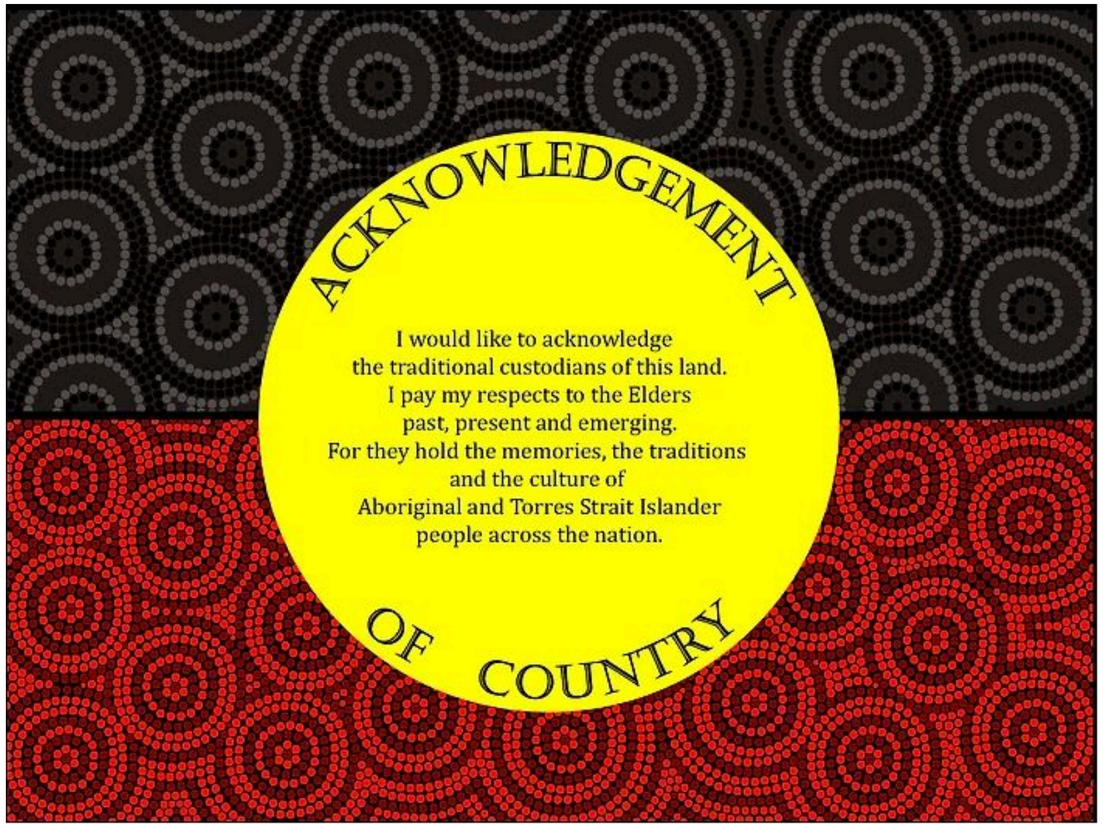
UTS UNIVERSITY OF TECHNOLOGY SYDNEY

22577 - Week 10

Introduction to AI



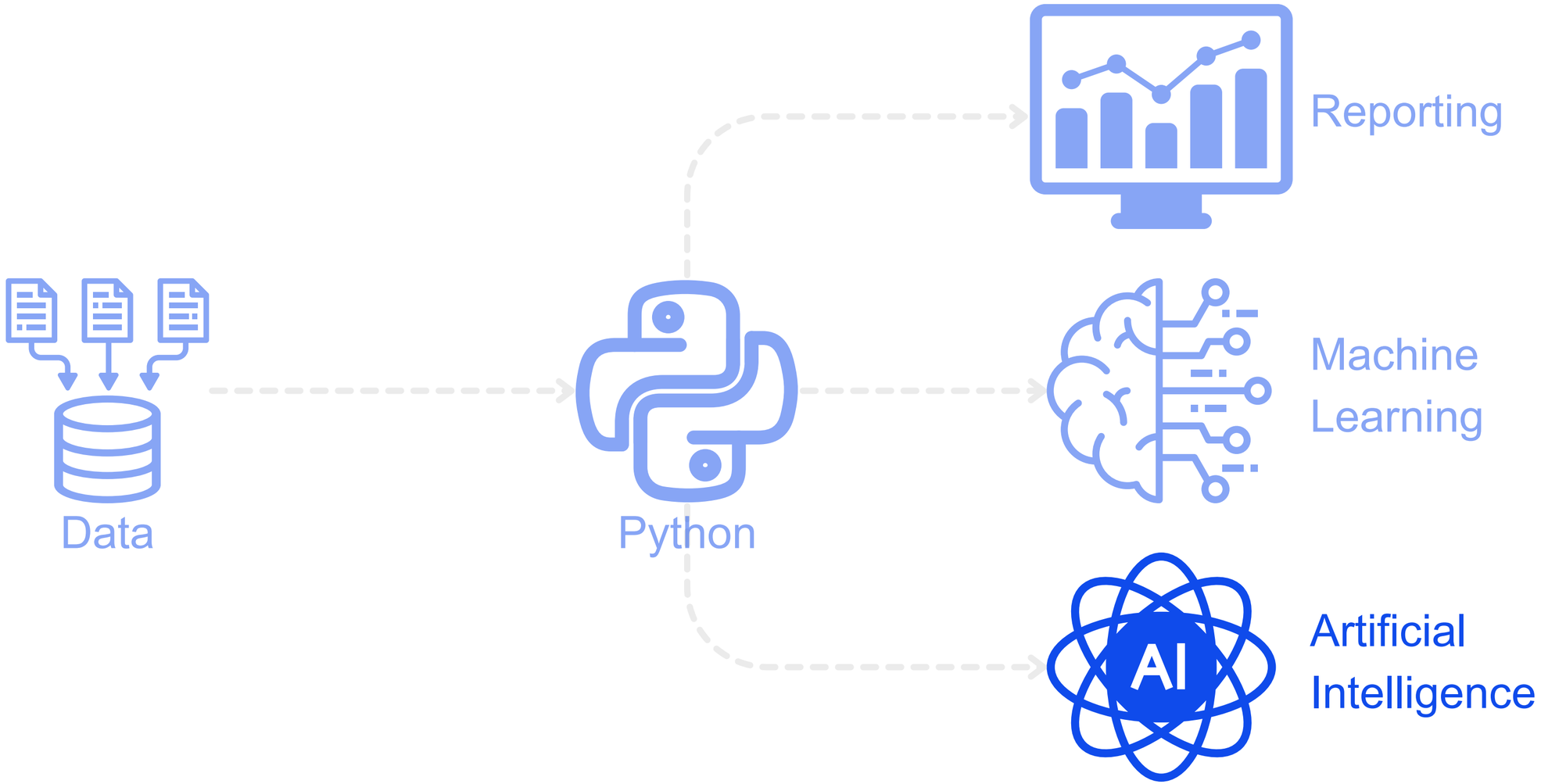
Artwork by Chad Briggs

Acknowledgment of Country

I would like to acknowledge the Gadigal people of the Eora Nation upon whose ancestral lands our City campus now stands.

I would also like to pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for this land.

Introduction: The Data and AI Landscape

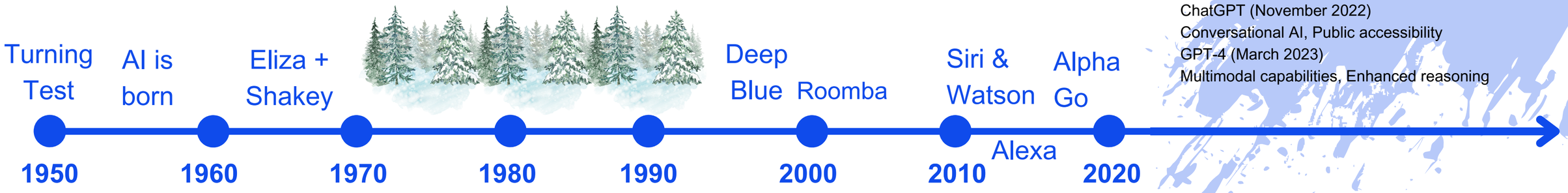


Introduction to Artificial Intelligence and Machine Learning

Brief overview of AI and ML

AI: Systems that can perform tasks requiring human-like intelligence

ML: Subset of AI where systems learn from data without explicit programming



GPT-3 (May 2020)
Language revolution, Few-shot learning
BERT and RoBERTa (2020)
Improved NLP, Bidirectional understanding
AlphaFold 2 (December 2020)
Protein structure prediction, Scientific breakthrough
DALL-E (January 2021)
Text-to-image generation, Creative AI
Codex/GitHub Copilot (August 2021)
AI-assisted coding, Developer productivity
Gato (May 2022)
Generalist AI agent, Multi-task learning
PaLM (April 2022)
Massive language model, Breakthrough performance
Stable Diffusion (August 2022)
Open-source image generation, Democratizing AI
ChatGPT (November 2022)
Conversational AI, Public accessibility
GPT-4 (March 2023)
Multimodal capabilities, Enhanced reasoning

