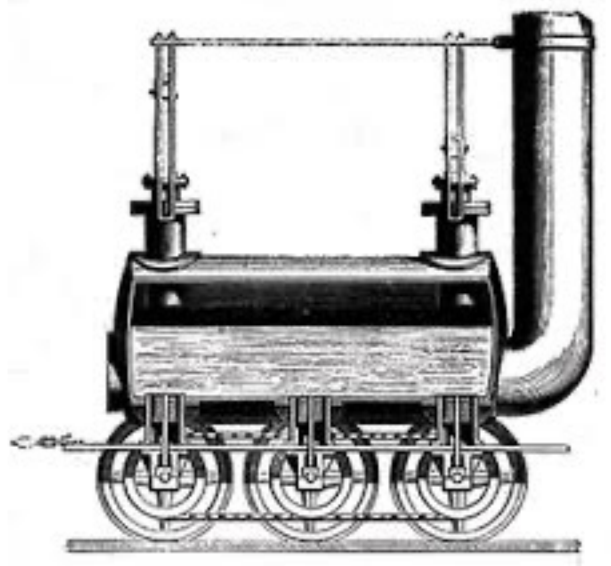
How Much Faster is Frontier than the Cray-1?

If a Frontier super computer can perform a calculation at 1.1 exaFLOPS in one second, how long would it take in years for the Cray-1 at 160 megaFLOPS to perform the same calculation?

Frontier (1.1 XFLOPS): 1 second (or 1mm)

Cray-1 (160 MFLOPS): 218 YEARS (or 6,875km)

About Denver to Brasillia



To put this in perspective, if we started this calculation on the Cray-1 in 1807 (around when the first steam locomotive was demonstrated), it would just be finishing now.

The Frontier is about 6.875 billion times faster than the Cray-1. This massive improvement in computing power over just a few decades demonstrates the remarkable progress in supercomputer technology.

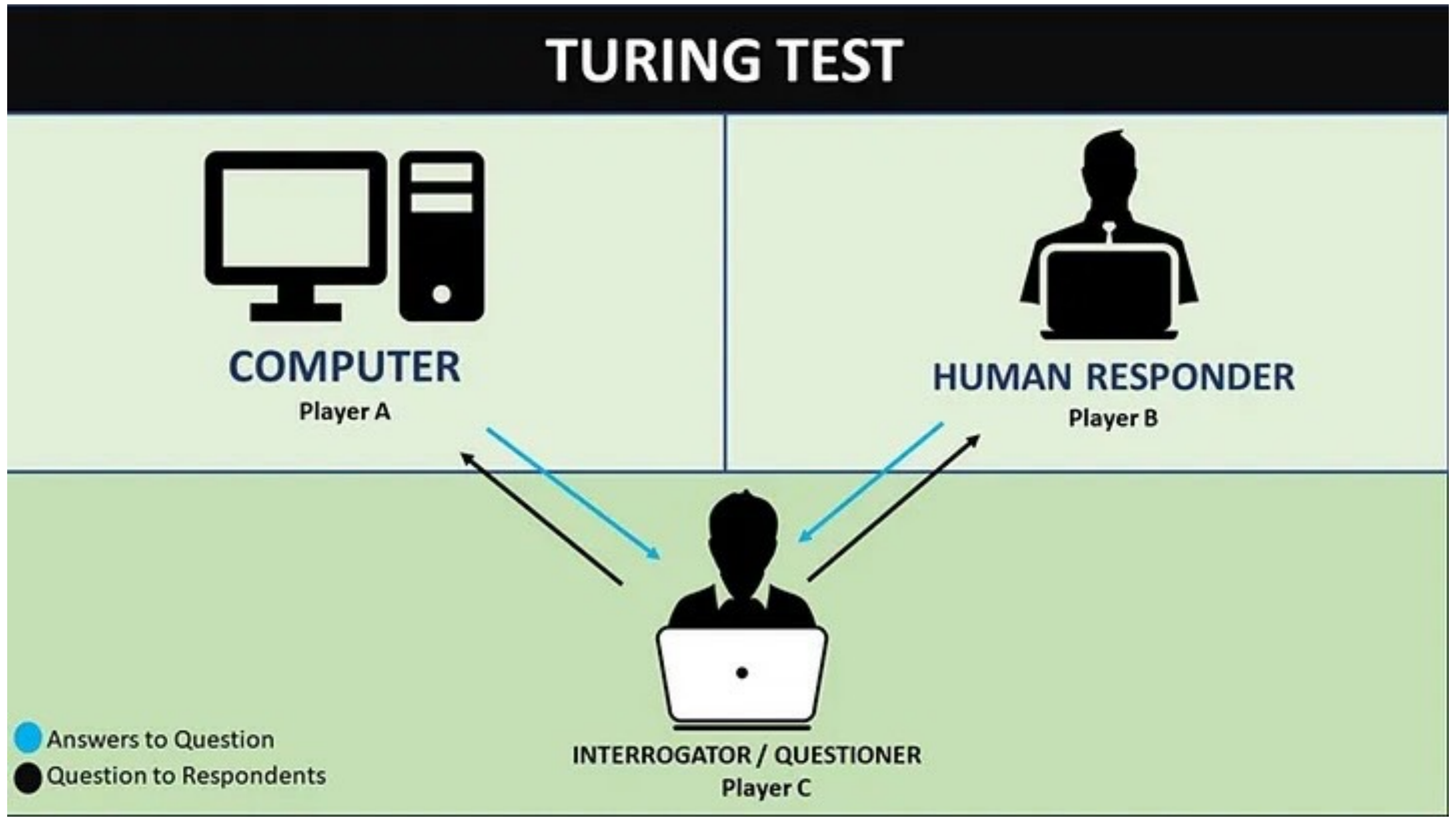
The difference in scale between 1.1 XFLOPS and 160 MFLOPS
is too large to show graphically on a screen! That arrow goes far!

Frontier 1mm → … → Cray-1 6,875 km

AI History

The Evolution of AI: From Expert Systems to Modern Deep Learning

The Turing Test*: Can a human evaluator distinguish between text conversations with a hidden human and a hidden machine?*



The *Turing Test* is named after the creator Alan Turing, about 1950.

