AI Fundamentals and Prompt Engineering

Mike Verdecanna
Aug 1, 2024

Agenda

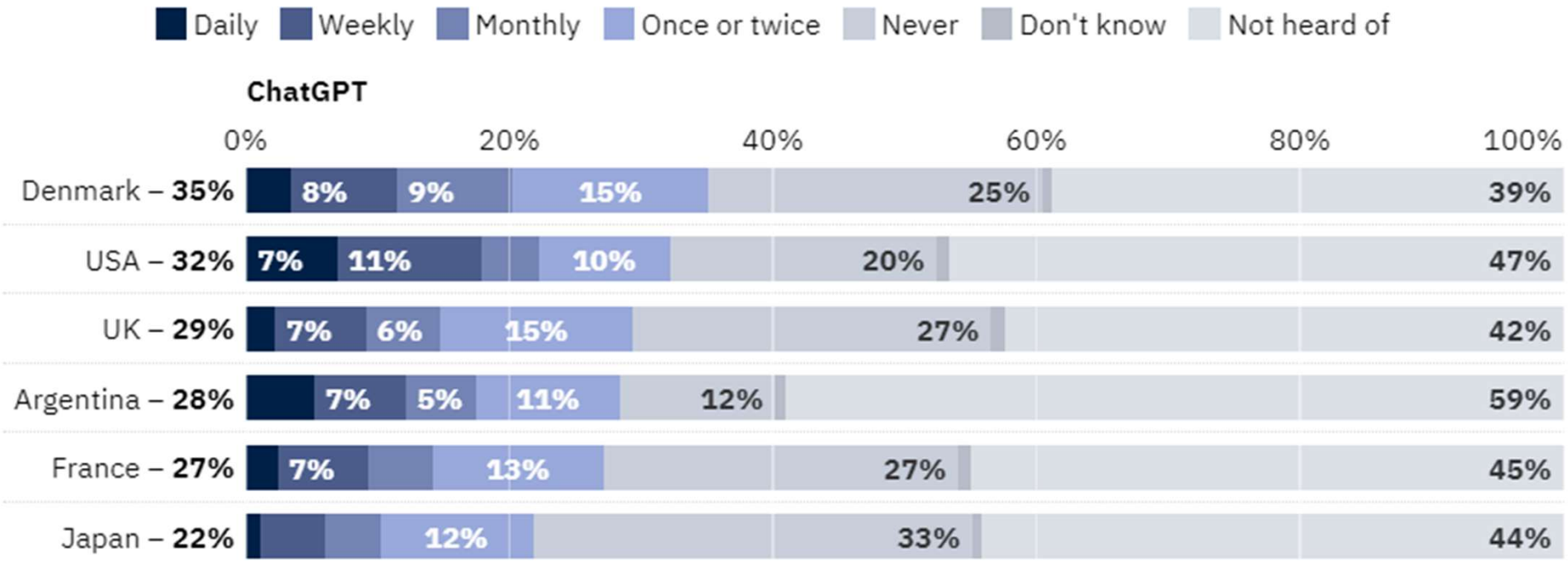
- Part 1: AI fundamentals
  - History
  - Machine Learning
  - Neural Networks - Deep Learning
  - Transformers
  - LLMs (Large Language Models)
- Part 2: Prompt Engineering
  - Basic Prompts
  - Prompt foundational structure
  - Advanced techniques: one shot, chain of thought, persona, self prompting, etc
- Part 3: Challenges and Future Trends
  - Challenges: Security, shadow prompting, bias, overfitting, etc
  - Trends: Multimodal, AI Agents, Edge AI, etc.

Audience Poll

How frequently do you use ChatGPT	Poll results
Daily	20%
Weekly	2%
Monthly	7%
Once or twice	10%
Never	32%
Don’t know	12%
Not heard of	17%

Reuters Online Survey May 2024

Sampled 2000 people each from 6 different countries



https://reutersinstitute.politics.ox.ac.uk/what-does-public-six-countries-think-generative-ai-news

What is AI

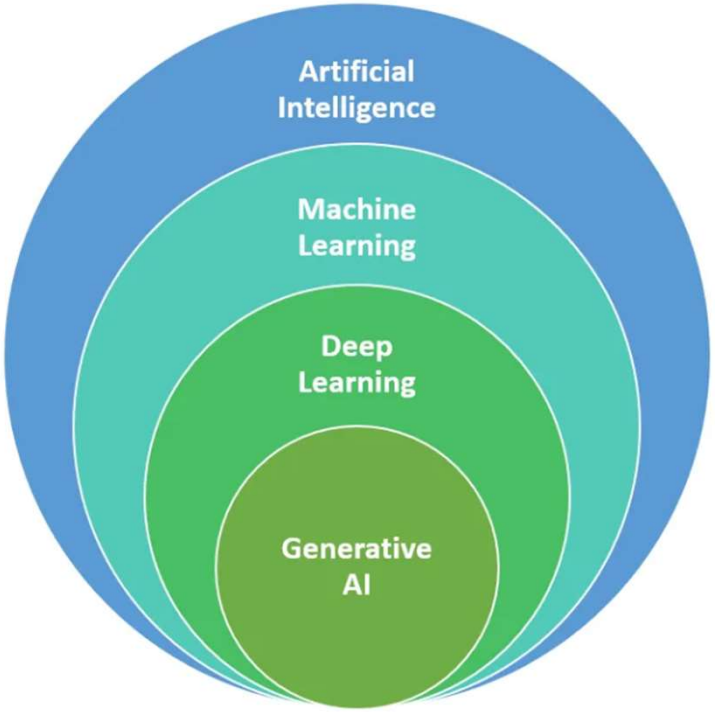
Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems, enabling them to perform tasks such as learning, reasoning, problem-solving, perception, and language understanding. They are designed to learn from data and improve performance over time.

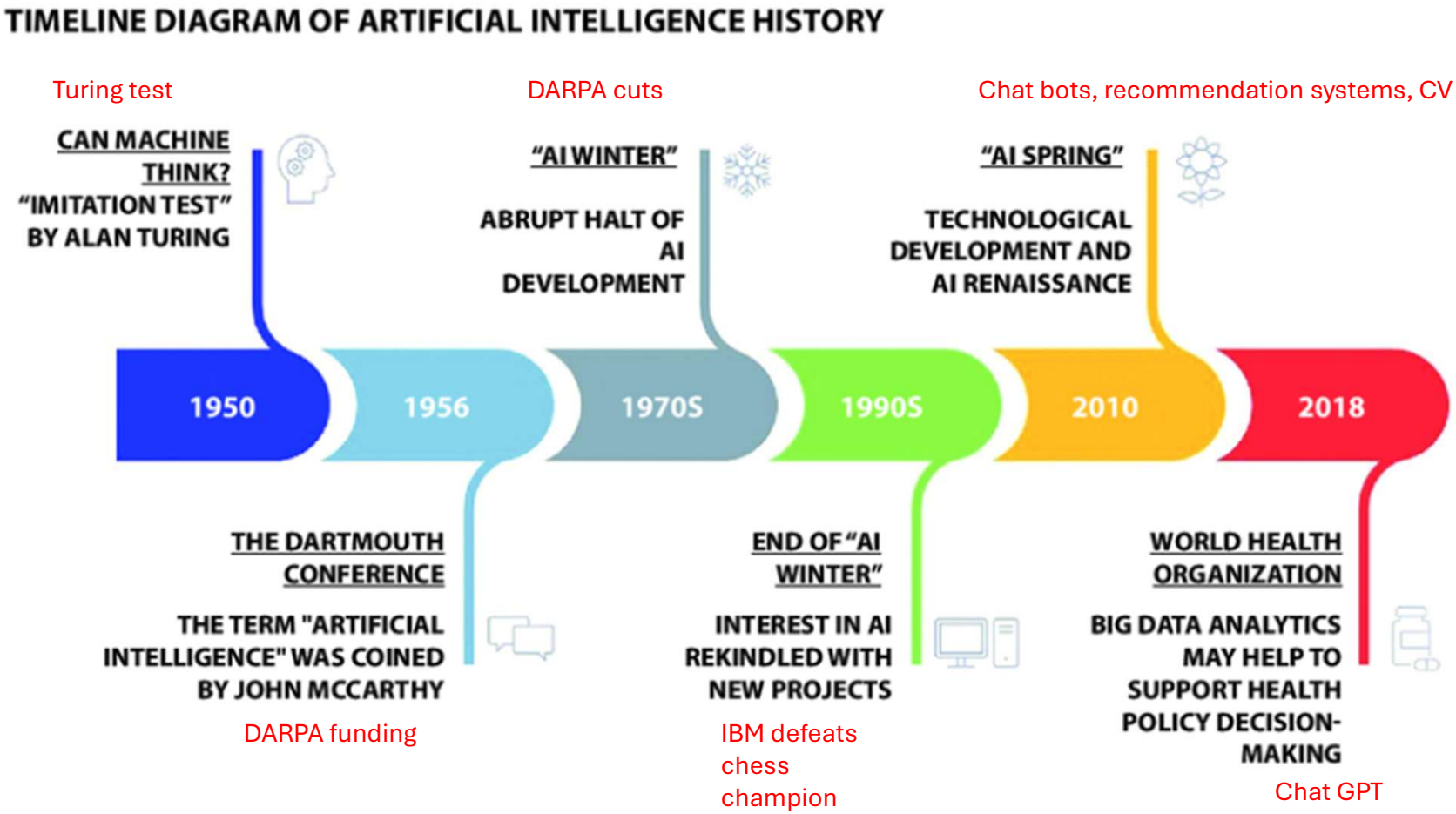
AI: imitate intelligent human behavior

ML: teaching machines to learn and make decisions with being explicitly programmed

DL: A ML technique that stimulates the human brain

GenAI: Creates new content by learning patterns from existing data





https://www.researchhgate.net/figure/Timeline-diagram-showing-the-history-of-artificial-intelligence_fig1_364826401