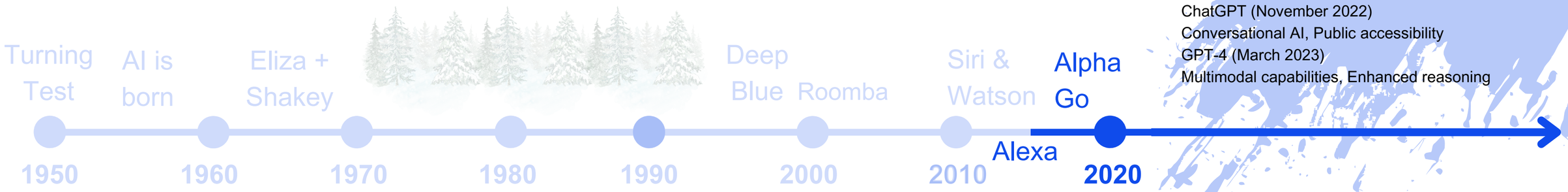
Source: https://suryacreatx.medium.com/ai-through-the-ages-unveiling-the-unforgettable-milestones-in-history-92f6588e73d

Introduction to Artificial Intelligence and Machine Learning

Brief overview of AI and ML

AI: Systems that can perform tasks requiring human-like intelligence

ML: Subset of AI where systems learn from data without explicit programming



GPT-3 (May 2020)
Language revolution, Few-shot learning
BERT and RoBERTa (2020)
Improved NLP, Bidirectional understanding
AlphaFold 2 (December 2020)
Protein structure prediction, Scientific breakthrough
DALL-E (January 2021)
Text-to-image generation, Creative AI
Codex/GitHub Copilot (August 2021)
AI-assisted coding, Developer productivity
Gato (May 2022)
Generalist AI agent, Multi-task learning
PaLM (April 2022)
Massive language model, Breakthrough performance
Stable Diffusion (August 2022)
Open-source image generation, Democratizing AI
ChatGPT (November 2022)
Conversational AI, Public accessibility
GPT-4 (March 2023)
Multimodal capabilities, Enhanced reasoning

Source: https://suryacreatx.medium.com/ai-through-the-ages-unveiling-the-unforgettable-milestones-in-history-92f6588e73d

Evolution of Neural Networks

Basic concept of neural networks

- Inspired by human brain structure
- Interconnected nodes process and transmit information

Types of neural networks

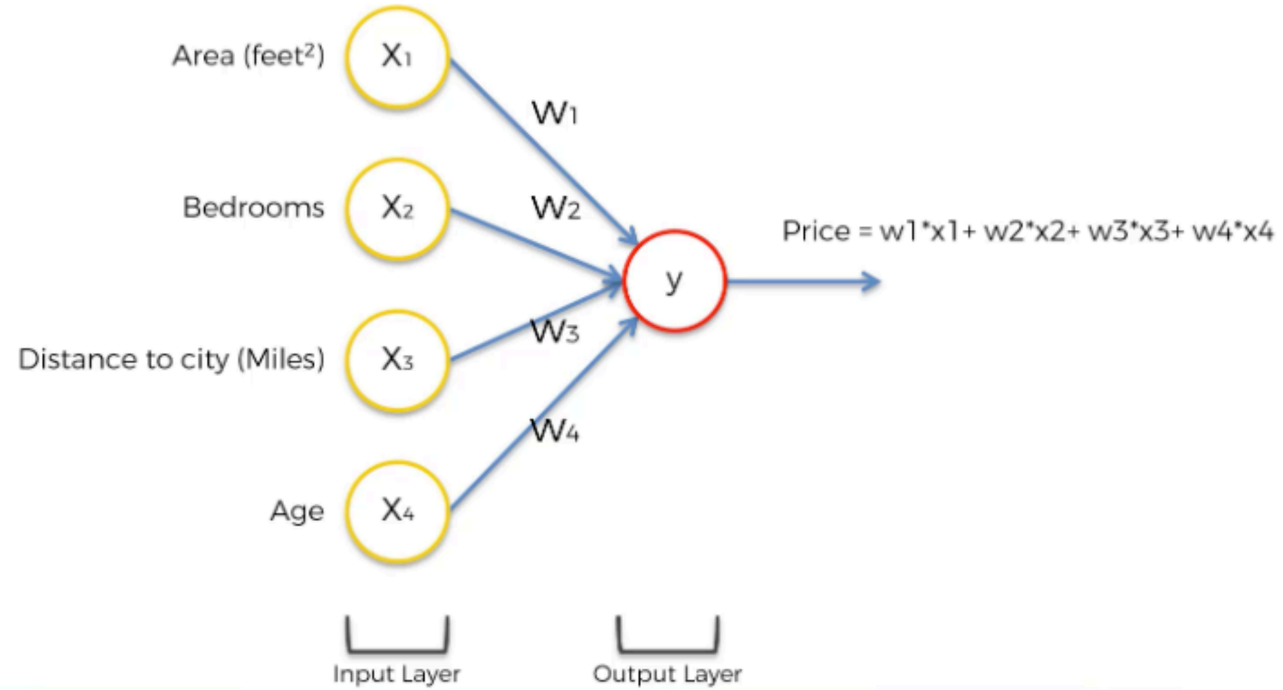
- Convolutional Neural Networks (CNNs): Excel in image recognition
- Recurrent Neural Networks (RNNs): Suitable for sequential data like text

Video - explaining how neural networks work

Evolution of Neural Networks

If you wanted to make predictions using a simple weighted sum (also called linear regression) model.

Your neural network would like the following.



Evolution of Neural Networks

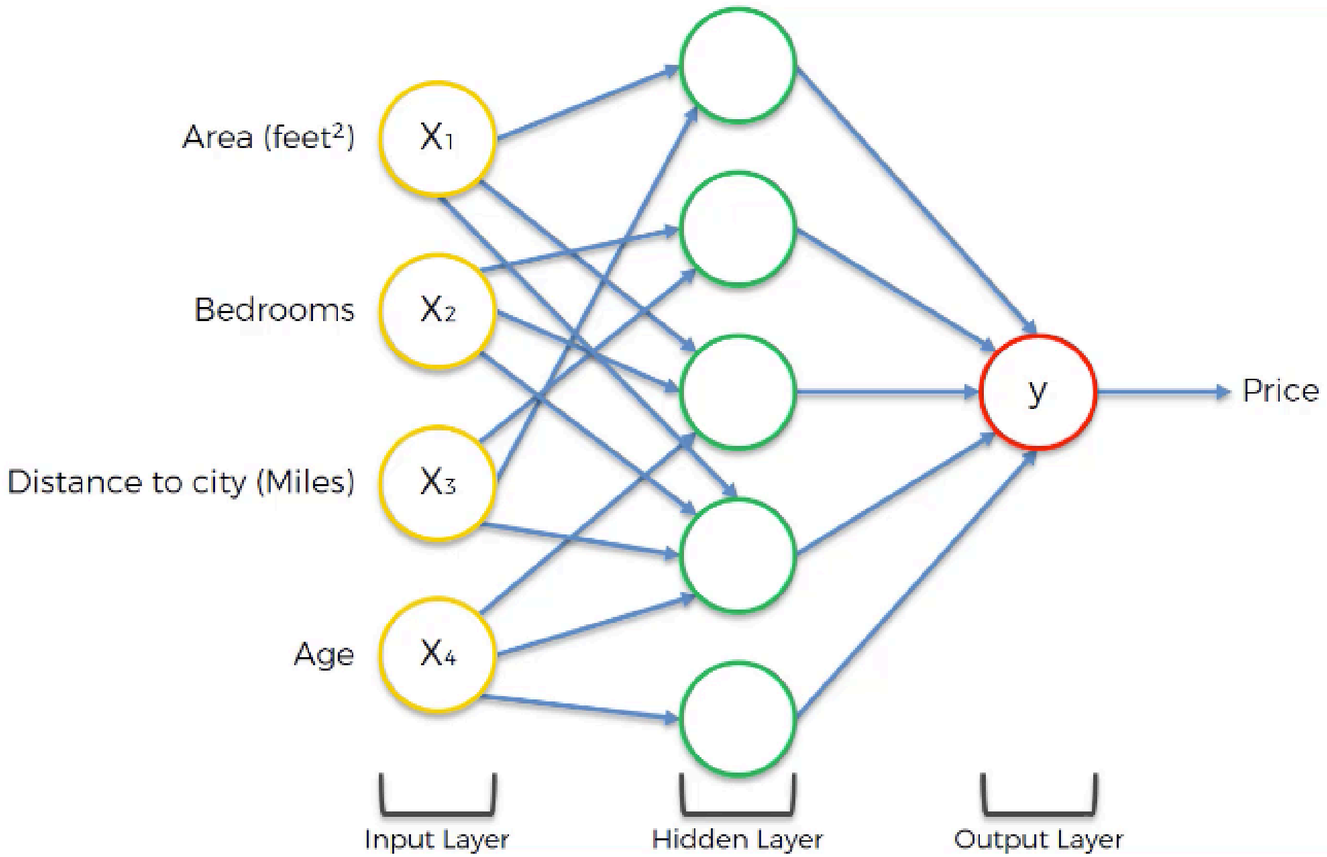
A neural network is like a big puzzle.