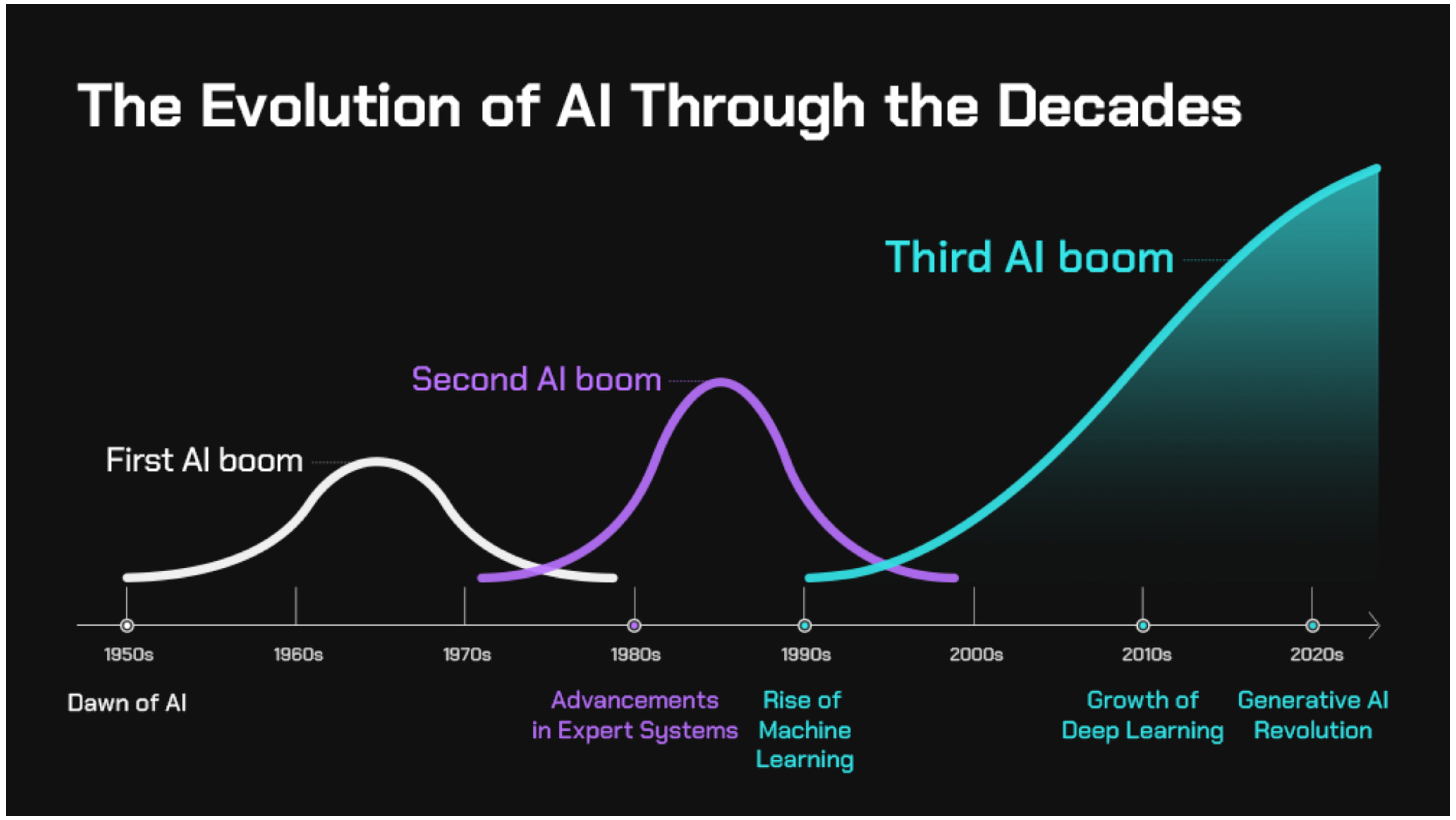
AI - The Beginnings

Between the 1950s through the 1980s

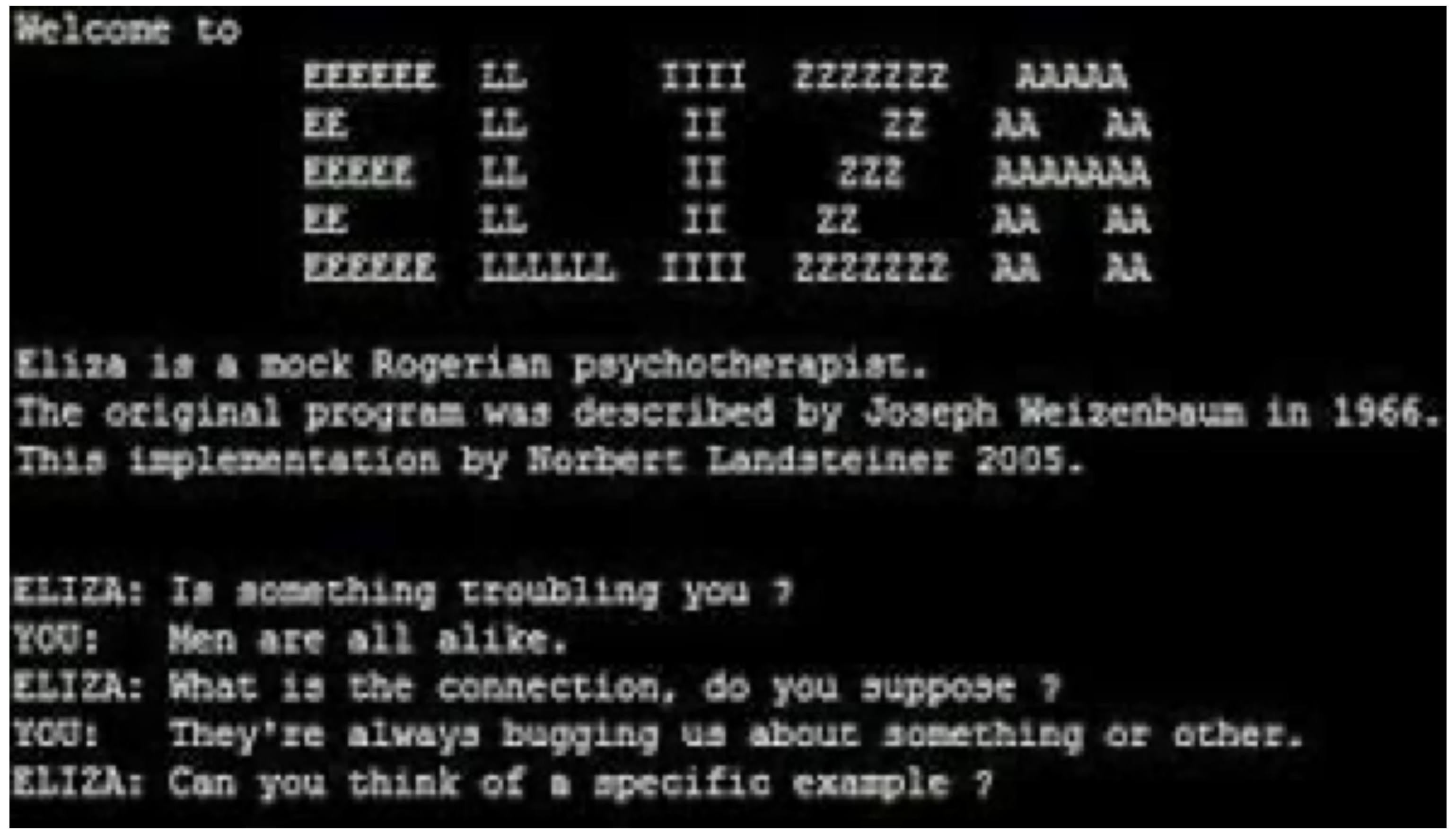
The term "artificial intelligence" (AI) was coined in 1956 by computer scientist John McCarthy at Dartmouth College.

Researchers understood what AI could be, but not much practical success.

A general consensus was “real AI” might not be possible.



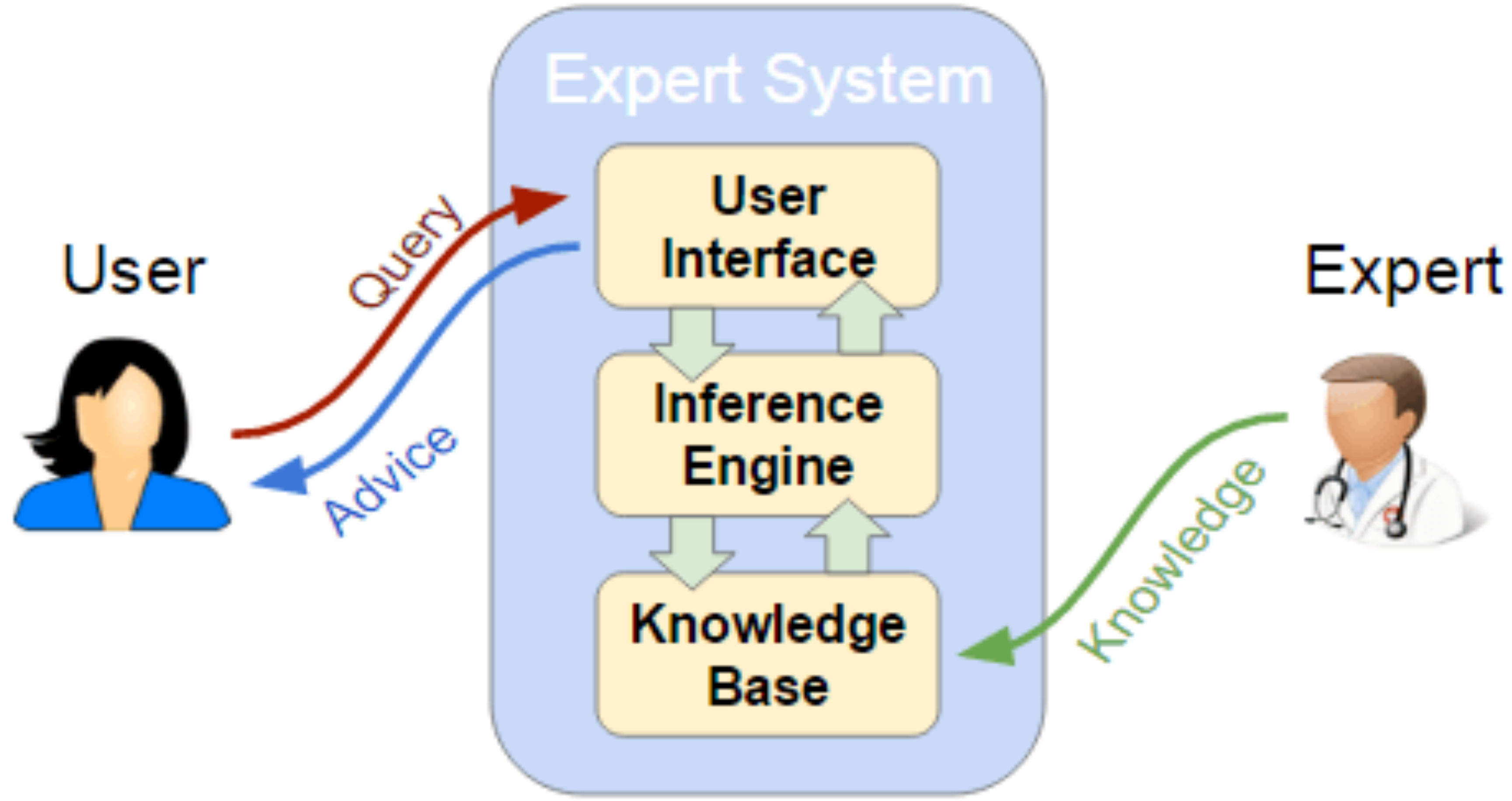
Anybody remember ELIZA?



“ELIZA—A Computer Program For the Study of Natural Language Communication Between Man and Machine” - Joseph Weizenbaum, 1966

The Era of Expert Systems (1980s)

- Rule-based systems encoding human expertise
- Knowledge bases + inference engines
- Narrow domain focus
- Explicit programming of decision logic



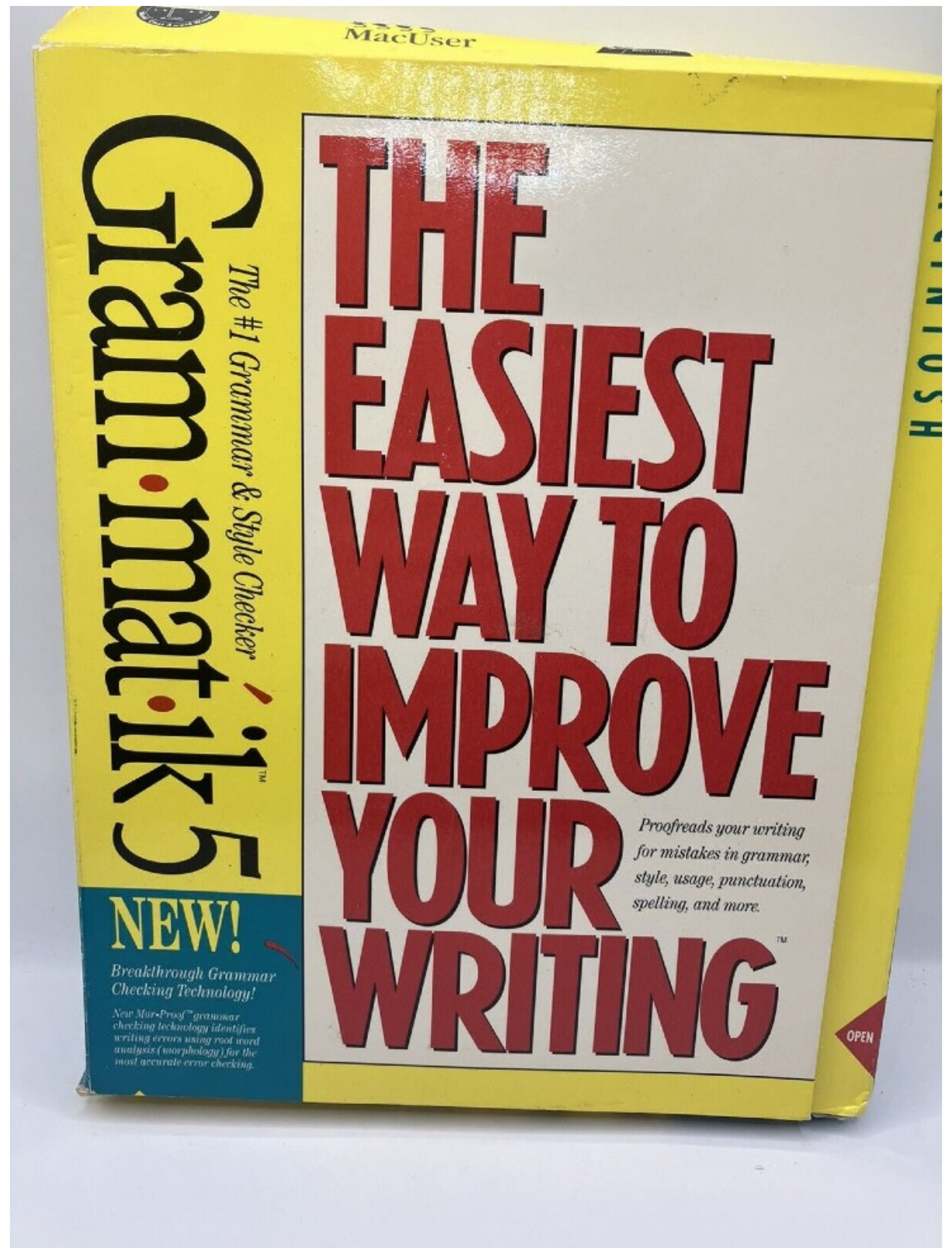
Notable Successes

- XCON/R1 (1980): Computer configuration
- PROSPECTOR: Mineral exploration
- DENDRAL: Chemical analysis
- Financial systems: Credit authorization, loan decisions