Machine learning

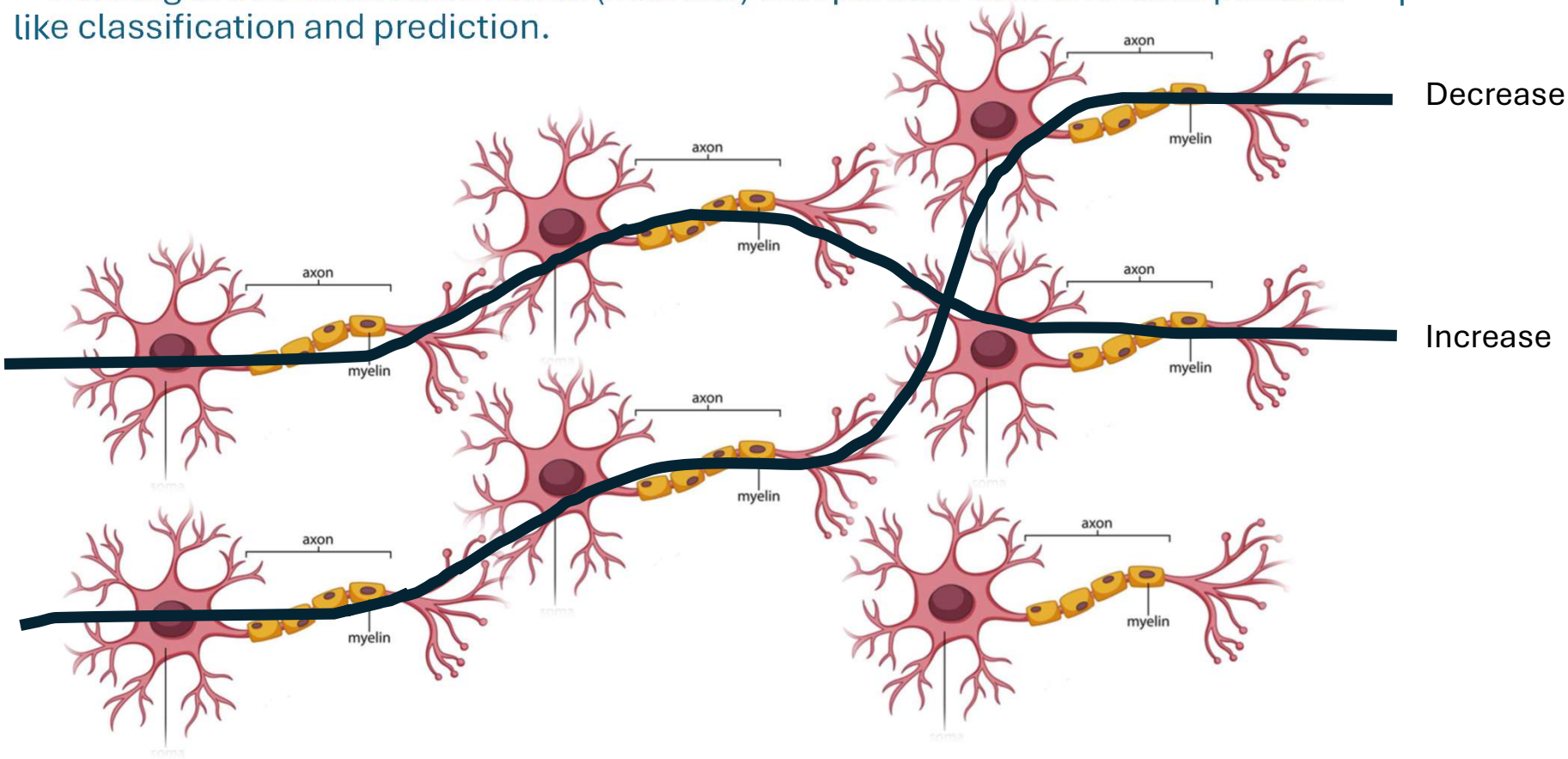
- Supervised learning
  - Massive labeled data sets required
  - During training the computer adjusts internal parameters to minimize the difference between its predictions and the correct answers
  - Success is measured in accuracy
  - ImageNet, 14M hand labeled images
  - Examples: image classification, spam filtering, facial recognition, speech recognition
- Unsupervised learning
  - Lots of data without any labels and ask computer to do the grouping
  - It can discover patterns not apparent by looking at the raw data
  - Clustering similar data
  - Examples: anomaly detection, recommendation systems, clustering customer segmentation

Machine Learning

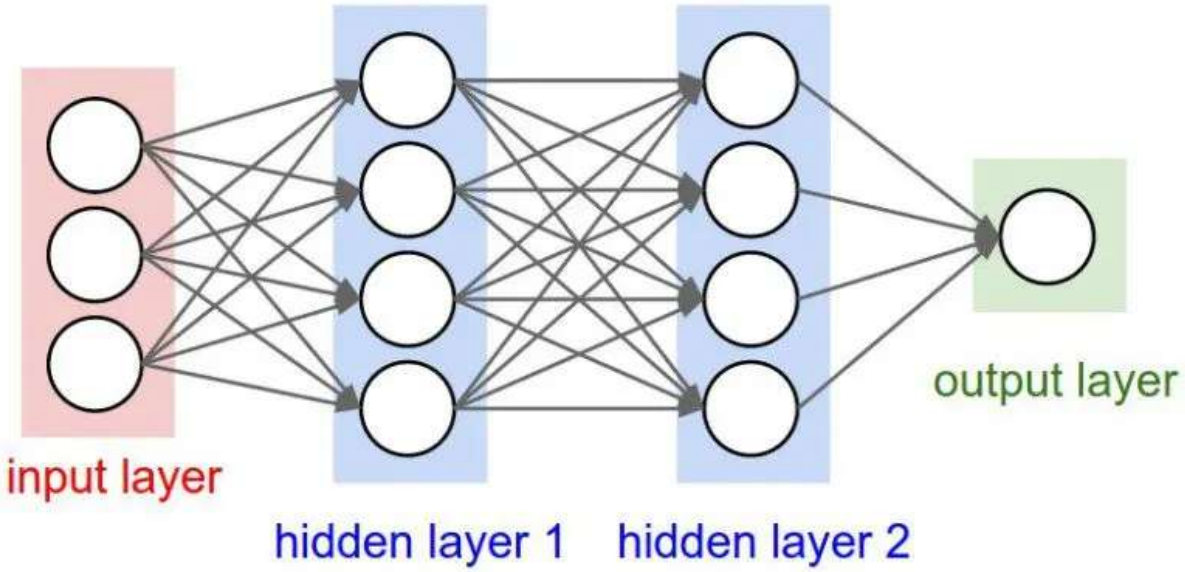
- Feature Extraction
  - Edges, corners, patterns used to distinguish different objects or scenes. Reduces the dimensionality of the data. Very useful in CV (computer vision).
- Transfer learning
  - Leveraging a pre-trained model as a starting point for a new dataset instead of training a model from scratch (ex, start with a model trained on imagenet, chest x-ray training example)
- Data augmentation
  - A technique to artificially increase the size and diversity of a training dataset (rotations, crops, flips, contrast, brightness)

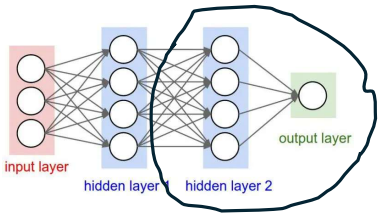
Deep Learning - What is a Neural Network

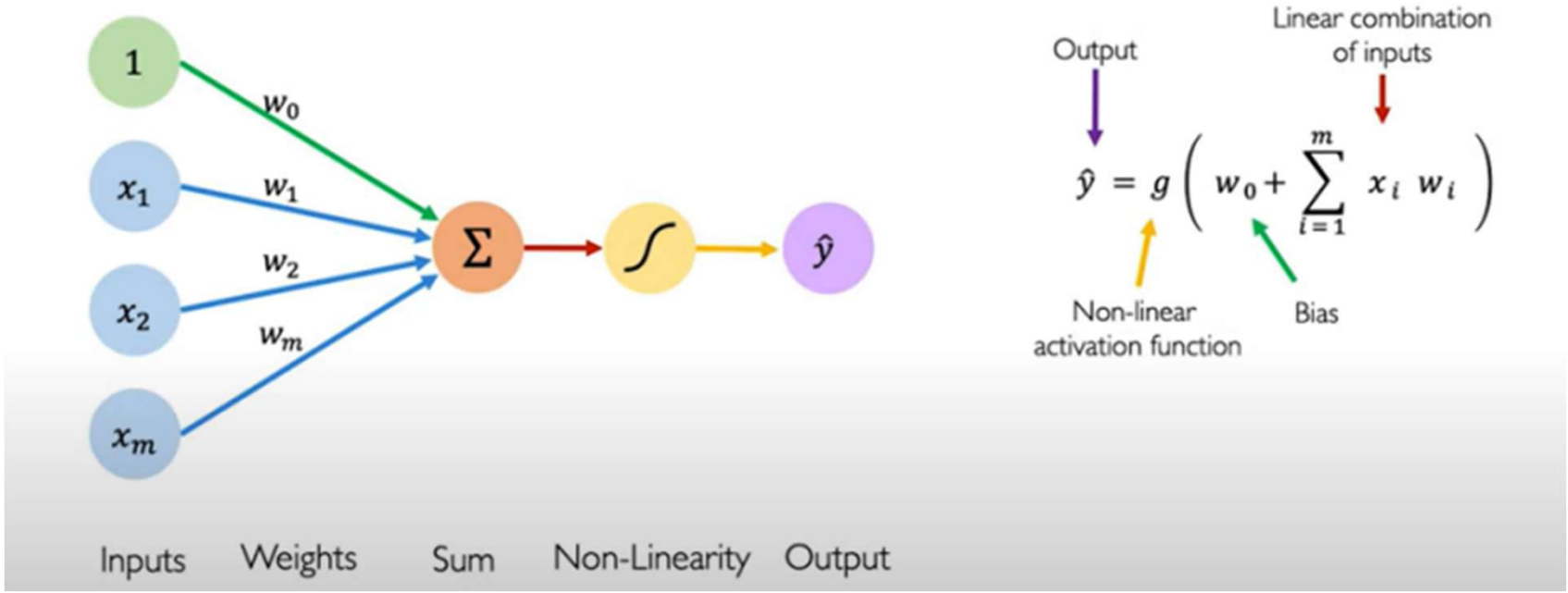
A neural network is a computational model inspired by the structure and function of the human brain, consisting of interconnected nodes (neurons) that process data and learn patterns to perform tasks like classification and prediction.