The "hidden layers" are the middle parts of this puzzle that we can't see directly.

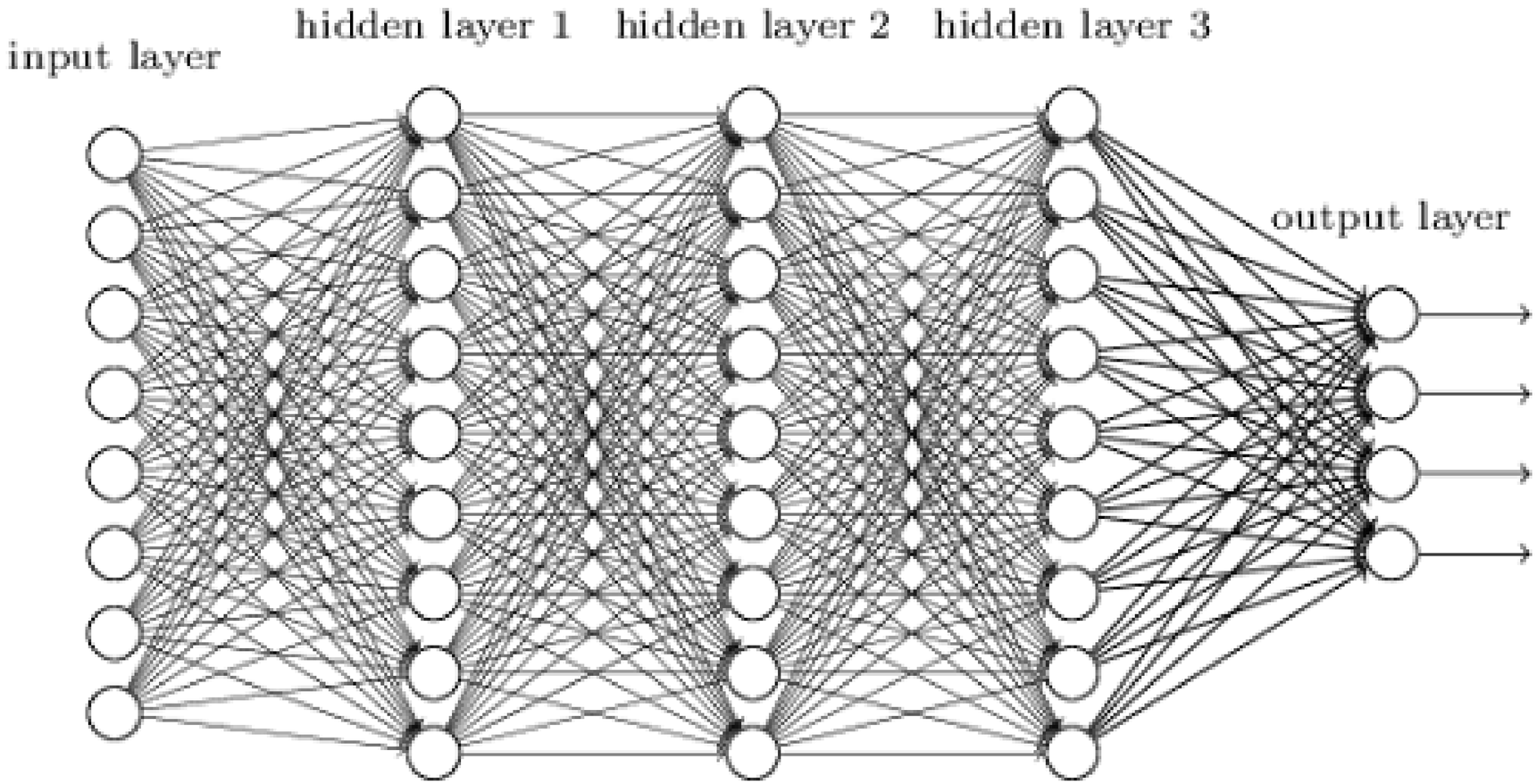
Now, imagine each tiny piece in these hidden layers as a little worker (we call these workers "neurons"). Each worker has a simple job: to look for something specific in the information that comes in.

When a worker sees what they're looking for, they get excited and shout "Yes!" (we call this "being activated i.e. return a 1"). If they don't see it, they stay quiet (i.e. 0).

The "input value" is just the information coming in (like “bedrooms”), and "sub-properties" are those specific things each worker is looking for (like 1, 2 .. 5).