Early "Smart" Applications (1980s-1990s)

- **Spelling and Grammar Checking**

  - Rules + Dictionary
  - Evolution from simple lookups to pattern matching
  - Integration of linguistic rules and parsing
  - Part-of-speech tagging



- **Digital Camera Intelligence**

  - Autoexposure (1990s): Algorithmic
  - Autofocus: Signal processing approaches
  - Face Detection (2005-2007): Early pattern recognition
  - Progression from pure algorithms to ML techniques

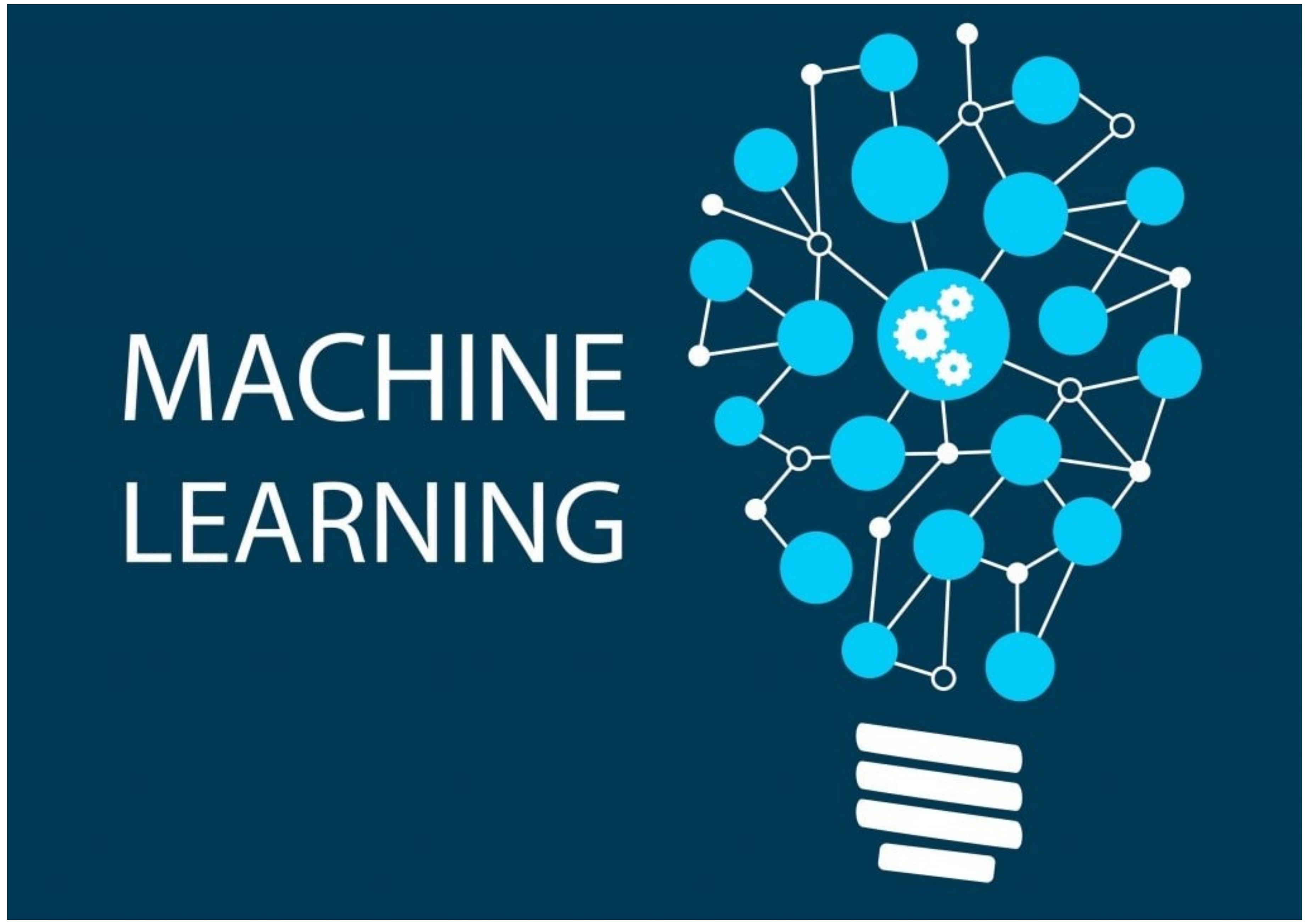
The Machine Learning Transition (1990s-2000s)

- Changes in Approach

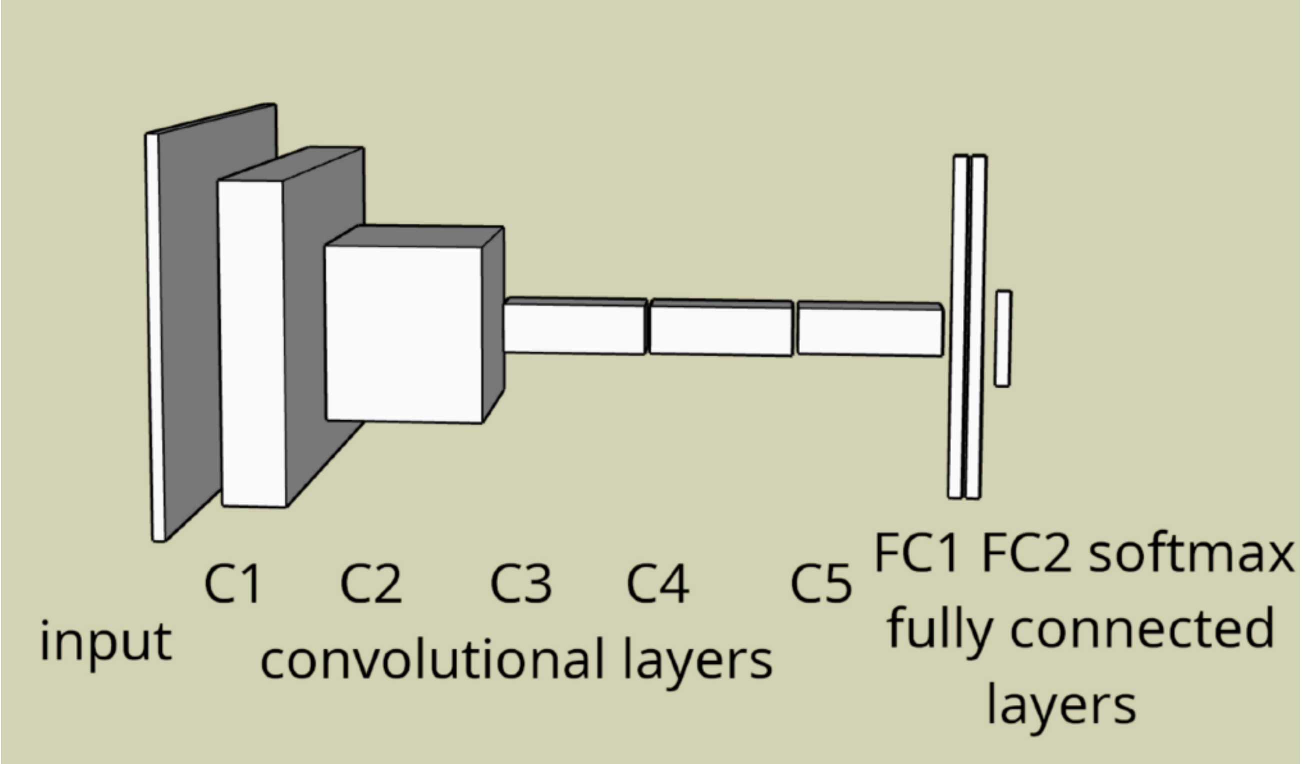
  - Learns rules by analyzing data
  - Integration of statistical methods
  - Rise of supervised and unsupervised learning
  - Limited by computational power and data availability

- Key Challenges

  - Small datasets
  - Limited computing power
  - Manual feature engineering
  - Lengthy training times



The Deep Learning Revolution (2012-2017)



- **First Breakthrough: 2012**

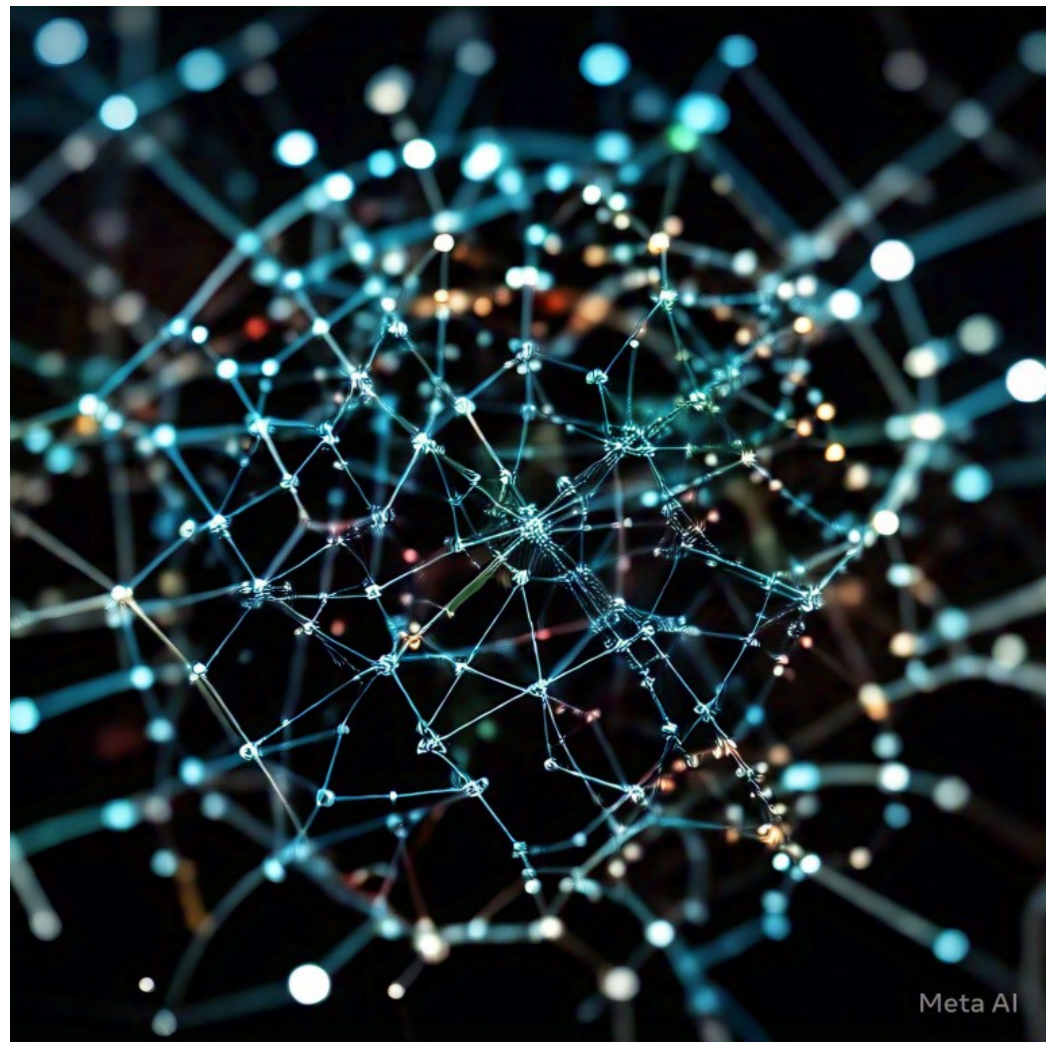
  - *AlexNet and ImageNet competition (Convolutional Networks)*
    - GPU acceleration enables practical deep learning
    - Convolutional Neural Networks prove viable *(Tensors)*
    - Marks shift from traditional Machine Learning to deep learning

- **Second Breakthrough: 2017**

  - *"Attention Is All You Need" paper*
    - Introduction of *Transformer* architecture
    - Enables efficient processing of sequential data
    - Foundation for modern language models

This page is the HARD one!