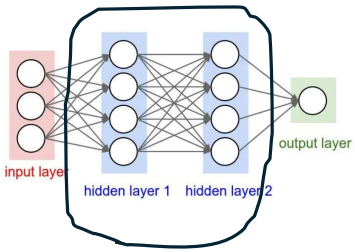
Computer Neural Network

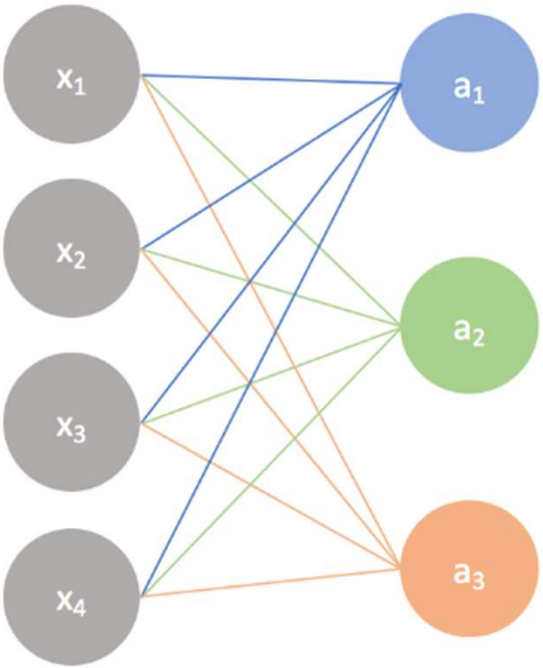






https://math.berkeley.edu/~bernd/vahid.pdf





$$\begin{bmatrix} w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow{activation} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

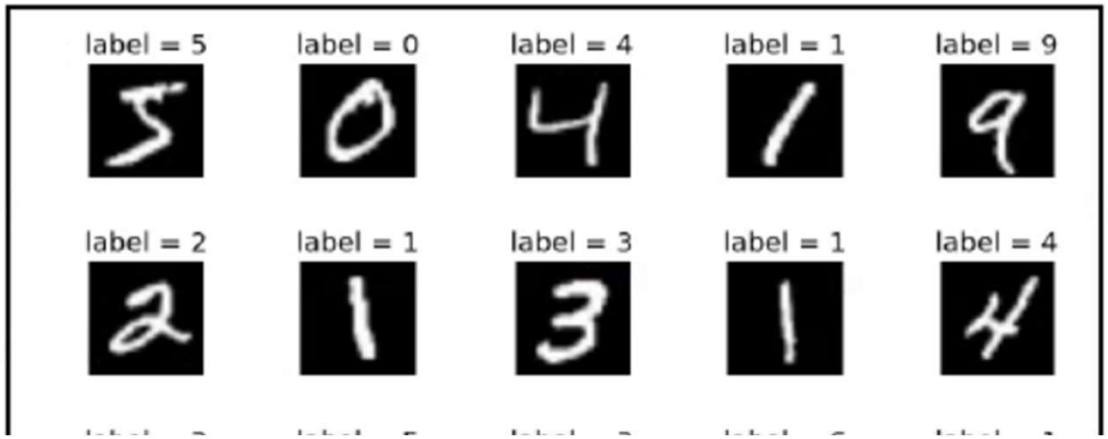
With large matrix math, its easy to see why GPUs are important

https://www.jeremyjordan.me/intro-to-neural-networks/

MINST Example

The MNIST datasets

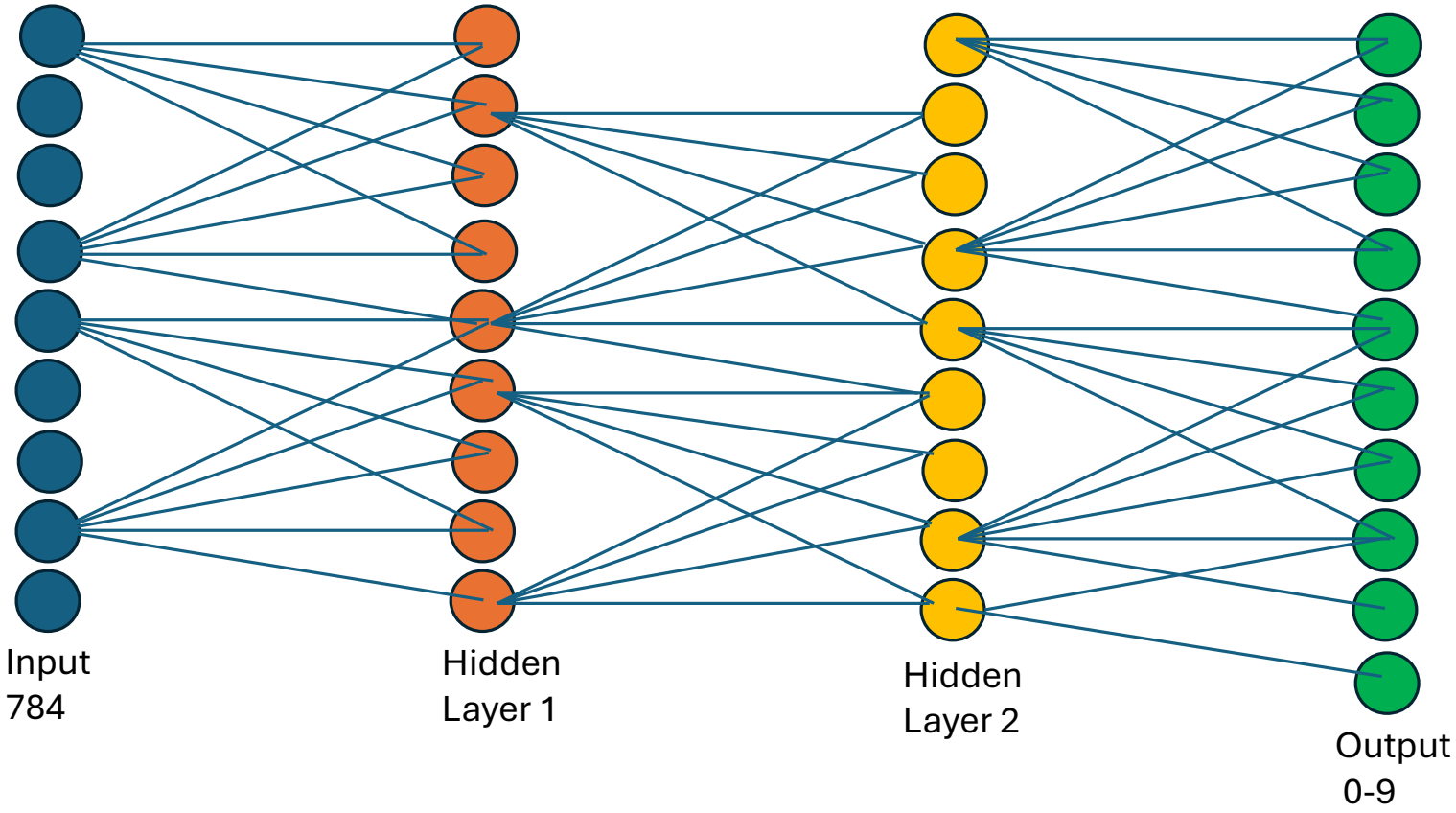
The MNIST data has handwritten digits from 0–9 with 60,000 images for training and 10,000 images for testing. This database is widely used to try algorithms with minimum preprocessing. It's a good and compact database to learn machine learning algorithms. This is the most famous database for image classification problems. A few examples are shown here:



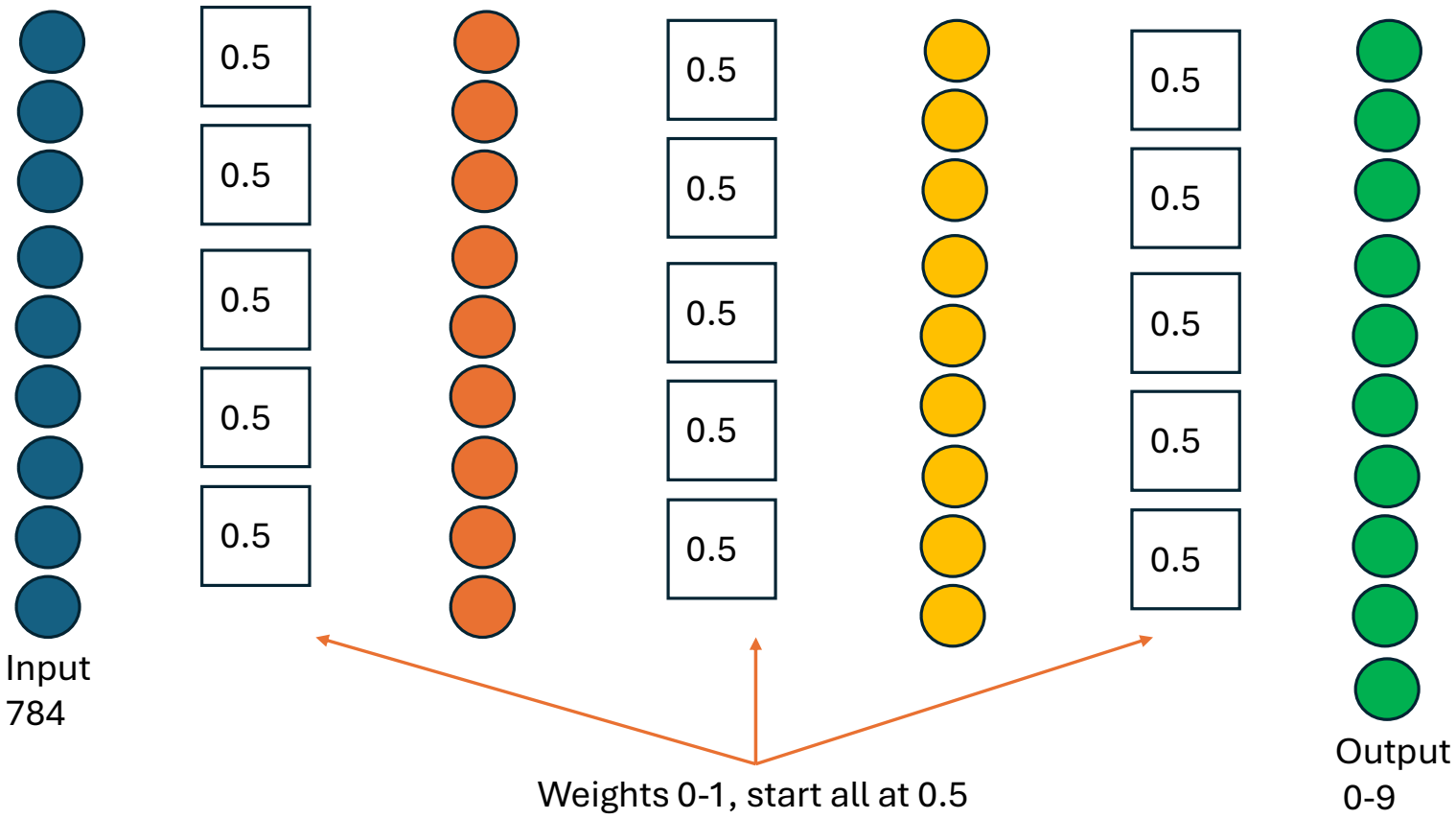
- 28x28 pixel images
- Greyscale integers between 0 and 255
- 784 total pixels