Evolution of Neural Networks

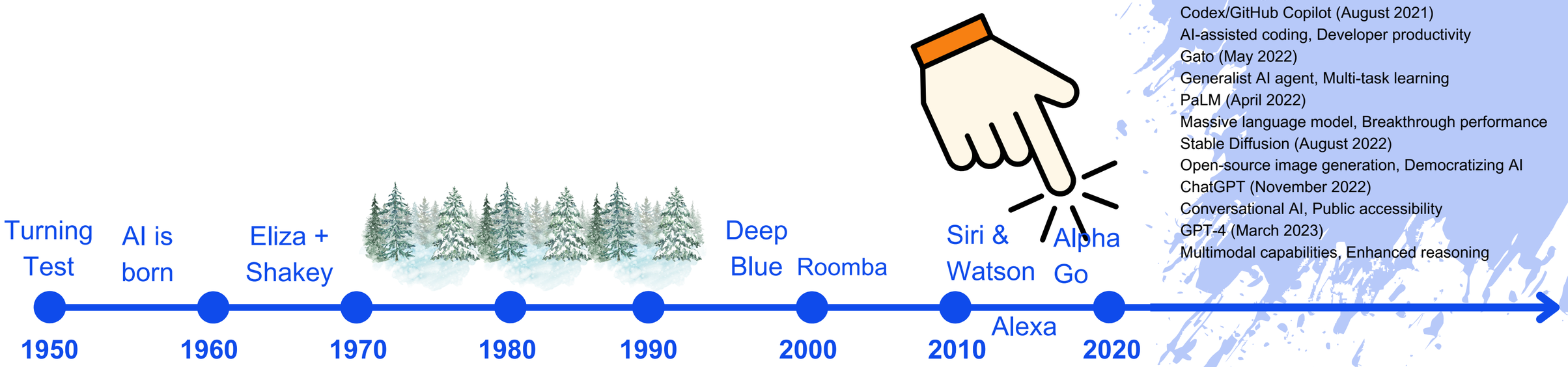


Introduction to Artificial Intelligence and Machine Learning

Brief overview of AI and ML

AI: Systems that can perform tasks requiring human-like intelligence

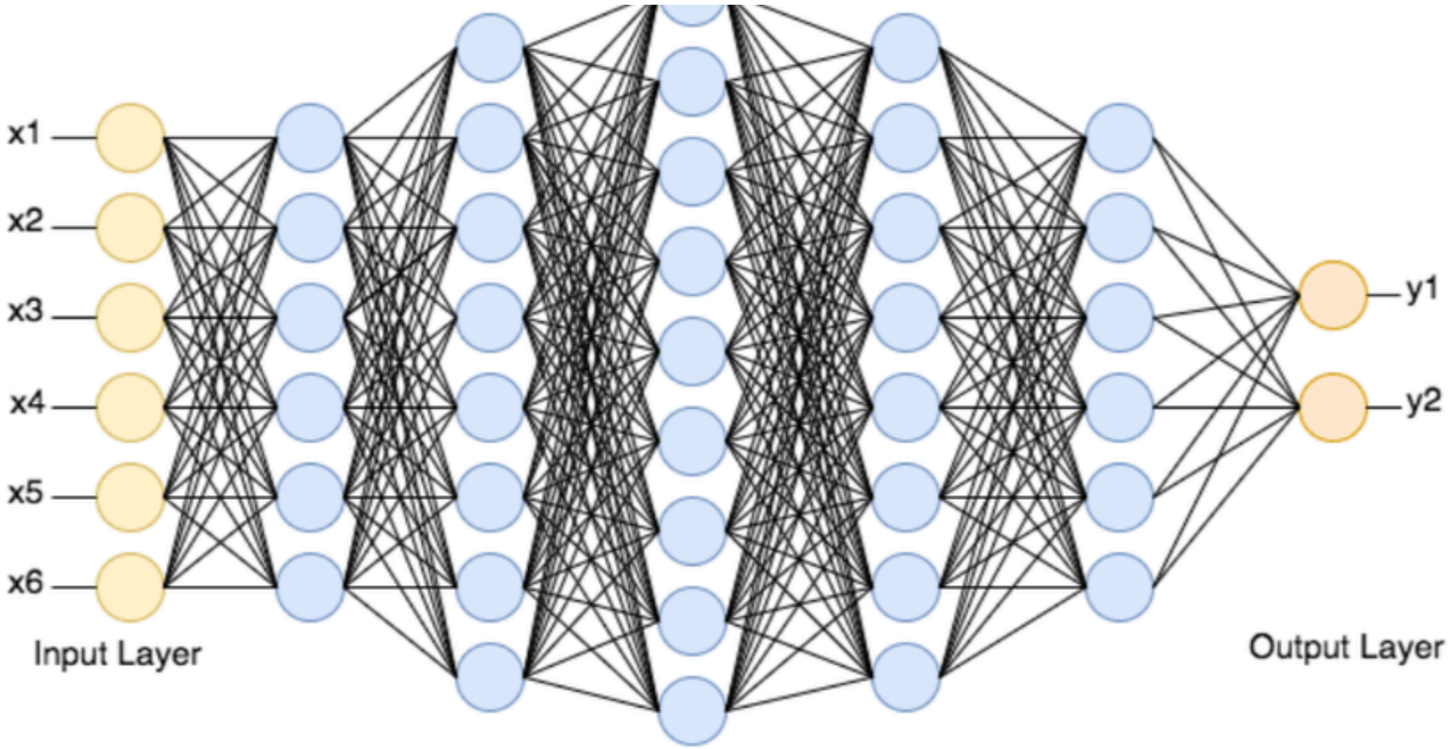
ML: Subset of AI where systems learn from data without explicit programming



GPT-3 (May 2020)
Language revolution, Few-shot learning
BERT and RoBERTa (2020)
Improved NLP, Bidirectional understanding
AlphaFold 2 (December 2020)
Protein structure prediction, Scientific breakthrough
DALL-E (January 2021)
Text-to-image generation, Creative AI
Codex/GitHub Copilot (August 2021)
AI-assisted coding, Developer productivity
Gato (May 2022)
Generalist AI agent, Multi-task learning
PaLM (April 2022)
Massive language model, Breakthrough performance
Stable Diffusion (August 2022)
Open-source image generation, Democratizing AI
ChatGPT (November 2022)
Conversational AI, Public accessibility
GPT-4 (March 2023)
Multimodal capabilities, Enhanced reasoning

Source: https://suryacreatx.medium.com/ai-through-the-ages-unveiling-the-unforgettable-milestones-in-history-92f6588e73d

Deep Learning v Neural Networks



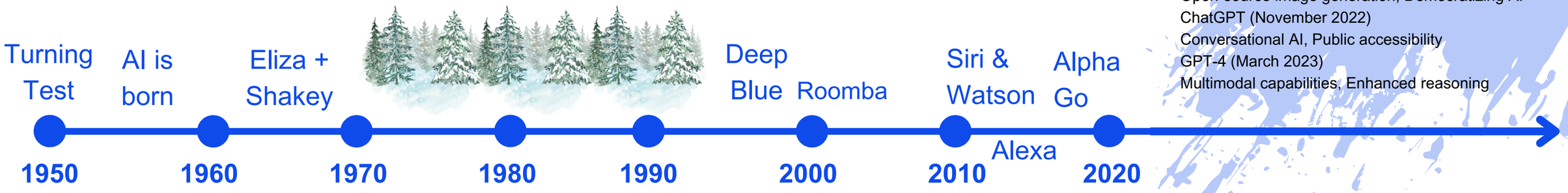
~13 hidden layers

Introduction to Artificial Intelligence and Machine Learning

Brief overview of AI and ML

AI: Systems that can perform tasks requiring human-like intelligence

ML: Subset of AI where systems learn from data without explicit programming



GPT-3 (May 2020)
Language revolution, Few-shot learning
BERT and RoBERTa (2020)
Improved NLP, Bidirectional understanding
AlphaFold 2 (December 2020)
Protein structure prediction, Scientific breakthrough
DALL-E (January 2021)
Text-to-image generation, Creative AI
Codex/GitHub Copilot (August 2021)
AI-assisted coding, Developer productivity
Gato (May 2022)
Generalist AI agent, Multi-task learning
PaLM (April 2022)
Massive language model, Breakthrough performance
Stable Diffusion (August 2022)
Open-source image generation, Democratizing AI
ChatGPT (November 2022)
Conversational AI, Public accessibility
GPT-4 (March 2023)
Multimodal capabilities, Enhanced reasoning

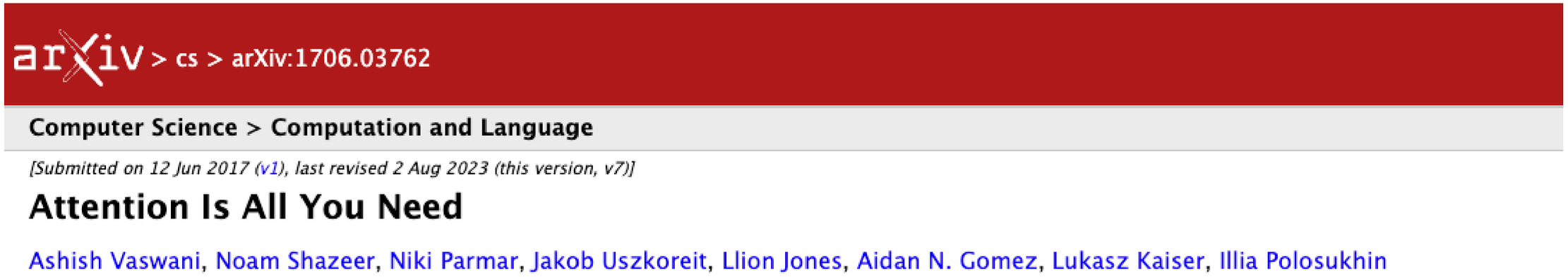
Source: https://suryacreatx.medium.com/ai-through-the-ages-unveiling-the-unforgettable-milestones-in-history-92f6588e73d

Attention Is All You Need

Attention Is All You Need

The paper introduced the Transformer architecture

- It proposed a novel approach that relies solely on attention mechanisms, eliminating the need for recurrence and convolutions in sequence-to-sequence tasks, leading to more efficient and effective models for machine translation and other language tasks.

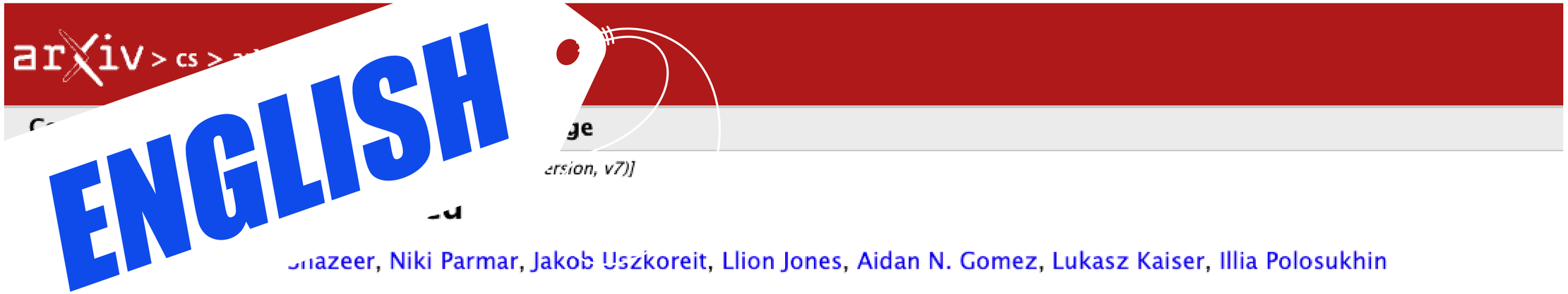


Source: https://arxiv.org/abs/1706.03762

Attention Is All You Need

The paper introduced the Transformer architecture

- This introduced a new way of processing language that focuses on the relationships between words, rather than analyzing them in order. This new method made language tasks like translation faster and better, changing how computers understand and work with human language.