What is modern Generative LLM AI?

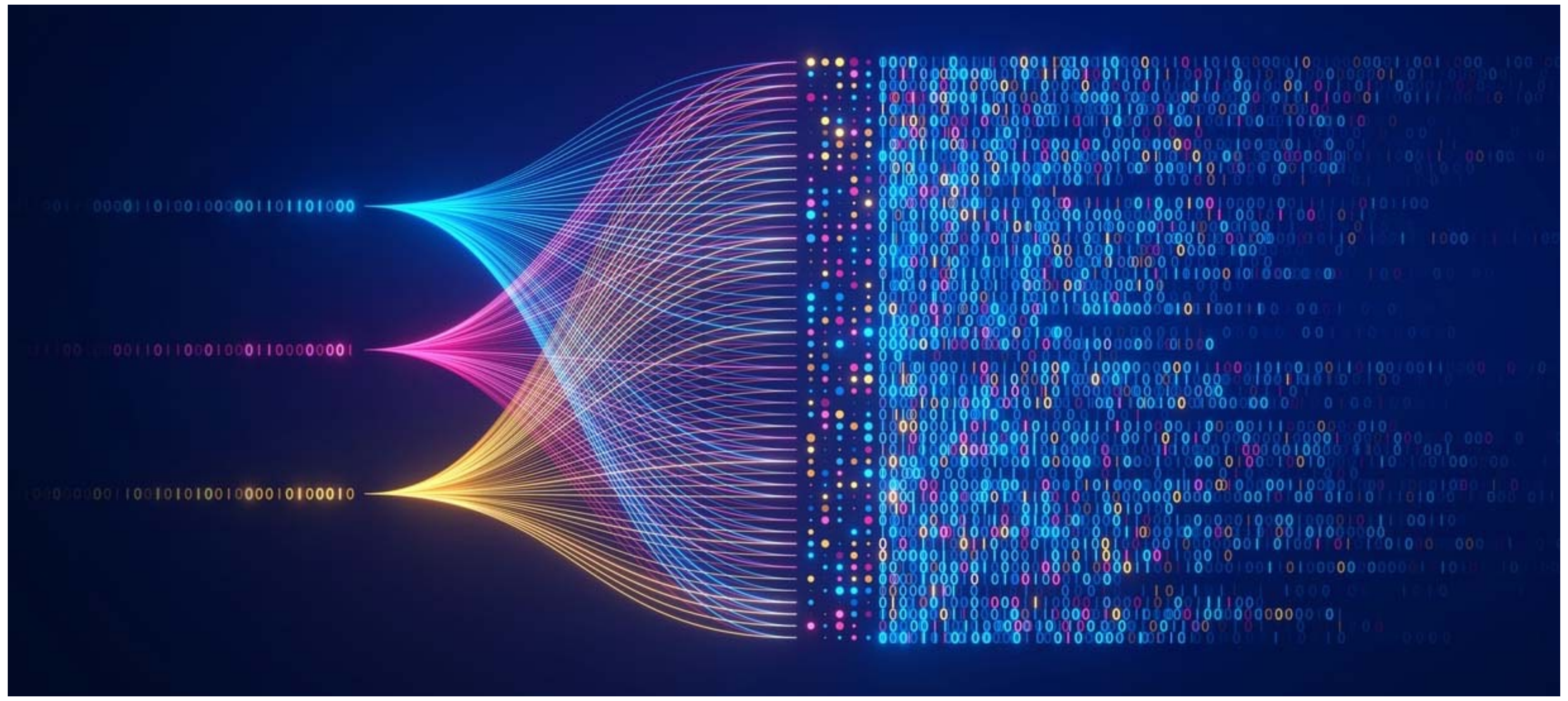


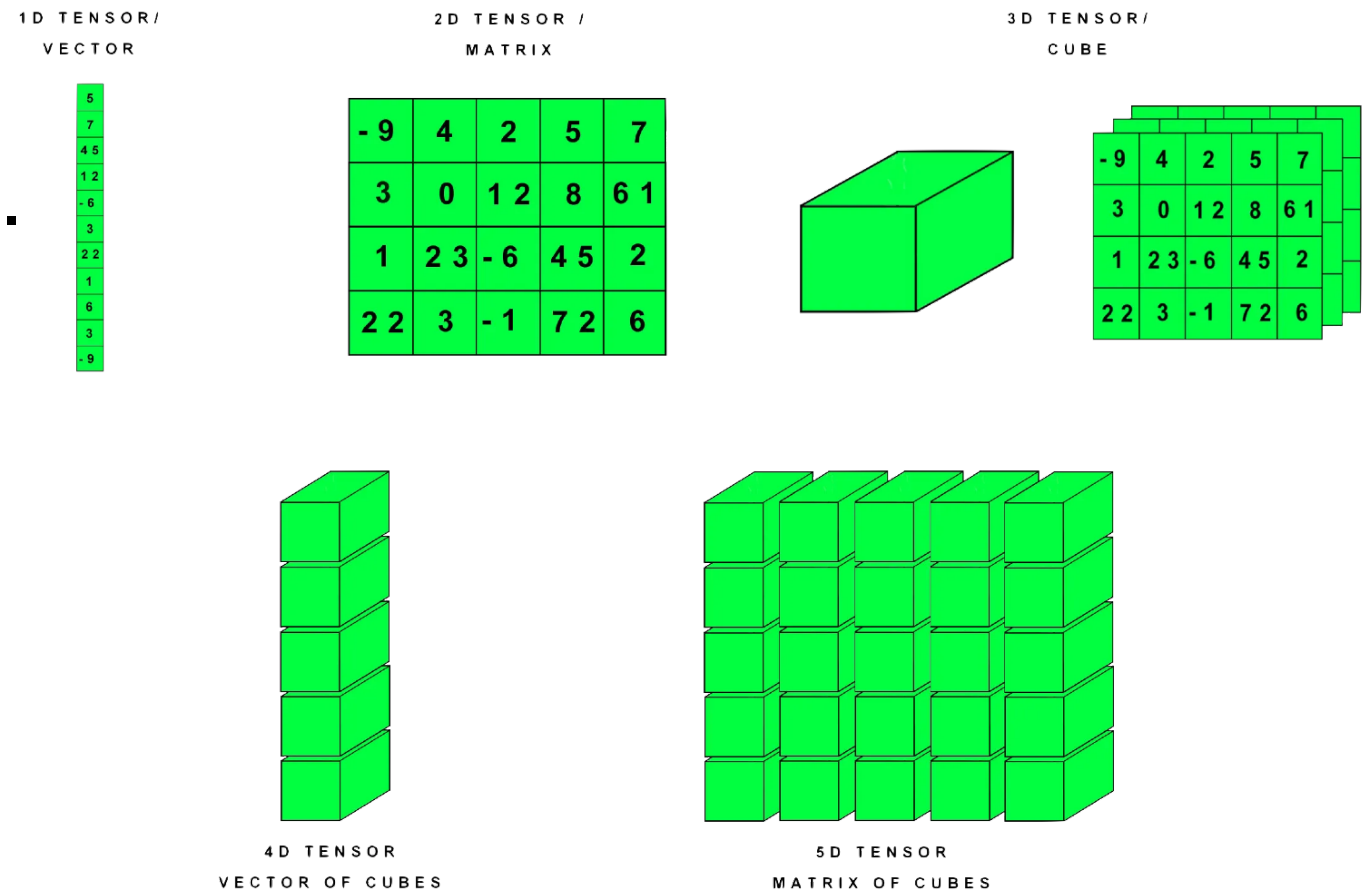
- AI systems process and learn from massive training datasets
- Large Language Models (LLMs) use the Transformer model
- Models generate novel content like text, code, and images
- Natural language enables human-like interaction
- Advanced reasoning capabilities handle complex problems
- Key limitations include potential errors and lack of true understanding
- Practical LLM AI Applications:
  - Answering questions on almost any topic
  - Content generation
  - Software development and coding
  - Medical diagnosis
  - Language Translation
  - Educational tutoring and personalized learning

This page is the HARD one, continued

What????

- Remember the big AI breakthroughs: *Neural Nets* and *Transformers*
- An LLM neural net is really a big *Tensor* with Billions or Trillions interconnected nodes.
- These nodes are mostly *probability* values generated using various formulas during training.
- With enough nodes, it turns out to be possible to generate answers using various very clever and mysterious algorithms using the probabilities to generate new information based on query input.
- Processing Billions or Trillions of nodes requires highly parallel computing, made possible with GPUs and TPUs.