https://math.berkeley.edu/~bernd/vahid.pdf

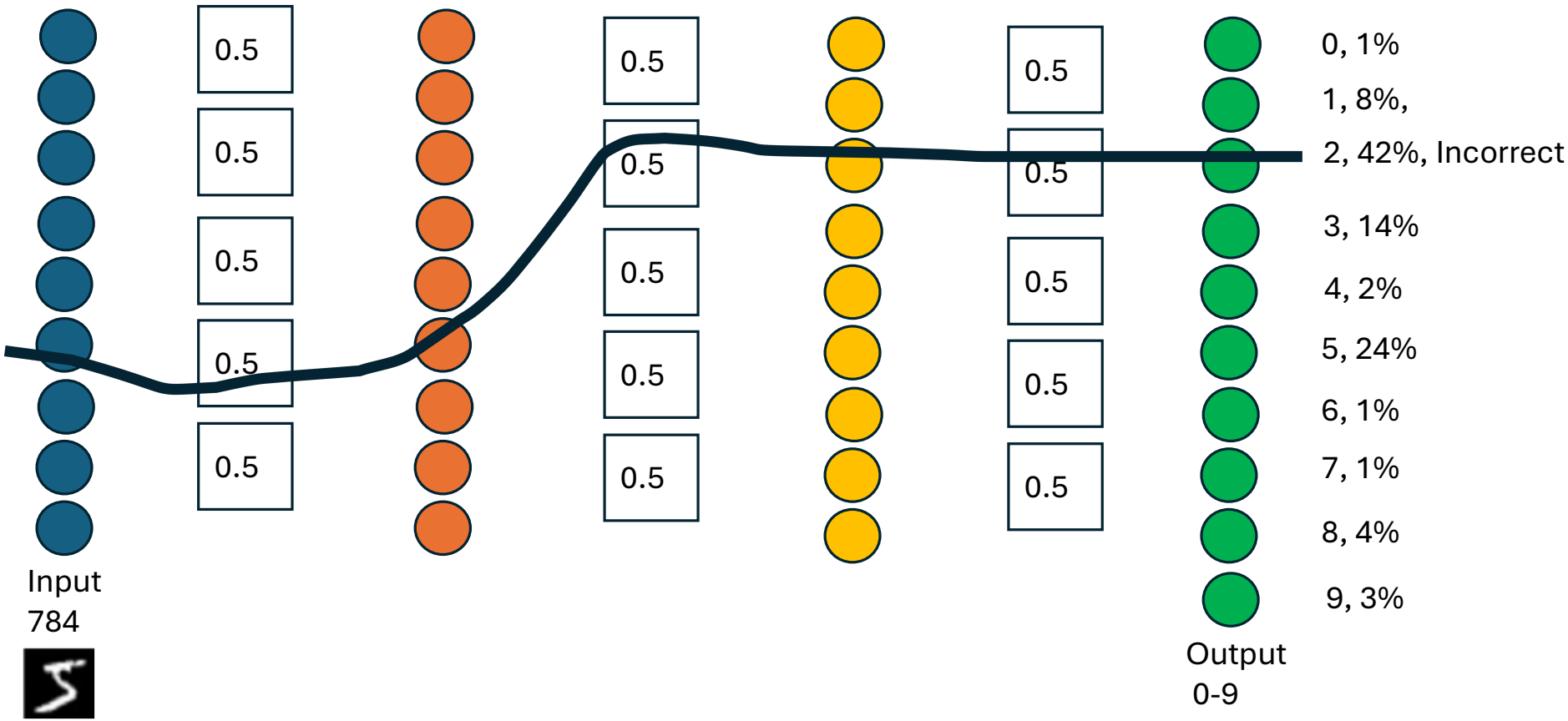
MINST Training example



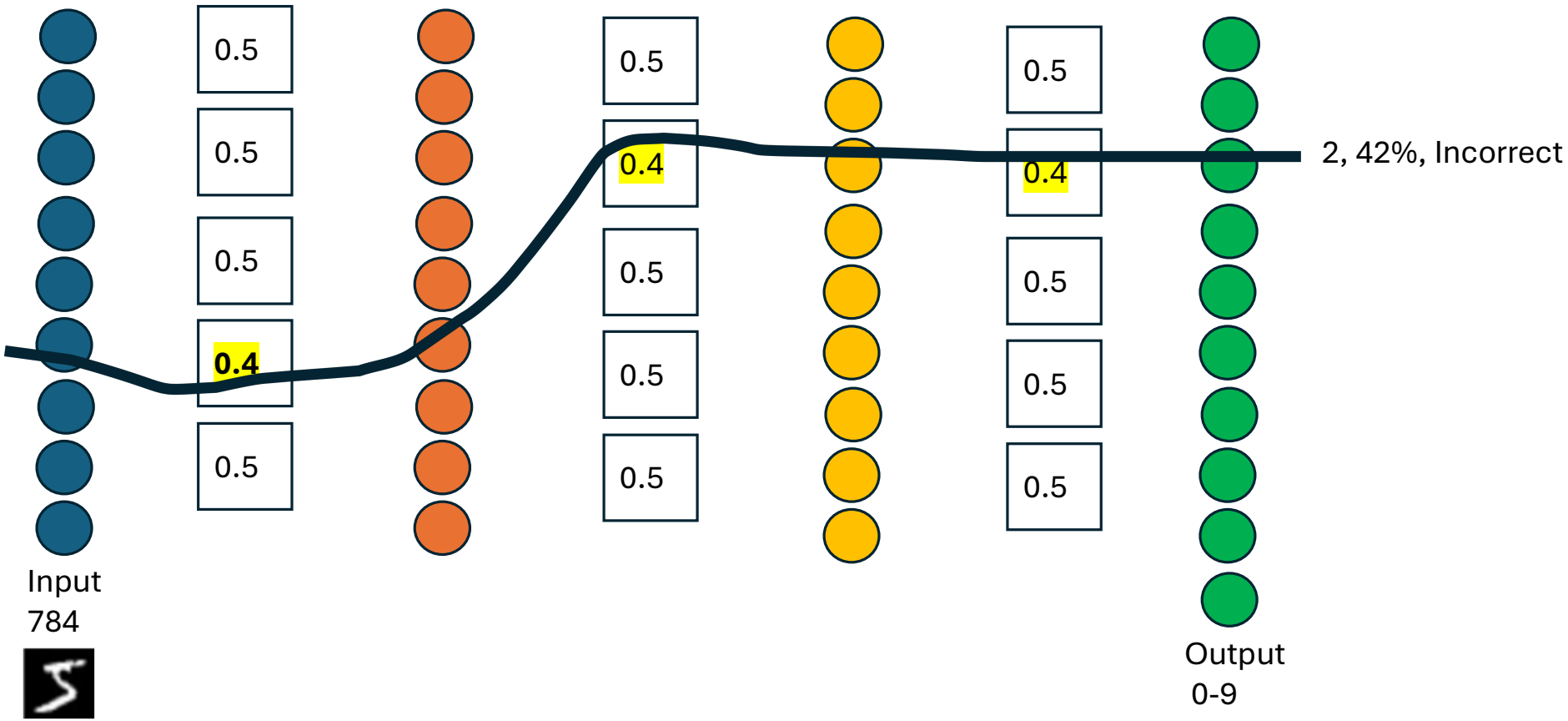
MINST Training example



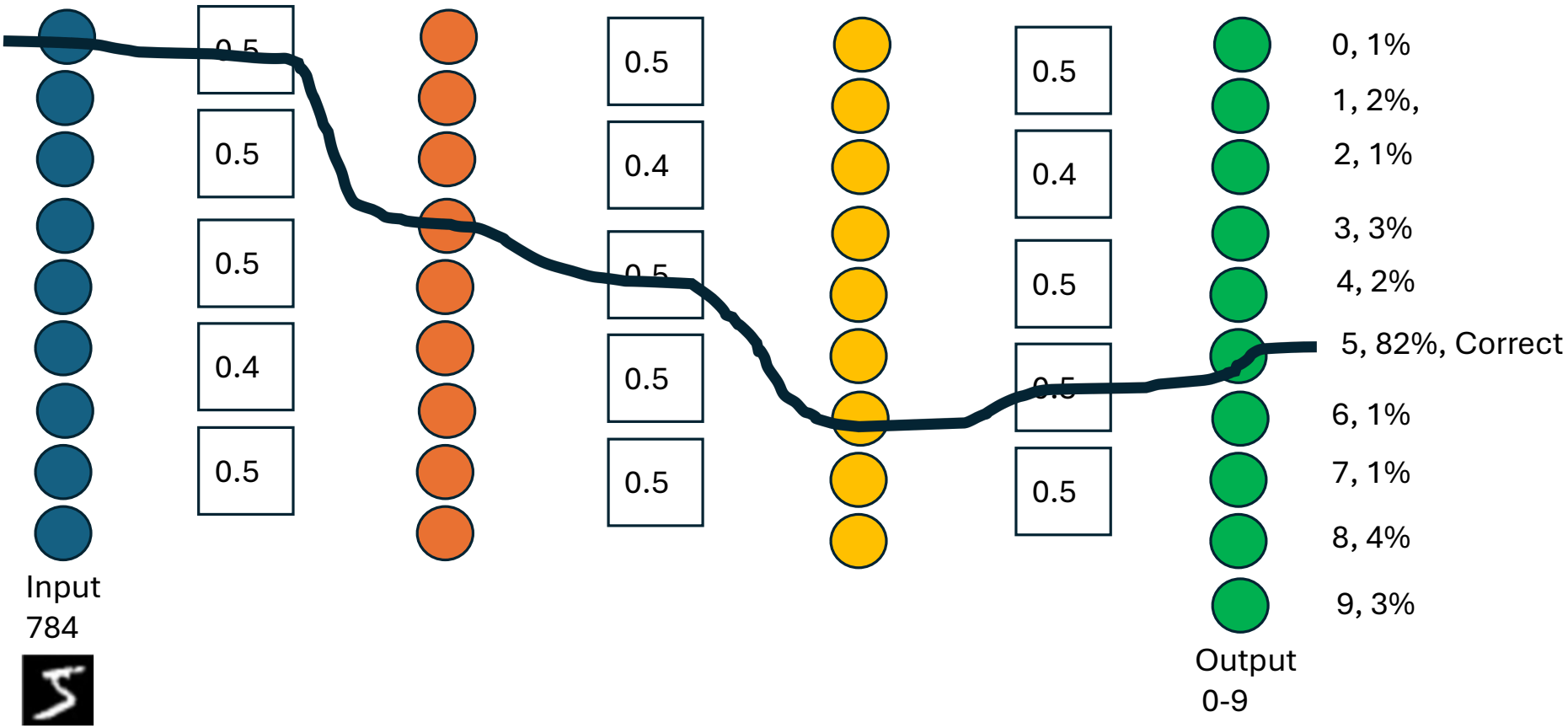
MINST Training example



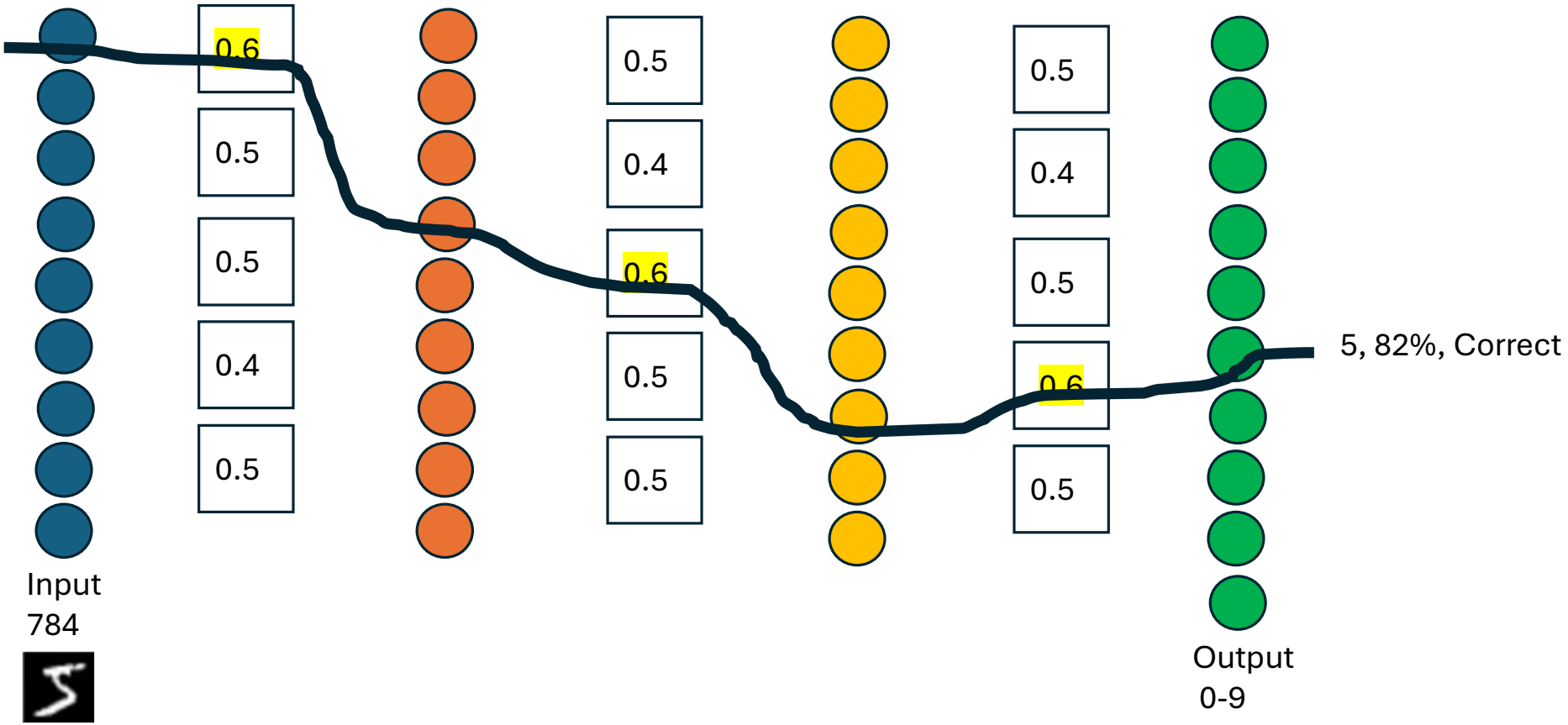
MINST Training example



MINST Training example

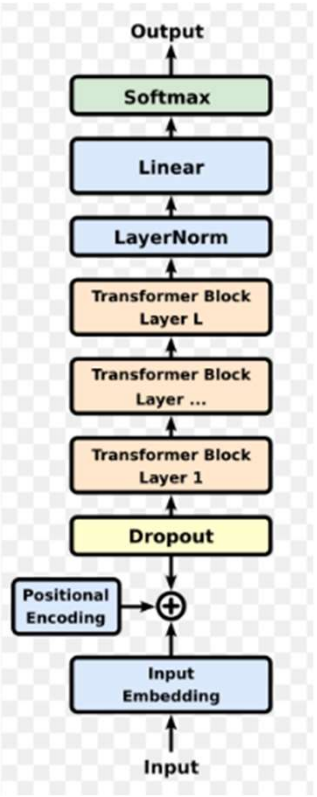


MINST Training example



Generative AI - Transformer architecture - 2017

Generative artificial intelligence (AI), or GenAI, is a type of AI that uses machine learning models to produce new content based on patterns learned from large amounts of data.



Attention

Neural networks are great for sequential data but struggle with long range dependencies.

- *She poured water from the pitcher to the cup until it was full.*
  - "it" = cup
- *She poured water from the pitcher to the cup until it was empty.*
  - "it" = pitcher

Self Attention is key to the transformer architecture and solves this

- Weighs the importance of different parts of the input
- Positionally encodes the input
- Can process more in parallel

Generative AI - Transformer architecture - 2017

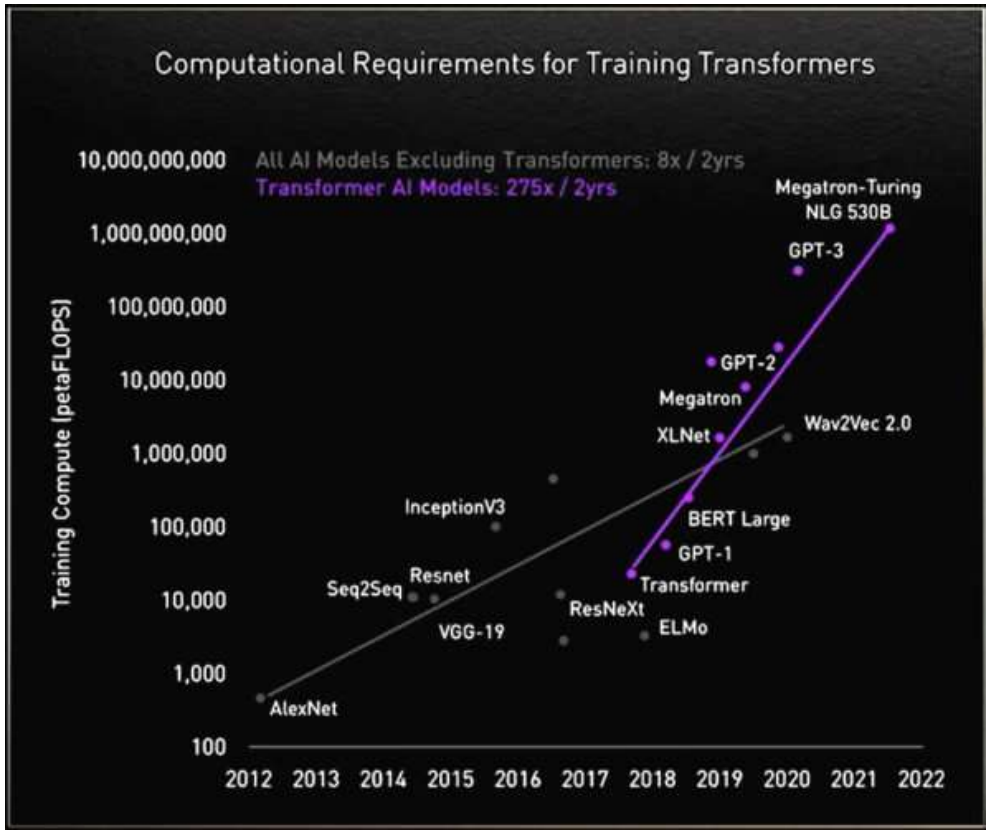
Self-attention, a mechanism within transformer models, allows AI to weigh the importance of different words in a sentence when interpreting meaning. It does this by transforming each word into a query, key, and value. The query from one word interacts with the keys of all words, calculating a score that represents how much attention should be paid to each. These scores are used to create a weighted sum of the values, effectively highlighting the most relevant context for understanding the word in question. This process is performed for every word, enabling the model to capture complex dependencies and relationships within the text.