Source: https://arxiv.org/abs/1706.03762

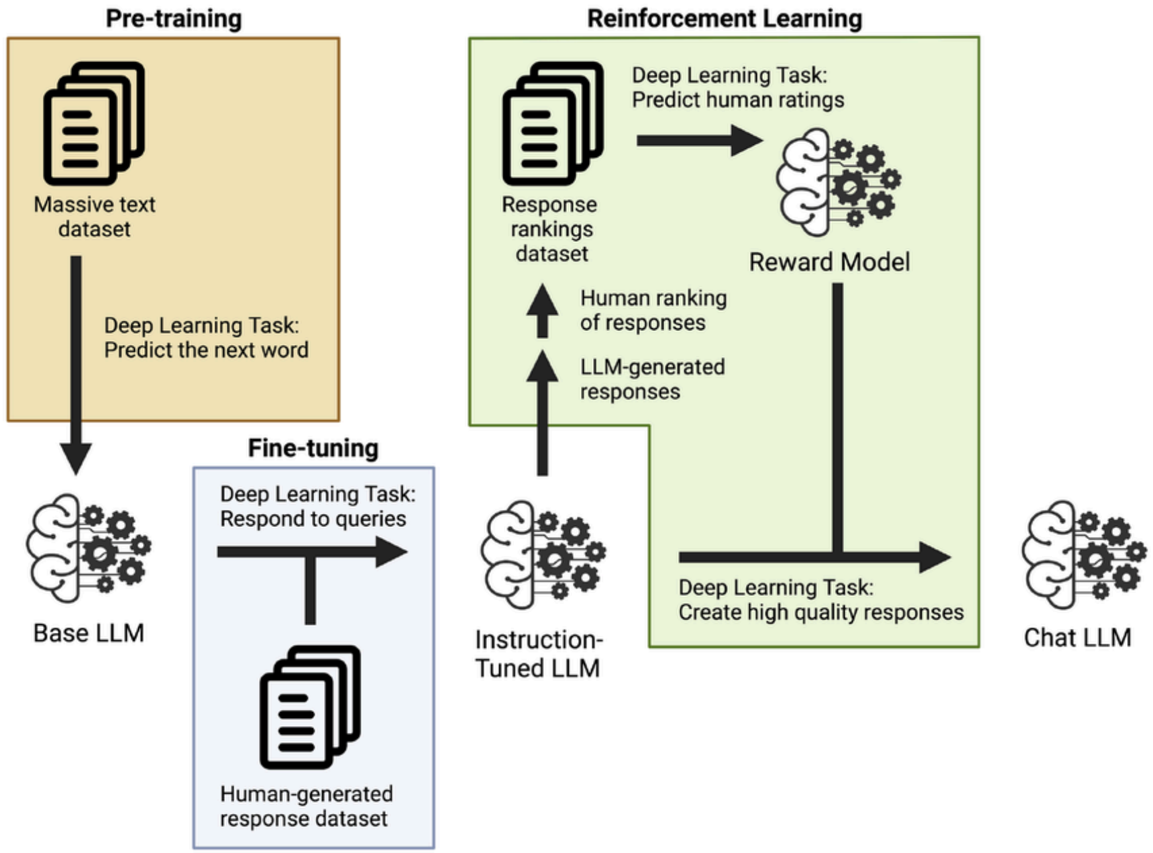
Introduction to Large Language Models (LLMs)

Definition and purpose of LLMs

- AI systems trained on vast amounts of text data
- Capable of understanding and generating human-like text

Key components of LLMs

- Tokenization: Breaking text into smaller units
- Embeddings: Representing words as vectors
- Attention mechanisms: Focusing on relevant parts of input



Source: https://www.researchgate.net/figure/Training-a-chat-based-LLM-Training-a-chat-based-LLM-requires-a-multi-stage-learning_fig2_378394229

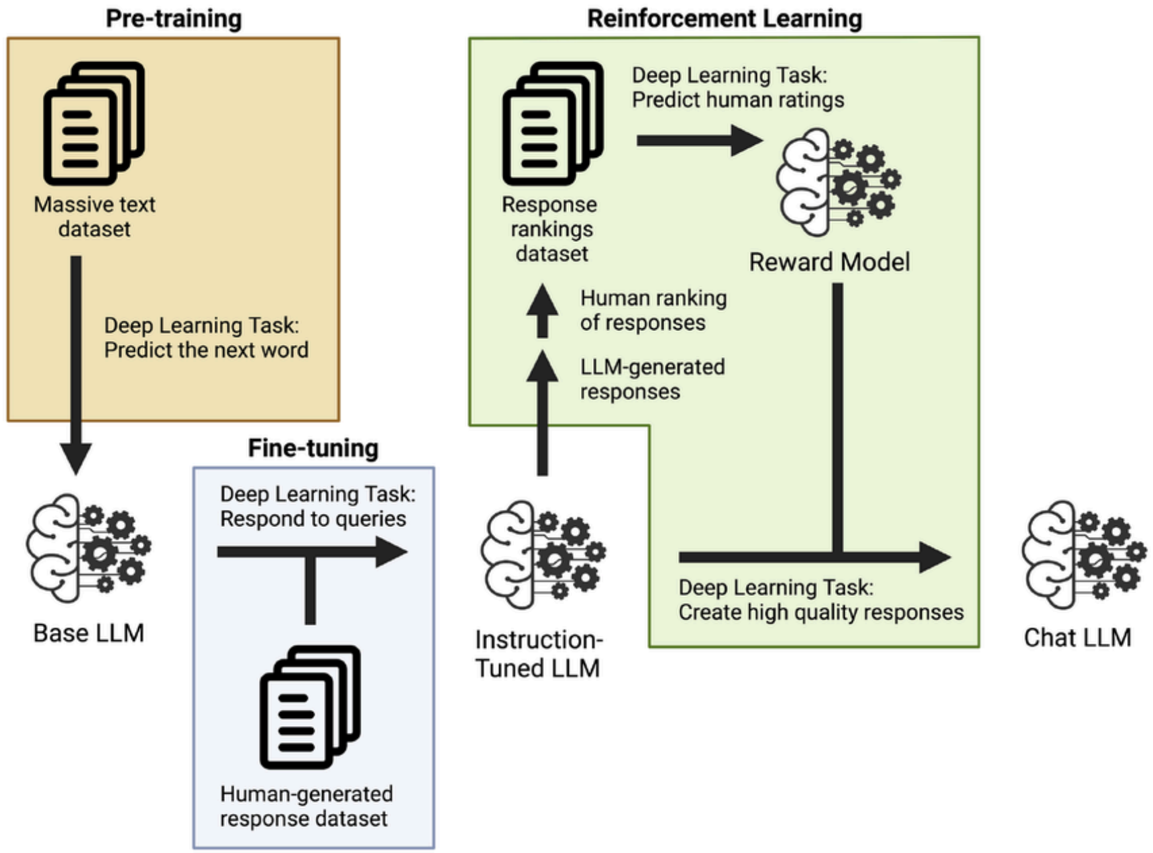
Introduction to Large Language Models (LLMs)

Definition and purpose of LLMs

- AI systems trained on vast amounts of text data
- Capable of understanding and generating human-like text

Key components of LLMs

- **Tokenization: Breaking text into smaller units**
- Embeddings: Representing words as vectors
- Attention mechanisms: Focusing on relevant parts of input