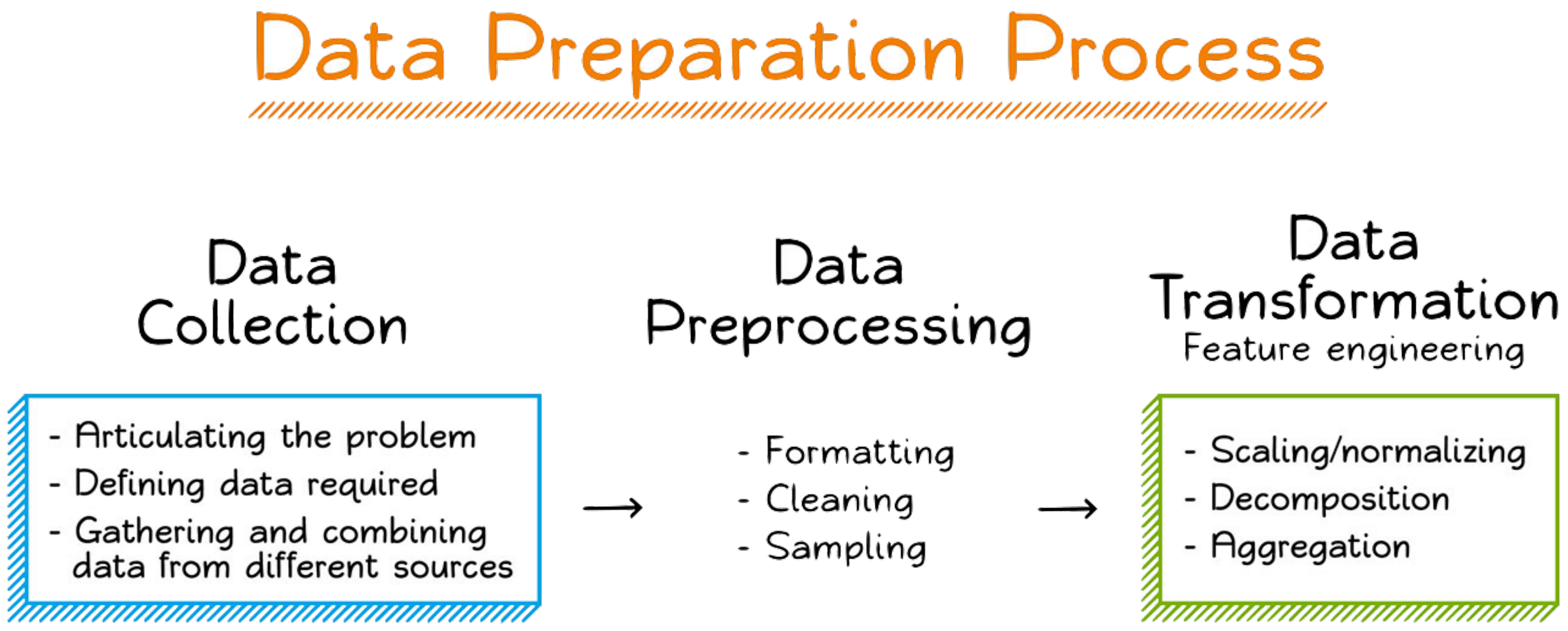




So, How is an AI Model Created?

Data Collection & Preparation

- Gather massive amounts of text data from books, websites, papers
- Clean and format the data (remove errors, standardize format)
- Break text into smaller pieces ("tokens") that the AI can process
- Create special markers to help AI understand context and structure



Training the Model

Phase 1: Initial Training

- Model starts with random values in its neural network
- Processes training data millions of times
- Compares its output to known correct answers
- Adjusts its internal values to improve accuracy
- Takes weeks/months on supercomputers

Note: Big LLMs such as ChatGPT require very high computational level.

Phase 2: Fine-Tuning

- Model is further trained on specific types of tasks
- Learns to follow instructions and maintain conversations
- Develops understanding of safe and helpful responses

Basic Description of AI Calculations

An AI model uses neural networks made of interconnected tensors - essentially multi-dimensional arrays of numbers. Each connection between nodes has a 'weight' that determines its importance, and these weights are adjusted through mathematical calculations during training to help the model learn patterns. A couple of simplified calculations are shown:

Basic Node Output: $output = \tanh(w_1x_1 + w_2x_2 + w_3x_3 + b)$

Numbers to probability: $probability\_i = e^\wedge z\_i \,/\, (e^\wedge z_1 + e^\wedge z_2 + ... + e^\wedge z\_n)$

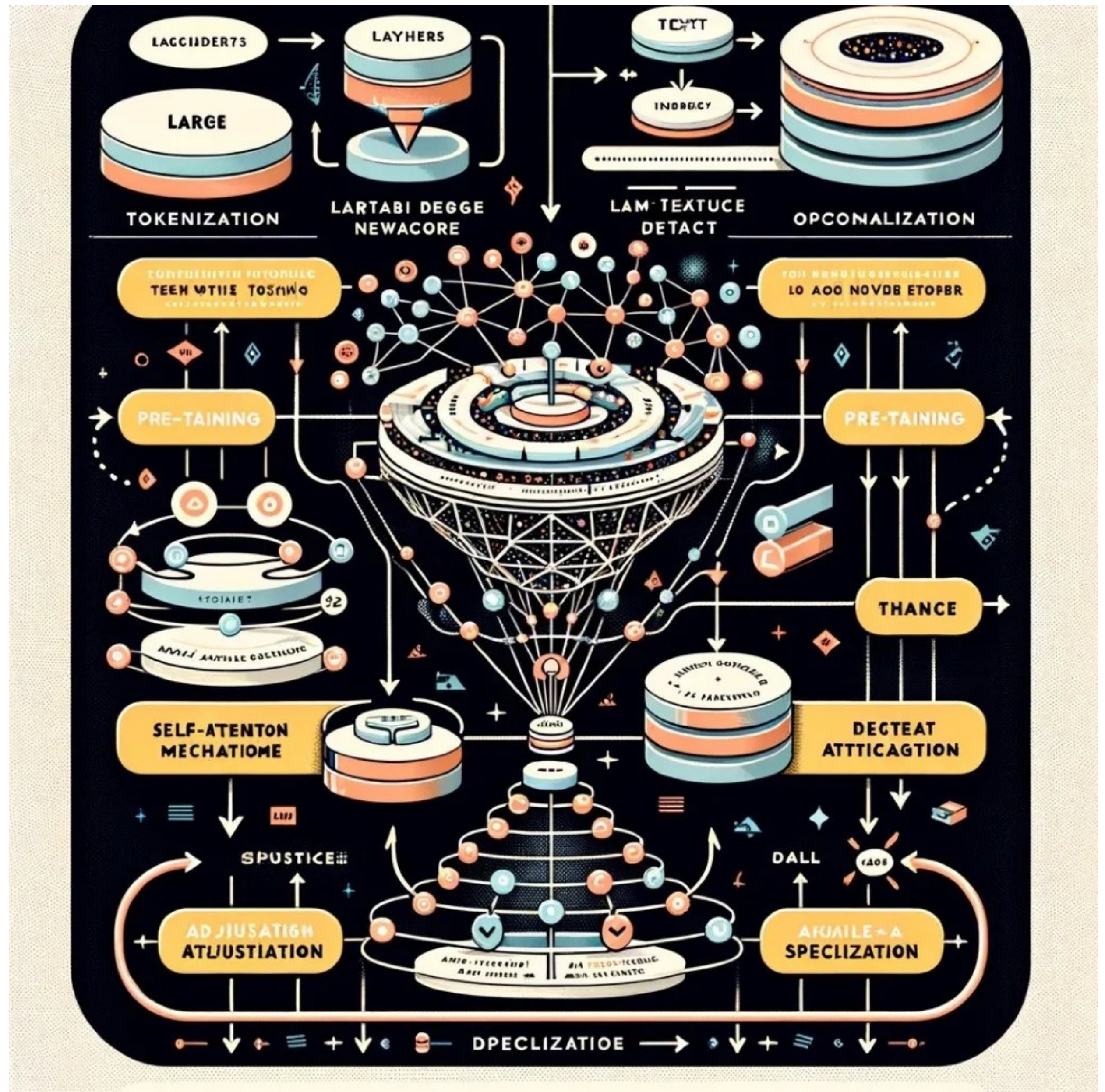
These formulas are relatively simple individually, but they become complex when:

- Applied millions of times per second
- Chained together in sequence|
- Used with multi-dimensional data (tensors)
- Combined with more advanced mathematical operations

The Scale of AI Training

- **Individual calculations are relatively simple mathematics**

  - Many parameter values in AI models are determined through probabilistic calculations and statistical methods during training.



- **The real power of an AI comes from:**

  - Processing *petabytes* of data (millions of gigabytes)

  - LLMs use *trillions* of parameters to enable advanced pattern recognition and response generation.

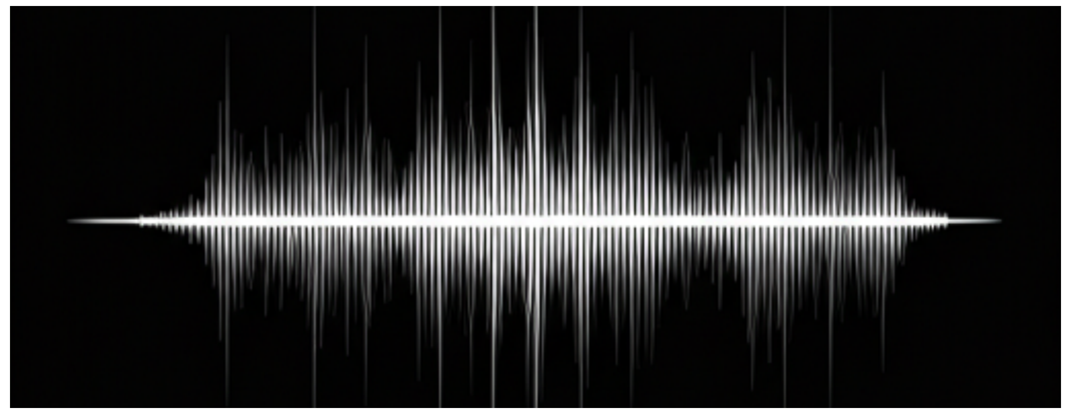
  - Performs *billions* of tensor calculations per second

  - Continuously adjusting *billions* of internal parameters during training

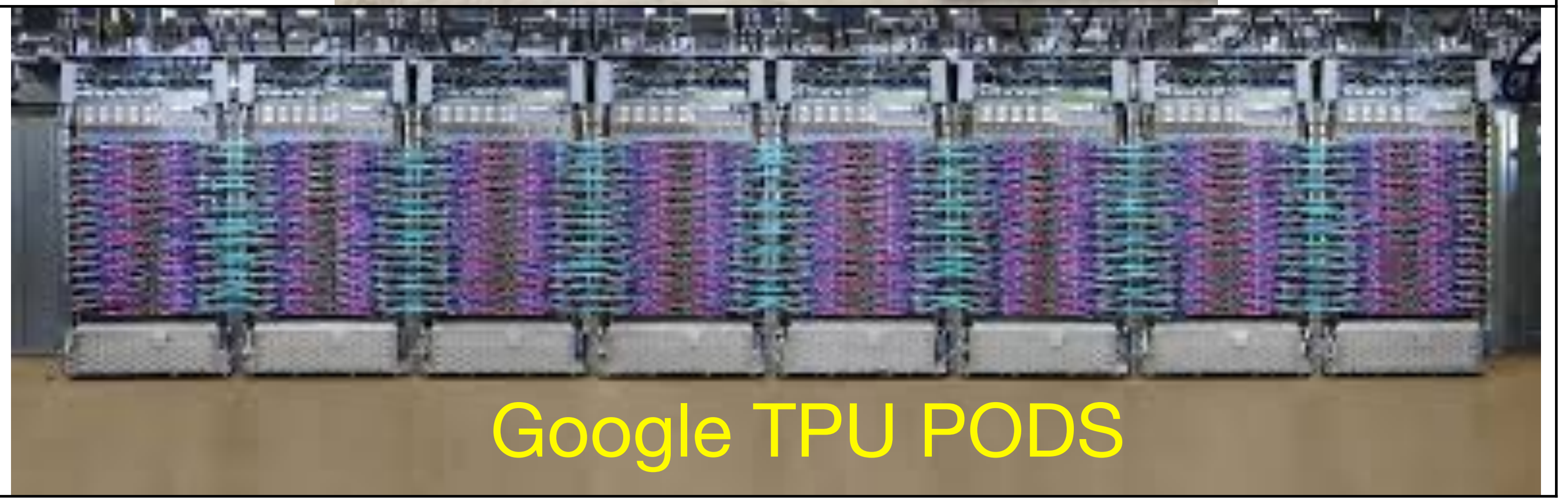
  - Running this process for weeks or months on powerful computers

What Computational Power does AI need?