Consider analyzing the sentiment of a movie review: "The movie started off slow, but the ending was surprisingly good."

- **RNN Approach:** An RNN would process the words sequentially, potentially getting stuck on the negative sentiment of "slow" and missing the positive sentiment of "surprisingly good" at the end.
- **Transformer Approach:** A transformer would consider all the words at once, recognizing that "surprisingly good" is more important for determining the overall sentiment of the review, even though it appears later in the sentence.

OpenAI - LLMs

- 2018 - GPT1 - 117M parameters, based on 7000 unpublished books
- 2019 - GPT2 - 1.5B parameters, 40GB of text + 8M documents from 45M web pages
- 2020 - GPT3 - 175B parameters, 570GB internet text
- 2022 - ChatGPT release using GPT3.5 – undisclosed training data size
- 2023 GPT4 - est. 1.7T parameters – undisclosed training data size
- 2024 - GPT4o and GPT4-mini
- TBD – GPT5



Part 2: Prompt Engineering

- Part 1: AI fundamentals
  - History
  - Machine Learning
  - Neural Networks - Deep Learning
  - Transformers
  - LLMs (Large Language Models)
- **Part 2: Prompt Engineering**
  - **Basic Prompts**
  - **Prompt foundational structure**
  - **Advanced techniques: one shot, chain of thought, persona, self prompting, etc**
- Part 3: Challenges and Future Trends
  - Challenges: Security, shadow prompting, bias, overfitting, etc
  - Trends: Multimodal, AI Agents, Edge AI, etc.

Audience Poll

- Have you heard of prompt engineering

- Poll results

No	78%
Yes	22%

Prompt Engineering

- Prompt engineering is not formally recognized as an engineering discipline in the traditional sense, such as electrical, civil, or mechanical engineering. However, it is a rapidly evolving field and that there is growing interest in establishing it as a formal discipline.

- As of now, there are no ABET-accredited programs specifically for prompt engineering.

- ABET does accredit programs in applied and natural science, computing, engineering, and engineering technology. These programs often include courses and specializations that cover aspects of artificial intelligence (AI) and machine learning (ML), which are foundational to prompt engineering

What is a prompt, entry level prompts

- What is a prompt
  - "A prompt is a carefully crafted input or question given to a language model to generate a relevant and accurate response based on the provided context.“

- Some very basic unstructured prompts
  - Tell me about the Eiffel Tower
  - Explain photosynthesis
  - Please summarize this document
  - Draft a letter about

- We are here to go beyond this level

Let’s start with the Foundation of a good Prompt

Foundation of a prompt – example 1

1.Task — The action verb + goal. Start prompts with a clear task.	Mandatory
2.Context — Background info to constrain possibilities. **3.Exemplars** — Examples to guide the AI.	Highly Recommended
4.Persona — Who you want the AI to be. **5.Format** — Visualize and specify desired structure. **6.Tone** — Casual, formal, excited, etc.	Nice to have

https://medium.com/@RickTaylar/the-only-prompt-formula-youll-ever-need-for-chatgpt-6946599c2912
https://www.youtube.com/watch?v=bkf3XBOj2PE

Foundation of a prompt – example 1

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Foundation of a prompt – example 1

Task

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Foundation of a prompt – example 1

Context

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Foundation of a prompt – example 1

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Exemplars

Foundation of a prompt – example 1

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Persona

Foundation of a prompt – example 1

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Format

Foundation of a prompt – example 1

Explain the importance of thermal management in electronic device design for an engineering blog focused on electronics and hardware. In your explanation, discuss why thermal management is crucial for performance, reliability, and longevity of electronic devices. Describe common techniques used in thermal management, such as heat sinks and cooling systems, and their role in maintaining optimal operating temperatures. You are a knowledgeable electronics engineer with a deep understanding of hardware design principles. The explanation should be a concise paragraph, around 150-200 words, written in an informative and professional tone, with an emphasis on clarity and practical relevance.

Tone

Foundation of a prompt – example 2

From Microsoft Copilot Training - https://support.microsoft.com/en-us/topic/learn-about-copilot-prompts-f6c3b467-f07c-4db1-ae54-ffac96184dd5

4 key elements

- Goal – what do you want
- Context – why do you need it and who is involved
- Expectations – how best to fulfill (tone/style)
- Source – What information or samples to use

Sample:
Source +Goal: Based on general best practices from leading tech companies such as Google, Microsoft, and Apple, provide advice on improving team productivity.
Context: You are drafting content for a professional development newsletter aimed at managers and team leaders. The audience is interested in actionable tips for enhancing productivity within their teams.
Expectations: Offer three easily implementable strategies, with brief explanations for each.

Advanced prompting techniques

- Persona
- Self prompting
- RAG
- Prompt Chaining
- Confirmation (do you understand)
- One shot / few shot
- Chain of Thought
- Panel tree of thought
- Metaprompts

Personas

Executive Admin

- *You are* an executive administrative assistant who has vast experience .....

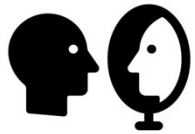
Tutor

- I would like to play *a game* with you. *You are* a Spanish teacher and assigned to help .....

System Engineer

- *You are* an experienced system engineer. Please review these files and *leveraging your vast experience* highlight the areas where our design may fall short of the pole stars identified

Self prompting



Example 1

- I need to build a presentation on using AI in the workplace to improve efficiency and effectiveness. *feel free to ask me any clarifying questions first*
- Okay, let's start with a few questions to help me tailor the presentation to your needs:
  1. Who is your target audience? Are they executives, middle management, or employees?
  2. What is the level of their understanding of AI? Is this a basic introduction or a deep dive?
  3. What industry is the presentation intended for? AI applications can vary greatly by industry.
  4. Are there any specific aspects of AI you want to focus on? For example, automation, data analysis, machine learning, etc.
  5. How long does the presentation need to be?
  6. .....

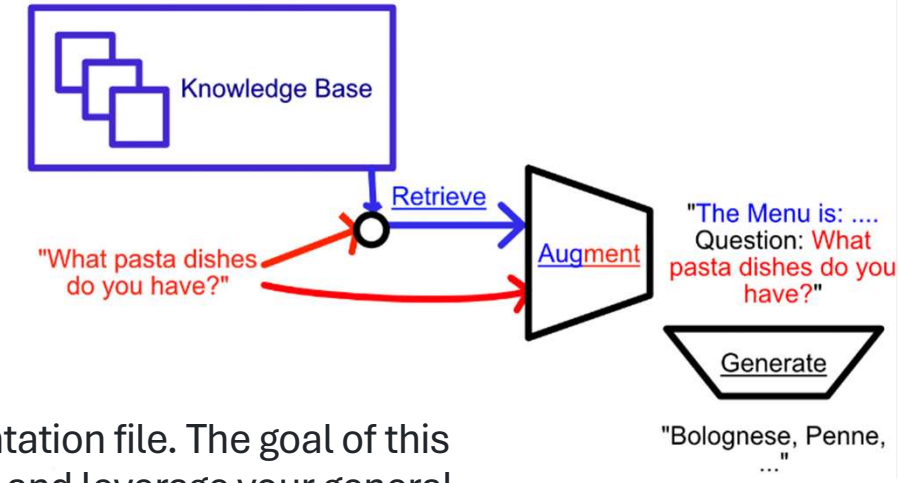
Example 2

- .... for each task, *ask me questions* about how this task is currently being completed that will *help you ascertain* if this task could be automated. *Only after I have responded to each of your questions* ....

Will self prompting evolve to a point where prompt engineering is not necessary?

RAG (Retrieve-Augment-Generate)

RAG is a method that combines the strengths of information retrieval and large language models. It first retrieves relevant information from a knowledge base based on the user's query and then uses it to generate more accurate and informed responses.



Example:

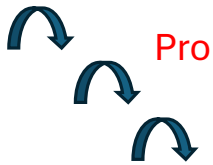
Attached is a presentation meeting transcript and the presentation file. The goal of this presentation is ..... You are a ..... Please review these items and leverage your general knowledge of ..... Please suggest specific follow up items for me to take and *provide feedback on my presentation (what could I have done better, what might I have missed, did I state anything incorrectly)*.

Google Gemini can link to google drive or attach files.

Some companies build their own RAG systems

RAG can be SQL entries into a database like Jira

Prompt Chaining



Prompt chaining is a technique where the output of one language model prompt is used as input for the next prompt. This allows for building upon previous responses and creating more complex interactions with the language model.