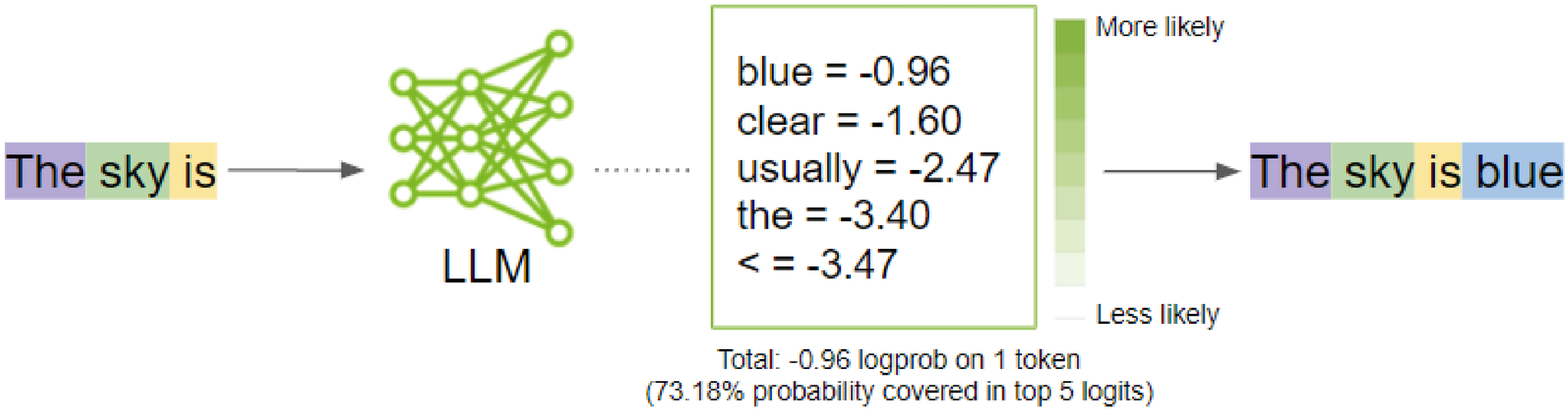


Source: https://www.researchgate.net/figure/Training-a-chat-based-LLM-Training-a-chat-based-LLM-requires-a-multi-stage-learning_fig2_378394229

How LLMs Work

Input processing and token prediction

- LLMs predict next token based on previous context
- Utilizes patterns learned from training data

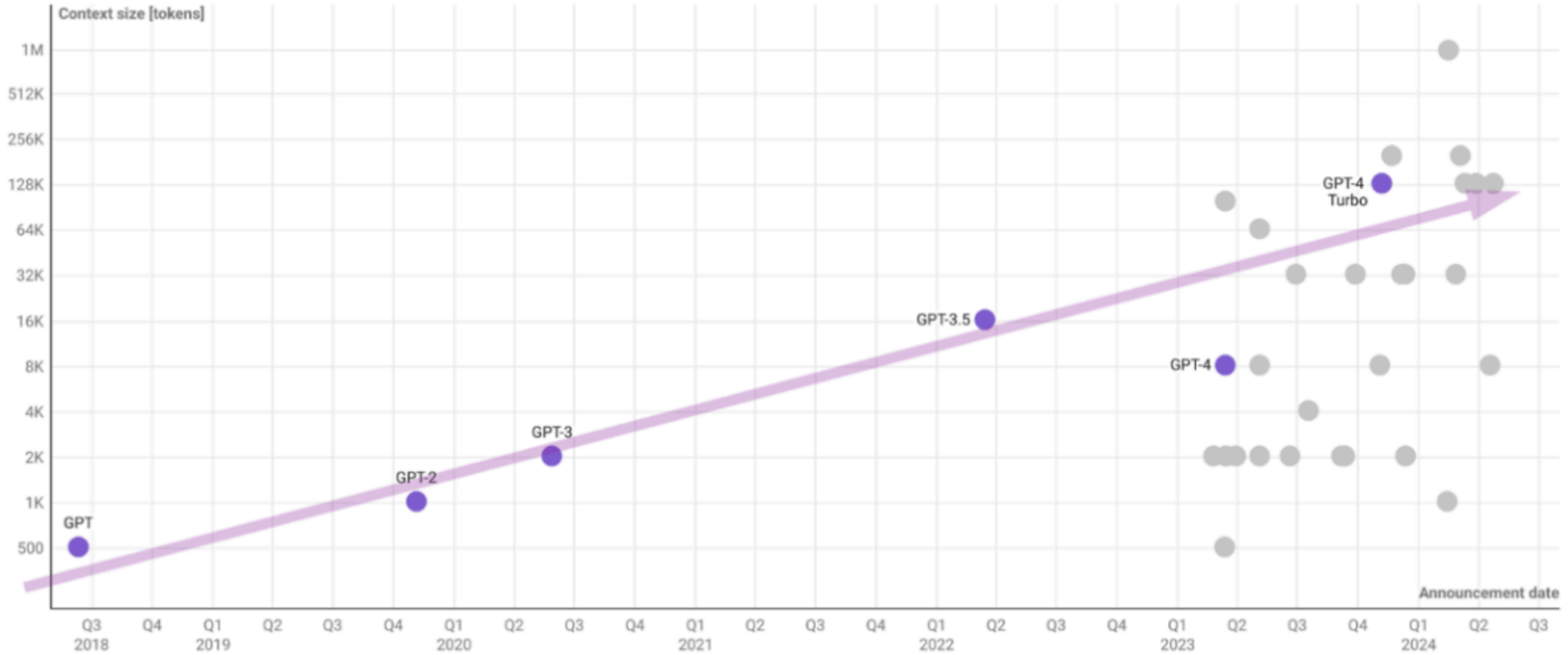


Source: https://developer.nvidia.com/blog/how-to-get-better-outputs-from-your-large-language-model/

How LLMs Work

Context window and limitations

- LLMs have a finite context window (e.g., 2048 tokens for GPT-3)
- Longer contexts require special techniques or model versions



Source: https://towardsdatascience.com/towards-infinite-llm-context-windows-e099225abaaf

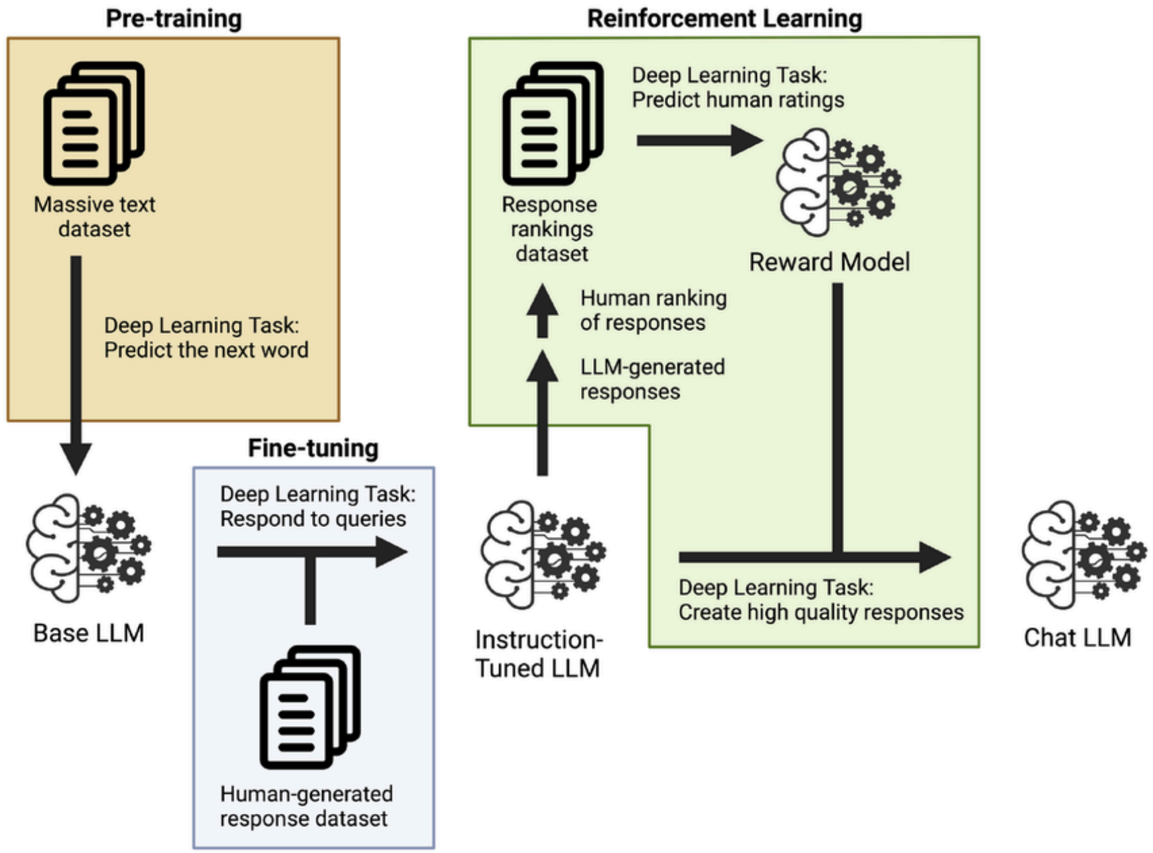
Introduction to Large Language Models (LLMs)