Application Domain	**Example**	**Computation Level**
Audio Processing	Noise Removal	Low
	Music Generation	High
Image Processing	Image Denoising	Low to Moderate
Text Processing	Autocorrect	Low
	Chatbots (interaction)	High
3D Modeling	3D Object Generation	High
Multimodal	Text-to-Image	Very High
LLMs	Research Assistance	Very High
Scientific	Protein Folding Prediction	Very High



Low, Moderate, High, Very High Computational Level

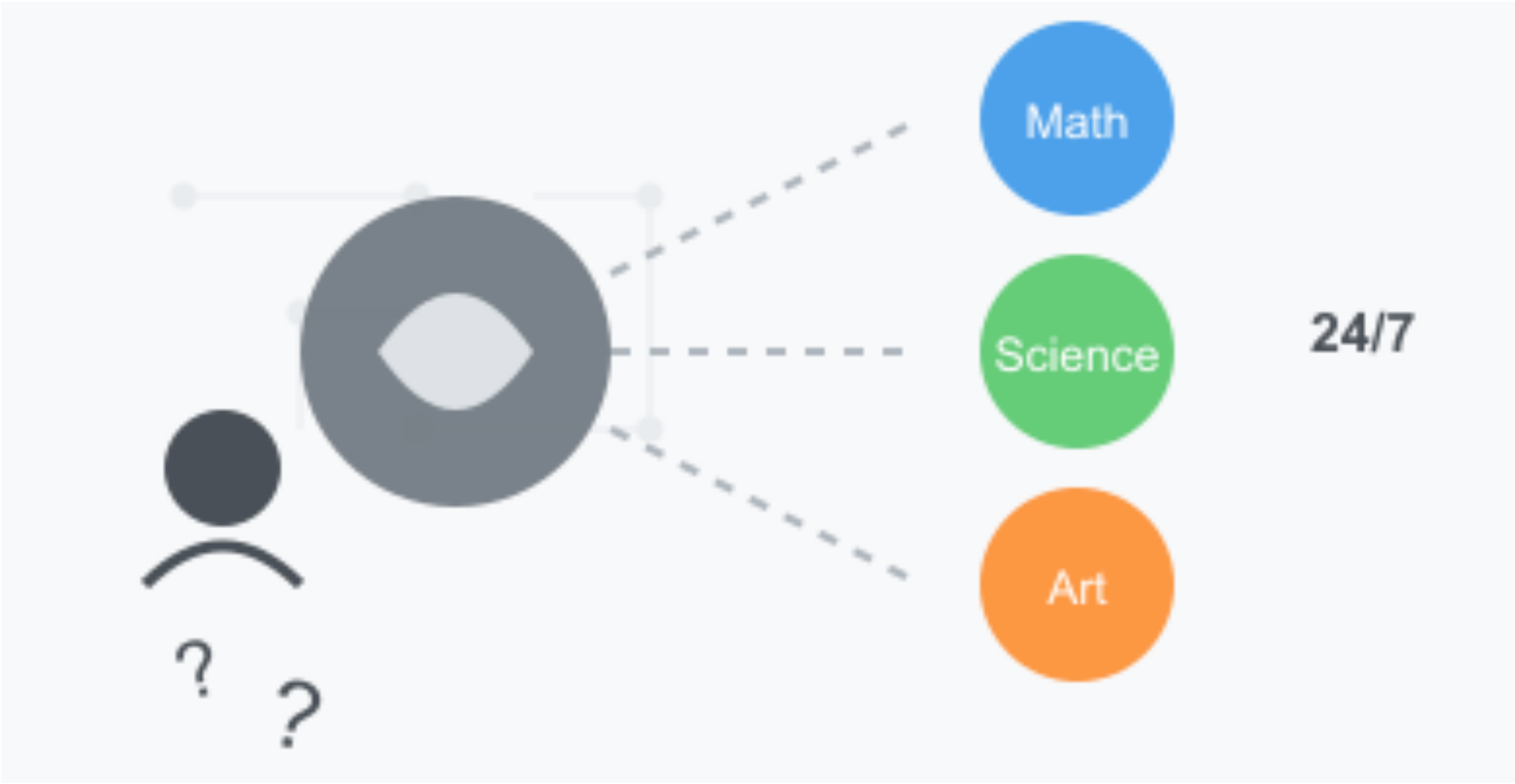
Computational Level	**Hardware Type**	**Example Systems**	
Low	High-End Desktop/Laptop	Intel Core i7, AMD Ryzen 5, NVIDIA GTX 1650. 20+ TFLOPS Smart Phone - 2+ TFLOPS	
Moderate	Workstation or Gaming PC	Ryzen 9, RTX 3060–4060, 32 GB RAM - 50+TFLOPS	
High	High-End Workstation/ ~~Server~~	AMD Threadripper, NVIDIA RTX 4090, or single A100. New nVidia Project Digits (2 PFLOPS) ($3000)	
Very High	Specialized AI Systems/Clusters	NVIDIA DGX SuperPOD Google TPU Pods multi-GPU/TPU systems (EFLOP scale)	

How can AI help you?

- **Use AI to learn about almost anything** *(To me, this is what it is all about!)*

  AI can be your personal expert teacher, available 24/7 to explain topics at your pace and answer follow-up questions until you truly understand the material. Simply start a conversation about what you want to learn, and treat the AI like a knowledgeable tutor who can break down complex ideas, provide examples, and even test your understanding. Remember that the key to effective learning is asking good questions and being specific about what you want to understand. Generated by Claude in response to section title

- **Answer General Questions**
- **Create Images & Diagrams**
- **Provide technical help**
  (for example, about RF circuits)
- **Design Projects**
- **Code (e.g. Raspberry Pi)**
- **Filter SSB Audio**



By Claude

By Microsoft Copilot

Some LLM AIs To Try

- ***ChatGPT***: Perhaps best known conversational AI chatbot released in November 2022. It can handle various tasks, including general questions, writing, coding, and solving complex math equations. Limited free service. (GPT = Generative Pre-trained Transformer)

- ***Gemini***: Google's conversational AI chatbot that sources its answers from the web, provides footnotes, and generates images within its chatbot. Liberal free service. There is a developer version called ***Google AI Studio*** with features (which can be ignored) to build alternate . This version lets you switch between models, and has fewer usage limits.

- ***Claude***: Anthropic's GPT focused on producing safer and more aligned interactions, with an emphasis on ethical AI usage. Free and Pro service.

Answers limited to last training date to increase safety.

- ***Copilot***: Microsoft's AI-improved chatbot offers capabilities similar to ChatGPT. Seems to have fairly generous free answers, but Pro available.

- ***Meta AI***: Developed by Meta. Gives nice answers. Liberal free service. Easy image generation.

- ***Perplexity AI***: A free AI chatbot. Live Answers. Provides sources. Interface tries to be modern and user friendly. But: Advertising supported - shows ads.

ChatGPT *Gemini* *Claude* *Copilot* *Meta AI* *Perplexity*

Most have "Pro" plans, commonly $20/mo.

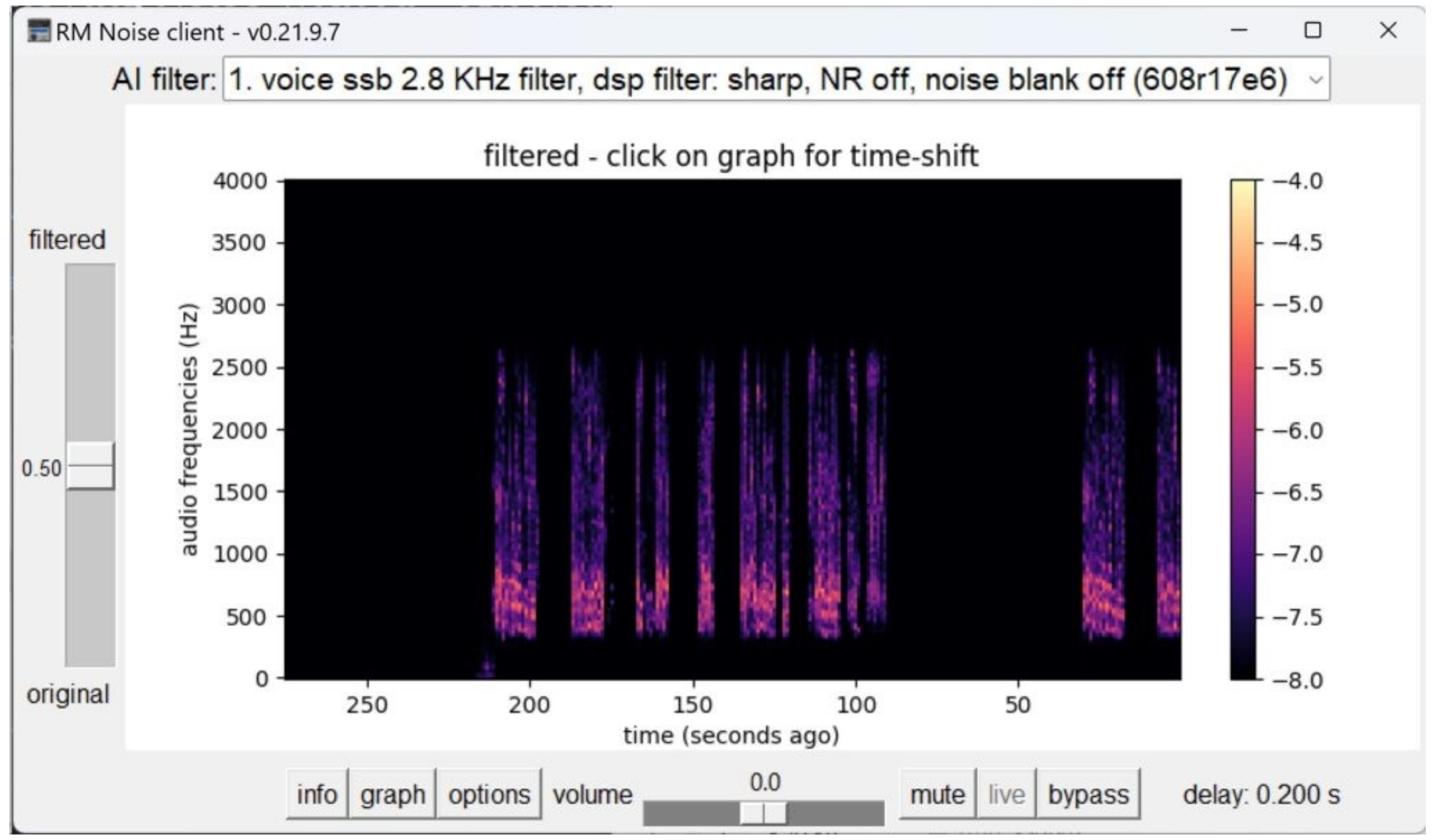
Hints for Interacting with an LLM AI

- Explain your end goal and use case for more relevant solutions
- Explain what you want. For example: “I’m creating a presentation for my amateur radio club about AI. I’d like to discuss the history of desktop computers. This presentation should be about 10 minutes long.”
- Break complex requests into smaller steps rather than asking everything at once
- State formatting and style requirements upfront
- Share your expertise level so it can explain at the right technical depth
- Be specific - "Help me program my Baofeng UV-5R to monitor local repeater frequencies" is better than "Help with my radio”
- For code, provide sample inputs/outputs when relevant - show example data and desired results
- When modifying previous work, specify exactly what needs changing
- Give specific feedback if my response isn't quite right - e.g. “Make it more technical”
- Reference previous prompts and answers, and specify refinements. Repeat as necessary.
- **<u>Beware of hallucinations!</u>** You can ask the AI to explain hallucinations for what one is. If you see a suspect answer is wrong, tell the AI.