Another word for iterative prompting

Break down the ask into smaller simpler steps

Example

- Can you tell me about the top challenges of this design?
  - Can you tell me a little more of the fingerprint sensor challenges?
    - Can you suggest some ways to address the dust ingress challenge with the fingerprint sensor?

Confirmation Words

- In prompt engineering, confirmation words like "Do you understand?" are used to elicit a clear response from the language model, indicating whether it has grasped the instructions or information provided in the prompt. This helps ensure the model is on the right track before proceeding with further tasks or generating output.

- Examples:
  - Do you understand
  - Only after I have
  - Confirm you understand the request and then I will proceed with supplying

One/Few shot learning

One/few-shot learning is a machine learning approach where a model learns to perform a task from a very limited number of examples (one or a few). It aims to mimic the human ability to quickly learn new concepts with minimal data.

Zero shot	No in-context training. General queries, simple tasks. language translation, basic classification tasks, sentiment analysis
One shot	Customer support, niche products. Legal drafting. Following a particular tone / style / level of detail. Identifying a new object in an image where not feasible to collect and label a large data set
Two shot	Drafting technical manuals, drafting product descriptions, personalized recommendations.
Many shot	Full automation, diagnosing medical conditions, detailed financial forecasting.

One/Few shot learning

Example 1

Zero shot: If ABBC4 changes to ABBD3, what does PQQR7 change to?

High probability it provides an incorrect answer

Two shot: Example1; DEEF4 changes to DEEG3. Example2; TUUV9 changes to TUUW8. What does PQQR7 change to?

Very high probability it is correct

Example 2

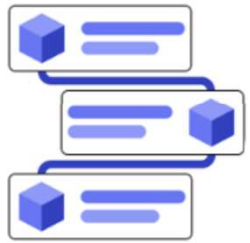
Here is an *example input* and *expected output* for summarizing a teams' set of status reports. Please analyze this transformation for level of detail, tone, what level of information get summarized, what level of detail gets omitted and overall summary style. *Confirm you understand the request and then* I will proceed with supplying a new set of inputs for you to perform a similar summary compilation

########Input:

########Output:

…. I understand the request and am ready to proceed with a new set of inputs to perform a similar summary compilation

Chain of thought



Chain of thought prompting is a method used in prompting language models that encourages them to break down complex problems into smaller, intermediate steps. This allows the model to better reason through the problem and provide a more accurate and detailed final answer

Q: i was born on nov 1 1972. what day of the week was i born?

High probability of an incorrect answer.

Q: i was born on nov 1 1972. what day of the week was i born? Please *think through step by step*

To find out the day of the week you were born on, you can use an online day-of-the-week calculator or do it manually with the Zeller's Congruence formula. Since you provided your birthdate, I can help you calculate it manually.

Zeller's Congruence formula:

h=(□+⌊13(□+1)5⌋+□+⌊□4⌋+⌊□4⌋−2□)mod 7h=(q+⌊513(m+1)⌋+K+⌊4K⌋+⌊4J⌋−2J)mod7

Much higher probability of a correct answer

Panel Tree of Thought

Panel tree of thought is a prompting technique where the language model generates multiple possible solutions to a problem, evaluates each solution, and selects the best one to continue exploring. It's like having a panel of experts brainstorm and debate to find the optimal path forward.

Example:

Consider a scenario where you want to plan a weekend trip

1. Generate multiple solutions: The language model generates several potential destinations, such as a beach town, a mountain retreat, a city with cultural attractions, or a national park.

2. Evaluate each solution: The model assesses each option based on criteria like budget, travel time, interests, and weather conditions. For example, a beach town might be great for relaxation but could be expensive during peak season.

3. Select and explore: The model might choose the city with cultural attractions as the most promising option. It would then generate more specific ideas for activities within the city, such as visiting museums, trying local cuisine, attending a concert, or exploring historical sites.

Metaprompts

A metaprompt is a high-level instruction or framework given to a language model to guide its response generation. Instead of providing specific content details, metaprompts focus on the structure, format, or desired outcome of the response.

Mostly for application developers

Can control response formats (list, table, code snippet)

Guide reasoning processes

Control outputs (ex. What it will not answer on)

Improve consistency (align with a style or brand)

Prompting wrap up

- Basics of a good prompt foundation go a long ways
- Get comfortable inserting key phrases:
  - *Act as if; You are a; Do you understand; Lets think through step by step; Leverage your expertise; Feel free to ask me any clarifying questions first*
- *Other*
  - Shouldn't be using the tool to get information on something you don't have some level of expertise in
  - Including too many unnecessary details is not necessarily a good approach
  - Use clear separators like "###" to separate the instruction and context, ask for separators in outputs. ChatGPT is color and font blind
  - Empathy matters.
  - If GPT gets stuck, you can always refresh the session. Every time you ask it will be a little different.
  - GPT can sometimes go too deep into the beginning of a large dataset – break up the task
  - Get comfortable with using shift-enter when entering your prompts.

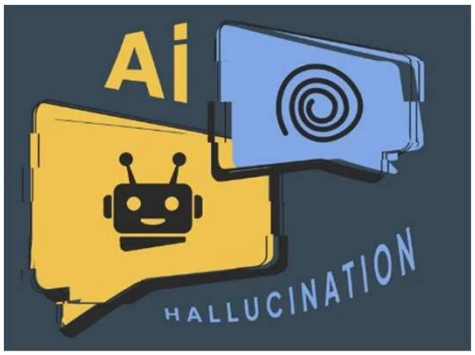
Part 3: Challenges and Future Trends

- Part 1: AI fundamentals
  - History
  - Machine Learning
  - Neural Networks - Deep Learning
  - Transformers
  - LLMs (Large Language Models)
- Part 2: Prompt Engineering
  - Basic Prompts
  - Prompt foundational structure
  - Advanced techniques: one shot, chain of thought, persona, self prompting, etc
- **Part 3: Challenges and Future Trends**
  - **Challenges: Security, shadow prompting, bias, overfitting, etc**
  - **Trends: Multimodal, AI Agents, Edge AI, etc.**

Challenges

- Hallucinations
- Security and Data Privacy
- Copyright and data theft
- Shadow AI
- Bias
- Overfitting
- Deep Fakes
- Intensive resource requirements

Hallucinations



An AI hallucination is when an AI model generates output that is plausible-sounding but factually incorrect or nonsensical.

Examples:

- Text-based hallucination: An AI-powered language model, when asked about historical events, confidently asserts that Abraham Lincoln was the first president of the United States, despite George Washington holding that title.
- Image-based hallucination: An AI image generator, prompted to create a picture of a realistic human face, produces an image where the person has three eyes instead of two, defying the anatomical norms of human beings.

Security and Data Privacy

AI security and data privacy focuses on safeguarding AI systems and the data they process from unauthorized access, misuse, and manipulation. It involves both protecting sensitive data used to train and operate AI models, as well as securing the AI algorithms themselves from vulnerabilities.

Examples:

- Data anonymization: Before training an AI model to diagnose medical conditions, patient data is stripped of personally identifiable information (PII) like names and social security numbers to protect patient privacy.
- Adversarial attack mitigation: Researchers develop techniques to detect and defend against adversarial attacks, where malicious actors intentionally manipulate input data to cause AI models to make incorrect predictions, potentially leading to harmful consequences.

Copyright and data theft

AI copyright and data theft involve the unauthorized use of copyrighted materials, including code, text, or images, to train AI models or the theft of sensitive data used for AI development.

Examples:

- Copyright Infringement: An AI art generator trained on copyrighted artwork without obtaining proper licenses, leading to the creation of derivative works that could potentially infringe on the original artist's rights.
- Code Plagiarism: A company copies proprietary code from a competitor's AI model to accelerate their own development, violating software licensing agreements and intellectual property rights.
- Data Scraping: A research group scrapes large amounts of copyrighted text from websites without permission to train their language model, potentially infringing on copyright laws and raising ethical concerns about fair use.

Shadow AI

Shadow AI refers to the use of artificial intelligence tools and applications within an organization without explicit approval or oversight from IT departments or management. This can create security, privacy, and ethical risks for the organization.

Examples:

- A marketing team member uses a free AI-powered writing tool to generate social media posts, bypassing the company's approved content creation process and potentially violating brand guidelines or compliance regulations.
- An employee utilizes a personal ChatGPT account to automate customer support responses, without informing IT or ensuring the AI's accuracy and alignment with company policies.

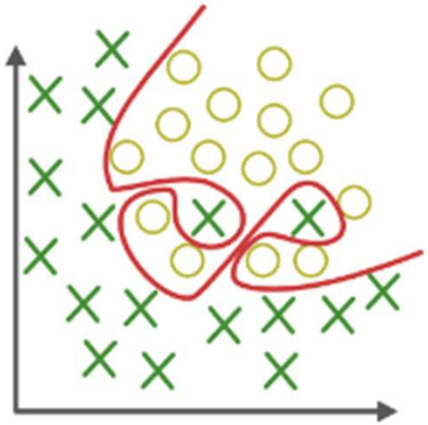
Bias

AI bias refers to the systematic errors or unfair outcomes in AI systems' decisions, predictions, or behaviors, often stemming from biases present in the data used to train them.

Examples:

- A facial recognition system struggles to accurately identify individuals with darker skin tones due to underrepresentation in the training data, leading to potential discrimination and misidentification.
- Suppose you need data about the travel patterns of people commuting to and from work in order to create public transportation schedules, so you gather information on the geolocations of smartphones during commuting hours. The problem is that 15% of Americans, or roughly 50 million people, don’t own a smartphone. Many simply cannot afford a device and a data plan. People who are financially less well off, then, would be underrepresented in the data used to train your AI. As a result, your AI would tend to make decisions that benefit the neighborhoods where wealthy people live.

Overfitting



AI overfitting occurs when a model learns to perform exceptionally well on its training data but fails to generalize to new, unseen data. This happens because the model has essentially memorized the training data instead of learning the underlying patterns.