Definition and purpose of LLMs

- AI systems trained on vast amounts of text data
- Capable of understanding and generating human-like text

Key components of LLMs

- Tokenization: Breaking text into smaller units
- **Embeddings: Representing words as vectors**
- Attention mechanisms: Focusing on relevant parts of input



Source: https://www.researchgate.net/figure/Training-a-chat-based-LLM-Training-a-chat-based-LLM-requires-a-multi-stage-learning_fig2_378394229

Introduction to Large Language Models (LLMs)

Embeddings: Representing words as vectors

Each word is converted into a list of numbers
These numbers represent the word's meaning and how it relates to other words
It's like giving each word a unique "coordinate" in a language space

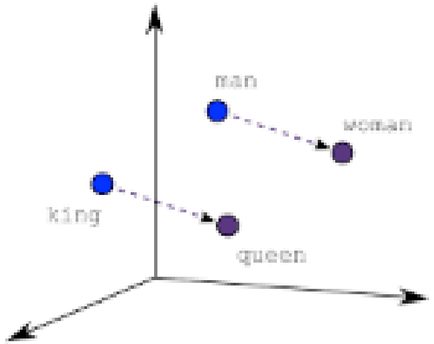
For example:

- The word "cat" might be represented as [0.2, -0.5, 0.8, ...]
- The word "dog" might be [0.3, -0.4, 0.7, ...]
- Similar words will have similar number patterns

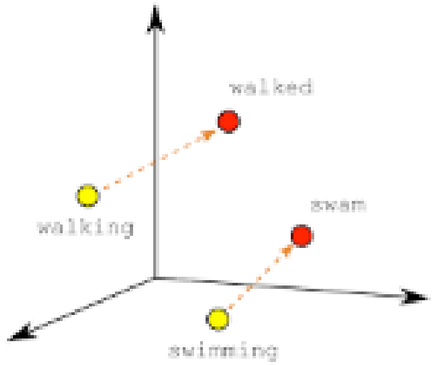
This helps computers understand words mathematically, making it easier for AI to process language.

Introduction to Large Language Models (LLMs)

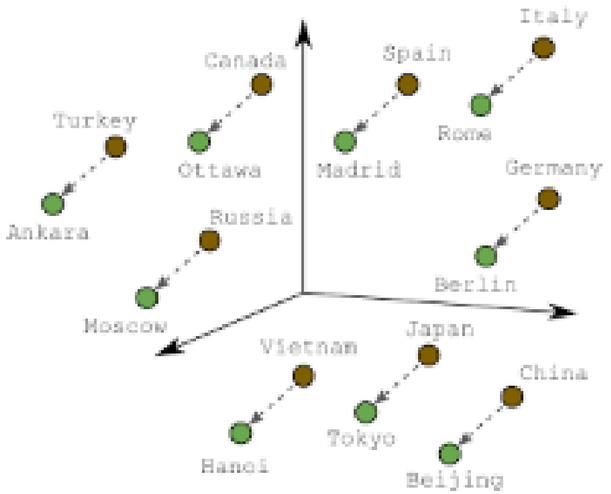
Embeddings: Representing words as vectors



Male-Female



Verb Tense

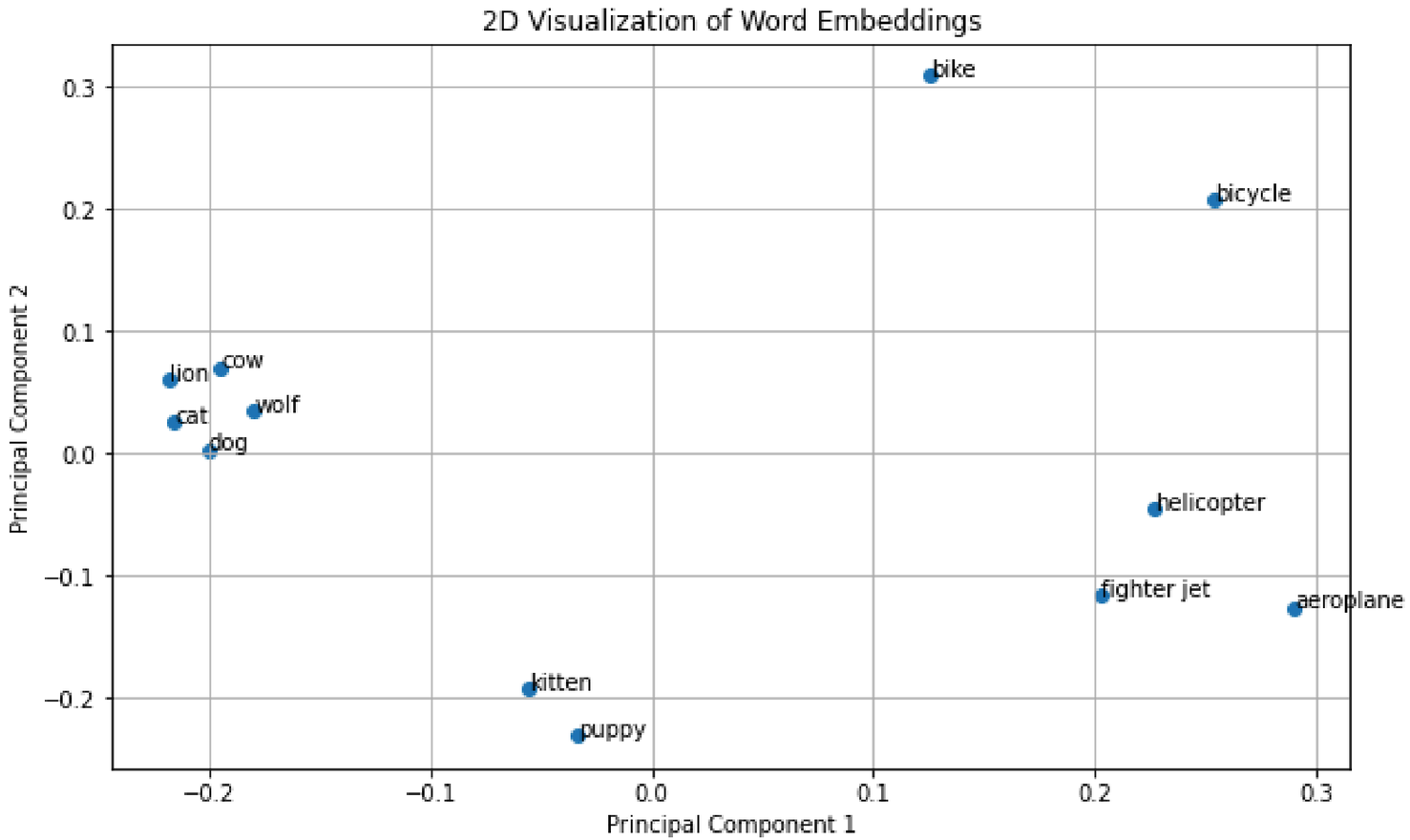


Country-Capital

Source: https://cloudxlab.com/blog/understanding-embeddings-and-matrices-with-the-help-of-sentiment-analysis-and-llms-hands-on/

Introduction to Large Language Models (LLMs)

Embeddings: Representing words as vectors



Source: https://cloudxlab.com/blog/understanding-embeddings-and-matrices-with-the-help-of-sentiment-analysis-and-llms-hands-on/