Useful AI tools for Amateur Radio

RM Noise

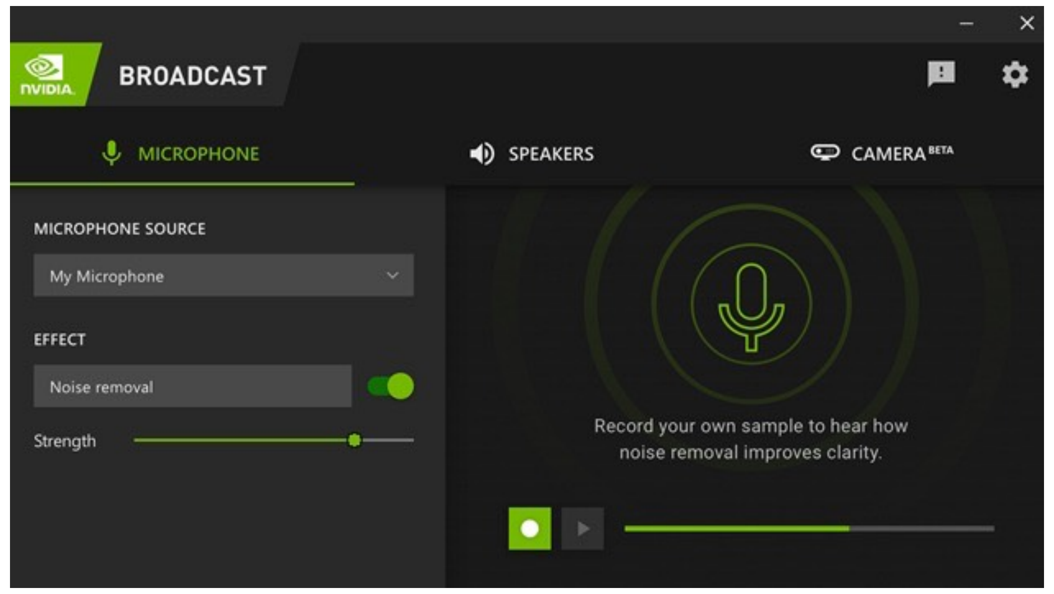
- The RM Noise project uses AI with the goal of removing noise from SSB voice or CW radio signals.
- The AI has been trained using noise recordings which are often unique to each setup and location.
- The client program sends the radio's noisy output to the AI servers, and the servers remove the noise in real-time and return the audio back to the client for listening.
- User can submit SSB recordings to help train the app.
- Has several servers world-wide
- Windows 11 only.
  RM Noise Website: https://ournetplace.com/rm-noise
- Only specific AI App for Amateur Radio I could find.



Alternate choice:

nVidia Broadcast App

Uses features of RTX video cards



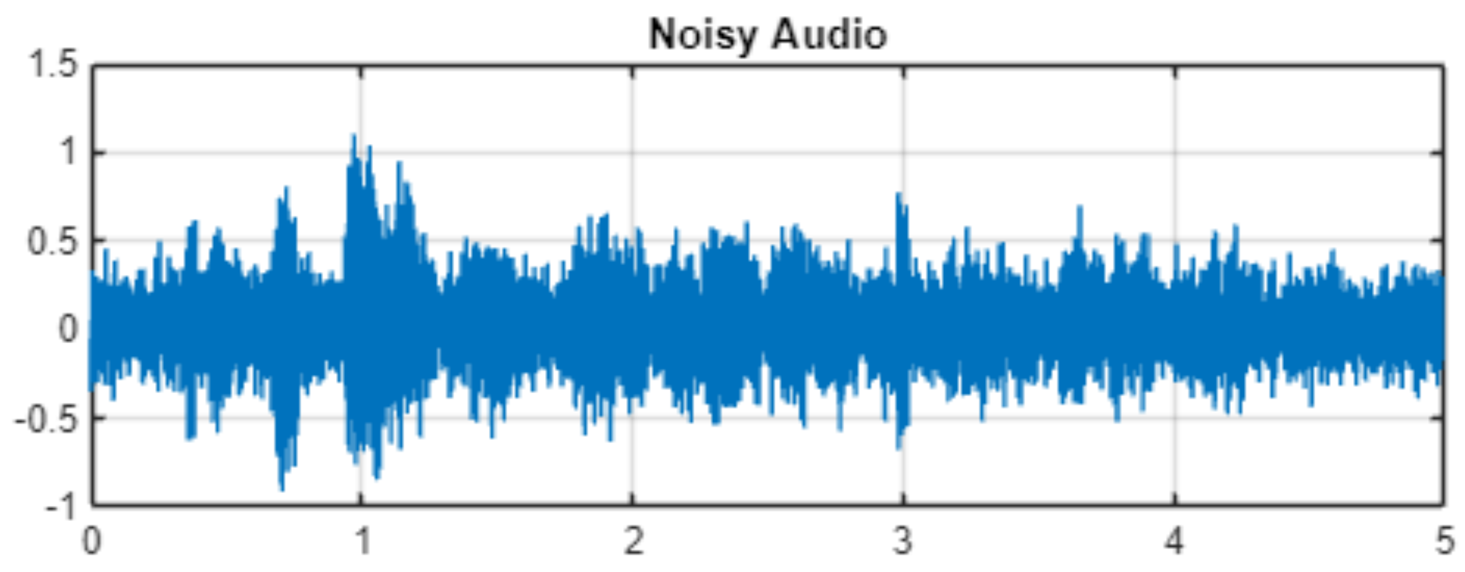
DSP vs AI for Noise Reduction

DSP

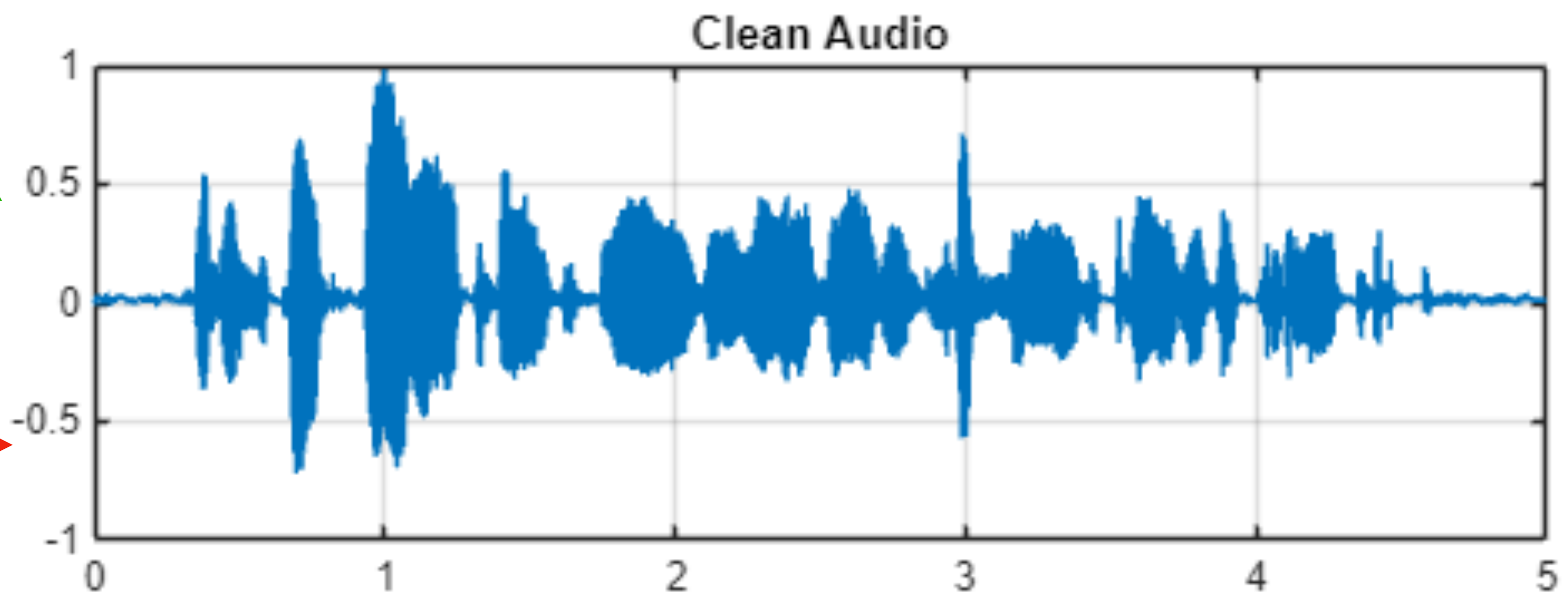
- Real-time processing of signals - fast
- Optimized for specific signal processing math operations
- Precise timing and deterministic performance
- Fixed performance, not as good as AI
- Locked in to chip
- Lower power consumption

AI

- Can have Latency issues
- Adaptive Learning - re-training and can get better
- Can process more than frequency components - context awareness
- Non-linear processing: more accurate
- Continuous improvement
- Higher power Consumption
- Requires AI capable computing



RM Noise



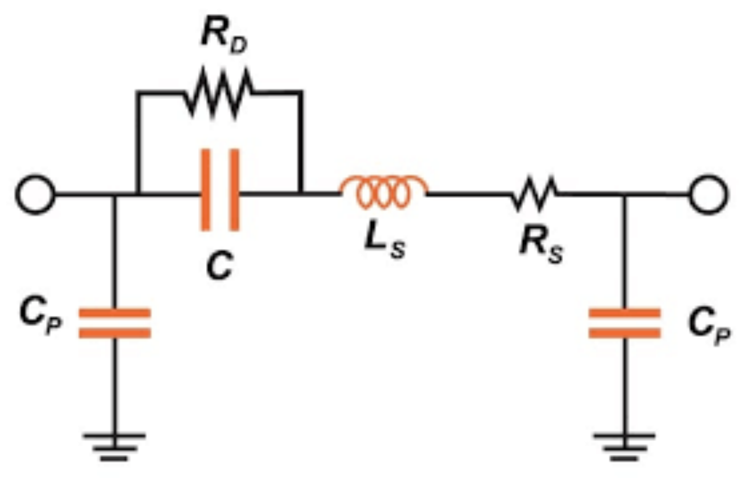
AIs Can Help Design RF Circuits

***Question asked:* Tell me about RF circuits suitable for use in various amateur radio projects?**

ChatGPT Answer: *Here are some common RF circuits and components that can be incorporated into projects, along with suitable off-the-shelf ICs and design considerations:*

(Edited to list of projects only, original answer provided more details not shown.)