Examples:

- Spam Filter Overfitting: A spam filter trained on a specific set of emails becomes so attuned to those examples that it starts classifying legitimate emails from new senders as spam, simply because they don't match the patterns it memorized.
- Stock Market Prediction Overfitting: An AI model trained on historical stock market data perfectly predicts past trends but fails to accurately forecast future market movements because it has over-relied on specific past events rather than learning broader market principles.

Deep fakes

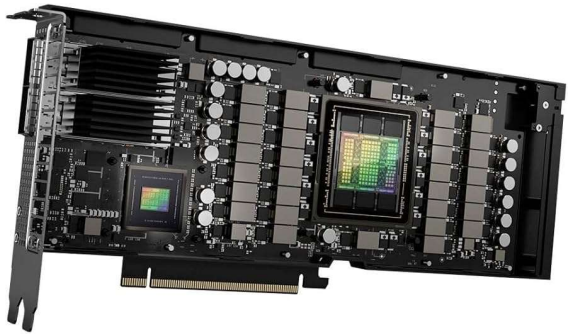
AI deep fakes are synthetic media, such as images, videos, or audio, created using advanced artificial intelligence techniques, particularly deep learning, to convincingly mimic real people and events.

Examples:

- Deepfake Videos: A video where a public figure’s face is swapped with another person's face, making it appear as though the public figure said or did something they never actually did.
- Voice Deepfakes: Audio clips where a person’s voice is synthetically generated to produce speech that they never spoke, often used in spoofing and impersonation attacks.

Intensive Resource Requirements

The challenge of intensive resource requirements for AI is that it demands significant investments of time, money, and energy to train and develop sophisticated models.



- Time: Training a complex AI model can take weeks or even months of continuous computation.
- Money: Building and maintaining the infrastructure necessary for AI development, such as powerful servers and GPUs, can be very costly.
  - Example. LLAMA3.1 405B. To Train 16,000 Nvidia H100 GPUs at $30,000 each (half a billion dollars)
- Energy: The computational power required for AI training consumes massive amounts of electricity, contributing to environmental concerns.
  - Example: A study showed GPT3 175B parameters = 552 tons of CO_2 equivalent which is roughly equivalent to the emissions from 120 average American cars. Using a linear extrapolation, GPT4 at 1.7T parameters is 10x that estimate.

Trends

- Rapid changes
- Multimodal
- AI agents
- Explainable AI
- Edge AI

Rapid Changes

AI is evolving at an unprecedented pace, with new breakthroughs and applications emerging constantly. For example, in just a few years, we've seen AI progress from generating simple text to creating realistic images and even writing code, and AI-powered medical tools can now diagnose diseases with accuracy rivaling human experts.

- Competition is driving changes: 2 weeks ago Meta announced LLAMA3.1 with its customizable features. Hours later OpenAI announced GPT-4o mini and its customizable features
- SearchGPT (announced last week, can join the waitlist)
- SiriAI update – awareness of what is on your screen and the ability to take action
- Friend – AI based pendant
- 7/31 - Fully-automatic robot dentist performs world's first human procedure

Suggestions for feeds with daily updates:

medium.com

tldrnewsletter.com

superhuman.ai

Multimodal

Multimodal AI is a type of artificial intelligence that can process and understand information from multiple sources, such as text, images, and audio.

- Examples,
- a multimodal AI model can analyze a photo and generate a caption describing its content,
- listen to a song and identify the genre and mood.
- Another example is a self-driving car, which uses multimodal AI to combine information from cameras, lidar, and radar to perceive and navigate its environment.

AI Agents

AI agents are autonomous software entities that can perceive their environment, make decisions, and take actions to achieve specific goals. They are becoming increasingly sophisticated and are being used in a wide range of applications.

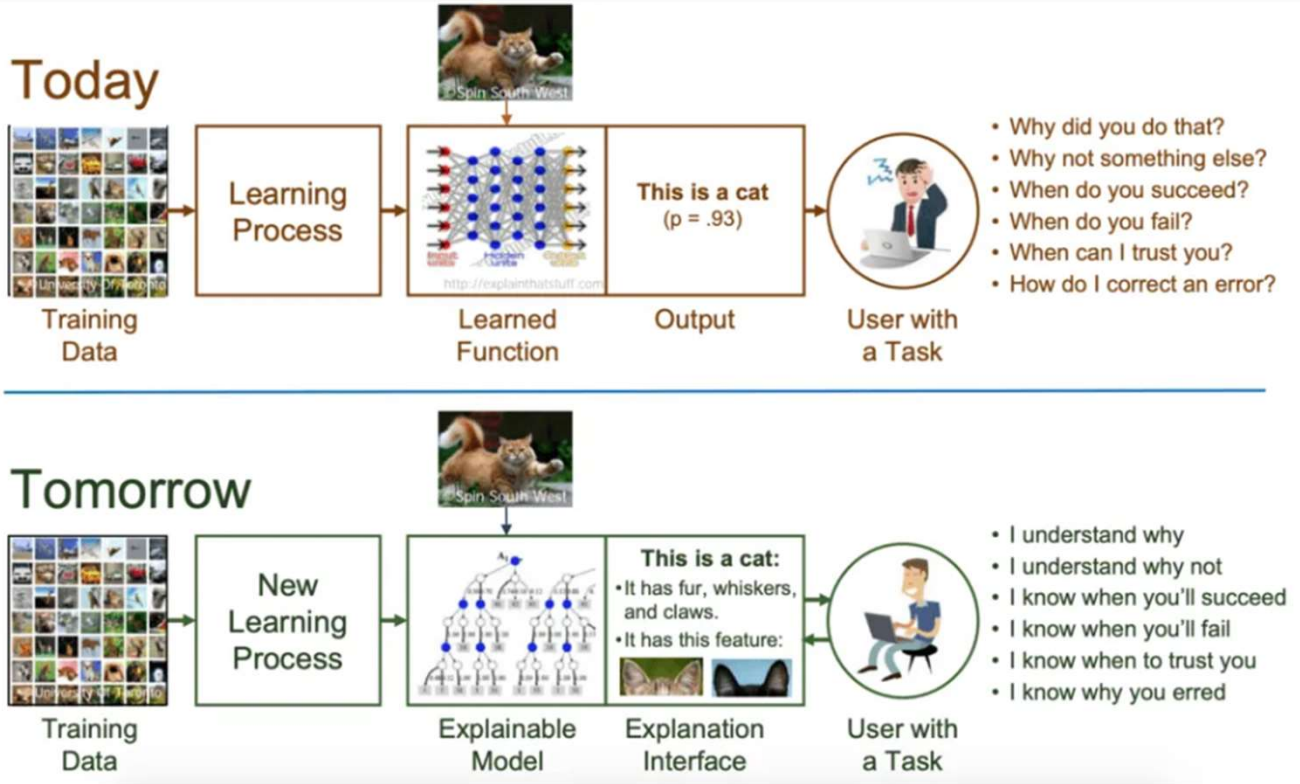
- For example, AI agents can be used to automate customer service interactions, manage complex supply chains, and even assist in scientific research.
- Lots of experimentation on GitHub: example: WebVoyger that can control clicking, typing, scrolling

https://towardsdatascience.com/ai-assistants-copilots-and-agents-in-data-analytics-whats-the-difference-2e63f8fb2384

Explainable AI

Explainable AI (XAI) is a growing trend that focuses on making AI systems more transparent and understandable. It aims to provide insights into how AI models arrive at their decisions, which can be crucial in critical applications.

For example, in healthcare, XAI can help doctors understand why an AI system recommended a particular treatment, while in finance, it can explain the reasoning behind investment decisions.



https://lawtomated.com/explainable-ai-all-you-need-to-know-the-what-how-why-of-explainable-ai/

Edge AI

Edge AI refers to the deployment of AI algorithms and models directly on edge devices, such as smartphones, sensors, or IoT devices, instead of relying on the cloud. This allows for faster processing, reduced latency, and improved privacy.

For example, a smart security camera with edge AI can analyze video footage in real-time to detect intruders without sending data to the cloud, and a self-driving car can make split-second decisions based on its onboard AI model, enhancing safety.

Thank You

Useful references

- Great book that sparked my interest: AI Superpowers by Kai-Fu Lee
- Fun GenAI overview video: https://www.youtube.com/watch?v=2IK3DFHRFfw&t=1s
- Short video on what an LLM is: https://www.youtube.com/watch?v=zKndCikg3R0
- 8 part GenAI article series: https://medium.com/@raja.gupta20/generative-ai-for-beginners-part-1-introduction-to-ai-eadb5a71f07d
- "The" prompt engineering encyclopedia: https://www.promptingguide.ai/
- Good article on advanced prompts: https://ai.gopubby.com/4-human-ai-interaction-patterns-for-experienced-chatgpt-users-9e49d4234013
- Article on few shot learning: https://www.promptingguide.ai/
- Article on prompt engineering is not the future: https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future
- Ideas for handling long prompts and large context windows: https://www.forbes.com/sites/jodiecook/2023/11/27/6-chatgpt-prompts-to-create-an-entire-ebook-and-grow-your-business/
- More on prompt foundations: https://medium.com/the-generator/the-perfect-prompt-prompt-engineering-cheat-sheet-d0b9c62a2bba
- Designing for AI which leads to why metaprompts are important: https://uxdesign.cc/designing-for-ai-beyond-the-chatbot-5edc0efe84a3
- Short video on 2024 AI trends: https://www.youtube.com/watch?v=sGZ6AlAnULc
- More on agent engineering: https://towardsdatascience.com/from-prompt-engineering-to-agent-engineering-f314fdf52a25
- AI agents impact on the workforce: https://medium.com/@sanguit/how-ai-agents-rewire-the-organization-3b18bf21912a
- CV and DL site with many good books and courses. Many free follow along examples in the blog section. All you need is a linux based machine: https://pyimagesearch.com/blog/