1. Oscillators
2. Mixers
3. Amplifiers
4. Filters
5. Frequency Synthesizers
6. RF Detectors
7. Modulators/Demodulators
8. Antenna Matching and Tuning



Example Projects

* Simple CW Transmitter: Crystal oscillator, buffer amplifier, and low-pass filter.
* Direct Conversion Receiver: NE602-based mixer, audio amplifier, and band-pass filter.
* RF Signal Generator: ADF4351-based synthesizer with microcontroller control.
* SDR Front-End: Band-pass filter, LNA, and mixer for signal processing.

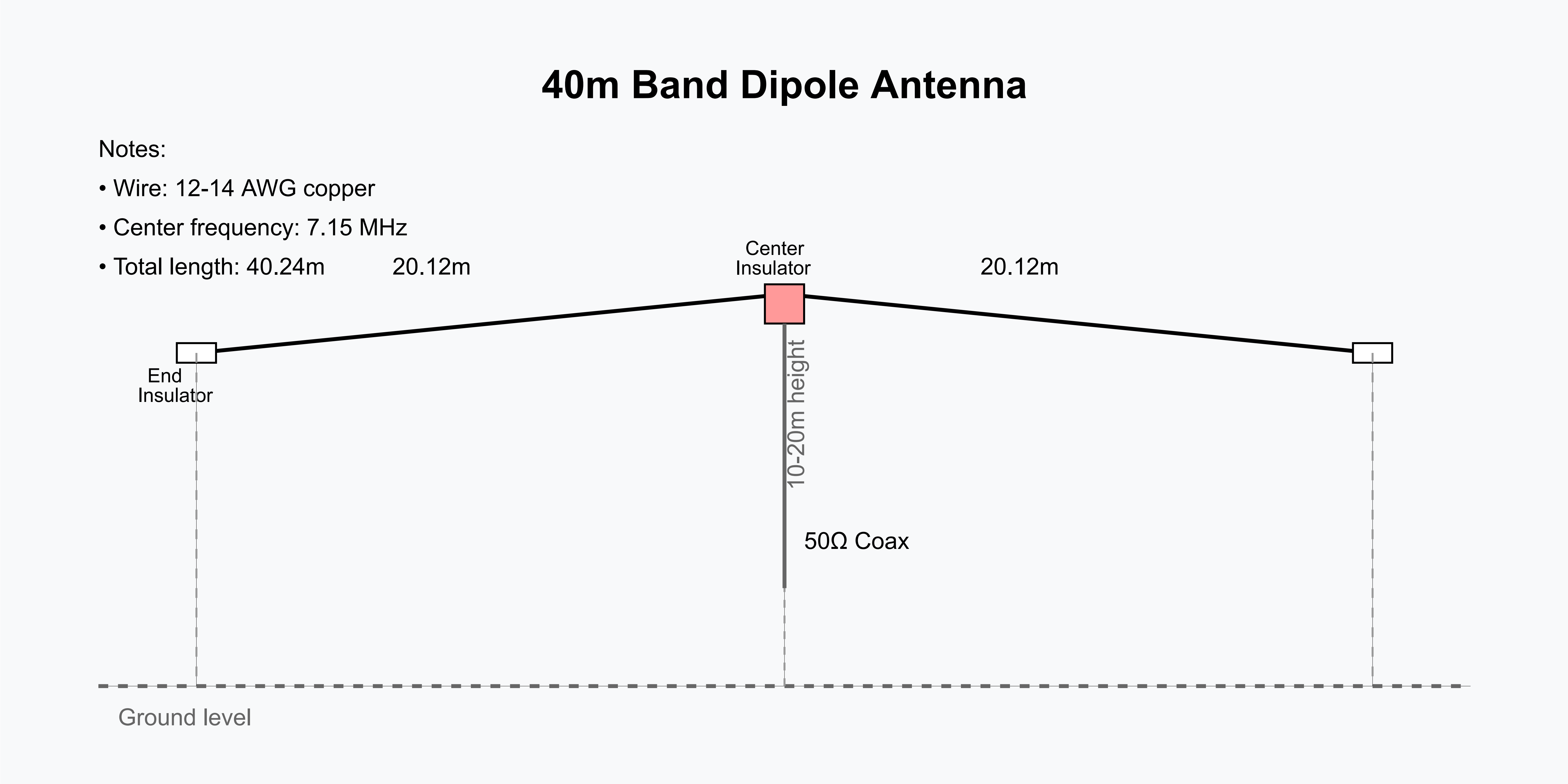
Let me know if you'd like detailed designs for any specific project or RF stage!

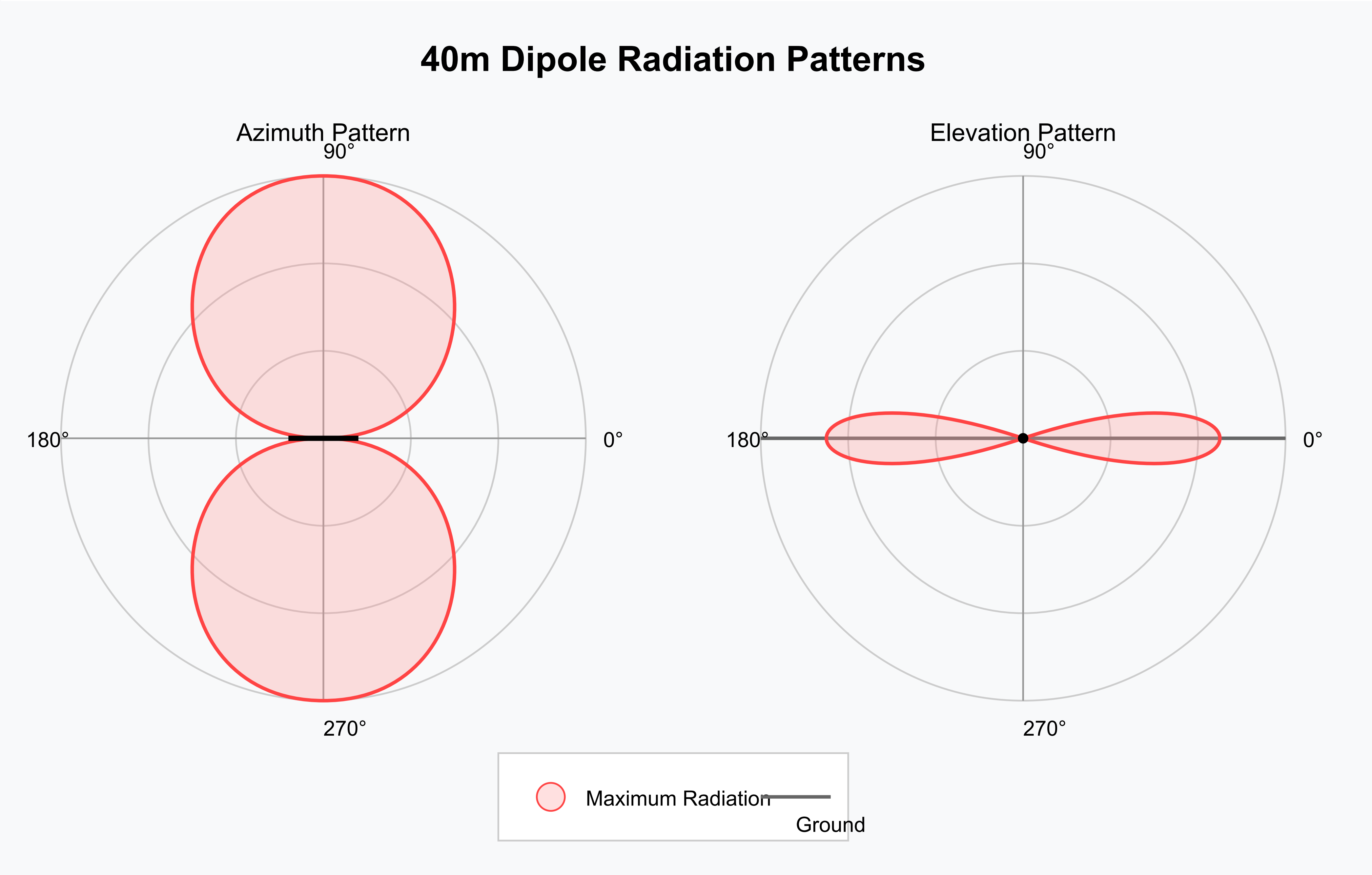
AIs Can Help Design Antennas

***Question asked:* Can you show me a design for a 40 meter band antenna for amateur radio?**

Claude Answer: I'll help you design a simple yet effective dipole antenna for the 40m band (7.0-7.3 MHz).

Claude provided a comprehensive list of materials, lengths, construction tips, and performance tips.
I then asked for a diagram for the antenna, and a radiation pattern diagram.





AIs Can Help Study FCC Exam Questions

Question: **Provide a sample question from the FCC Amateur Radio Extra license exam. Include an explanation.**

Meta AI Answer: Here's a sample question from the FCC Amateur Radio Extra license exam:

Question E7C09 - What is the primary advantage of using a pi-L network for impedance matching?

A - It provides a constant impedance match at all frequencies

B - It provides a high degree of selectivity

C - It allows for a wide range of impedance transformations

D - It is relatively simple to construct and adjust

Correct Answer - C

Explanation

A pi-L network is a type of impedance matching network that uses a combination of inductors and capacitors to match the impedance of a load to the impedance of a source. The primary advantage of using a pi-L network is that it allows for a wide range of impedance transformations.
This is because the pi-L network can be designed to match a wide range of load impedances to a fixed source impedance, or vice versa. This makes it a very versatile impedance matching network that can be used in a wide range of applications.
The other options are incorrect because:

- A: A pi-L network does not provide a constant impedance match at all frequencies. Its impedance matching characteristics vary with frequency.
- B: A pi-L network does not provide a high degree of selectivity. It is primarily used for impedance matching, not filtering.
- D: While a pi-L network can be relatively simple to construct and adjust, this is not its primary advantage.

Conclusions

- **Is AI Magic?** (well… pretty close)
- **Could Modern LLMs pass the Turing Test?**
  (Probably… AI generated images are still pretty obvious!)
- **Questions about this presentation? Ask an AI.**
  - Try asking Chat GPT about any topic covered here.
- **My personal take:**
  - I used AIs to help create many parts of this presentation.
  - I believed it would not happen in my lifetime.
  - This might be a seriously huge thing, like electricity, radio, automobiles, airplanes, telephones, TV, internet, cellphones.
  - My biggest hope is that it can improve how we learn.
  - AI is not self-aware, it does not have intent, nor judgement. It can be misused by people. It is, as always, the people with bad intent misusing the technology to create fake, false, or altered content that is the danger. Hopefully, AI will bring way more good than bad.