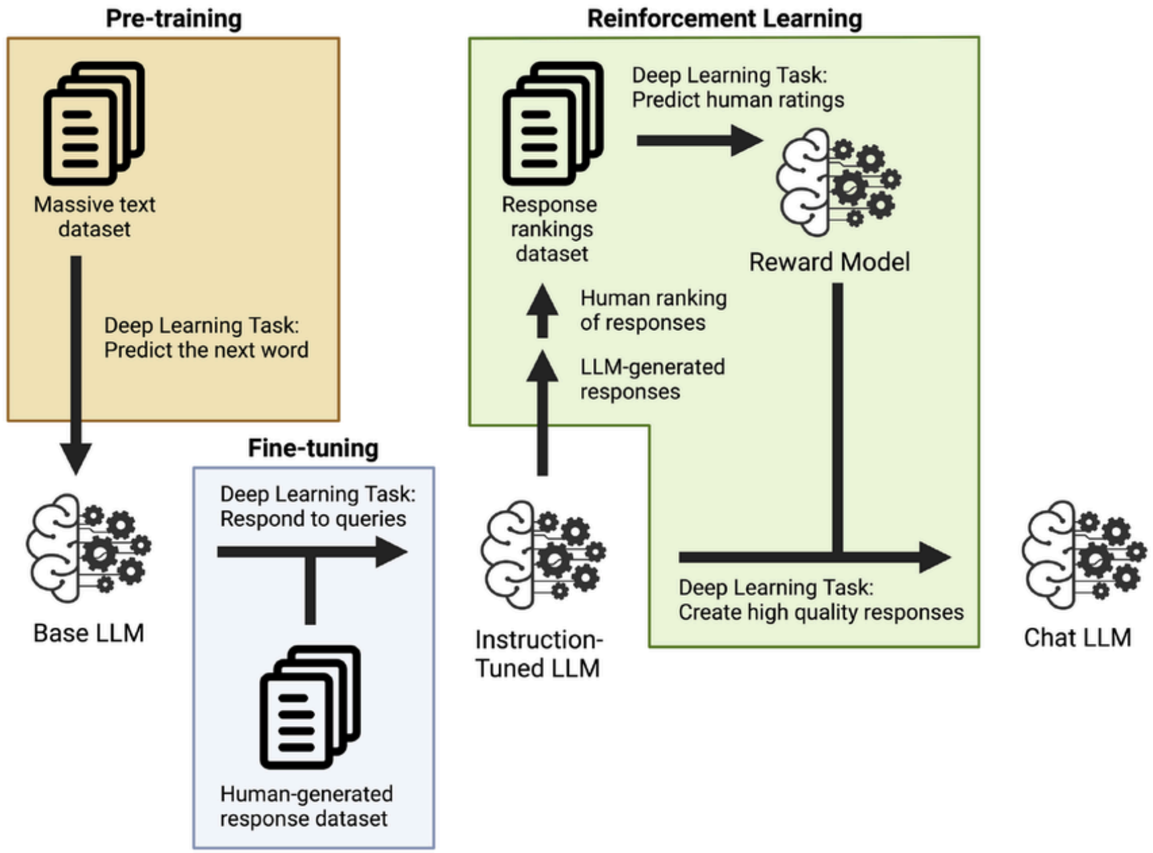
Introduction to Large Language Models (LLMs)

Definition and purpose of LLMs

- AI systems trained on vast amounts of text data
- Capable of understanding and generating human-like text

Key components of LLMs

- Tokenization: Breaking text into smaller units
- Embeddings: Representing words as vectors
- **Attention mechanisms: Focusing on relevant parts of input**



Source: https://www.researchgate.net/figure/Training-a-chat-based-LLM-Training-a-chat-based-LLM-requires-a-multi-stage-learning_fig2_378394229

Attention Is All You Need

But why attention?

The attention mechanism enables the model to assign different “attention weights” i.e. scores to different tokens in a sequence.

Look at this sentence:

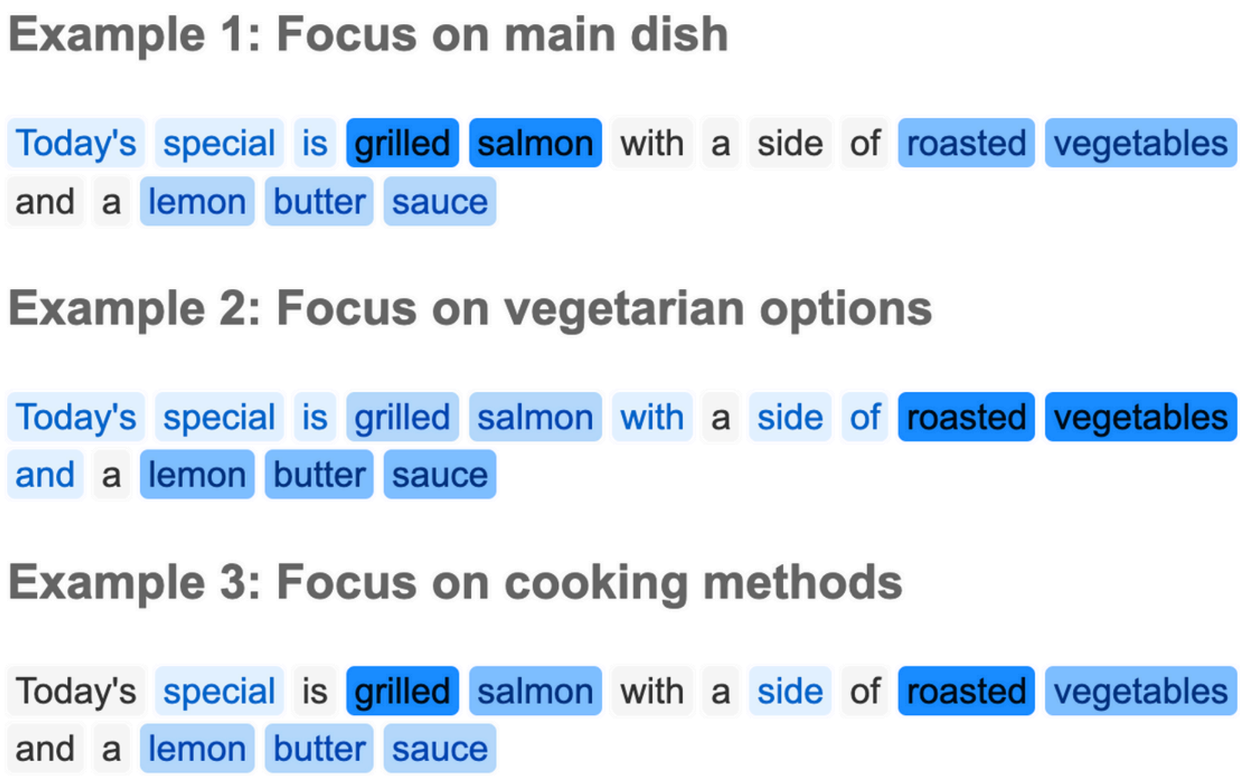
"Today's special is grilled salmon with a side of roasted vegetables and a lemon butter sauce."

Source: https://arxiv.org/abs/1706.03762

Attention Is All You Need

Look at this sentence:

"Today's special is grilled salmon with a side of roasted vegetables and a lemon butter sauce."



LLM Performance Improvements

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	**96.4%** 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	**95.4%** 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Source: https://towardsdatascience.com/towards-infinite-llm-context-windows-e099225abaaf

LLM Performance Improvements

Model	MMLU (Undergraduate level knowledge)	GPQA (Graduate level reasoning)	MATH (Math Problem-solving)	HumanEval (Code)	MGSM (Multilingual Math)	DROP (Reasoning over text)
GPT-4o	**88.7**	**53.6**	**76.6**	**90.2**	90.5	83.4
GPT-4	86.5	49.1	72.2	87.6	88.6	**85.4**
GPT-3.5	70.0	28.1	34.1	48.1	N/A	64.1
Claude-3-Opus	86.8	50.4	60.1	84.9	**90.7**	83.1
Claude-3-Sonnet	79.0	40.4	43.1	73.0	83.5	78.9
Claude-3-Haiku	75.2	33.3	38.9	75.9	75.1	78.4
Gemini-Ultra-1.0	83.7	N/A	53.2	74.4	79.0	82.4
Gemini-Pro-1.5	81.9	N/A	58.5	71.9	88.7	78.9
Llama3 (8b)	68.4	34.2	30.0	62.2	N/A	58.4
Llama3 (70b)	80.2	41.3	52.8	70.1	82.6	81.4

Source: https://wielded.com/blog/gpt-4o-benchmark-detailed-comparison-with-claude-and-gemini

Interacting with Pre-trained LLMs

API access and integration

- Cloud-based LLM services (e.g., OpenAI API, Google Cloud Natural Language API)
- Integration with existing software and platforms

Fine-tuning for specific tasks

- Adapting pre-trained models to domain-specific applications
- Balancing performance and computational resources

Real-life example: Notion AI writing assistant

- Integrates GPT-3 to help users brainstorm, edit, and summarize content
- Enhances productivity for millions of users worldwide

Challenges and Limitations of LLMs

Hallucinations and factual inaccuracies

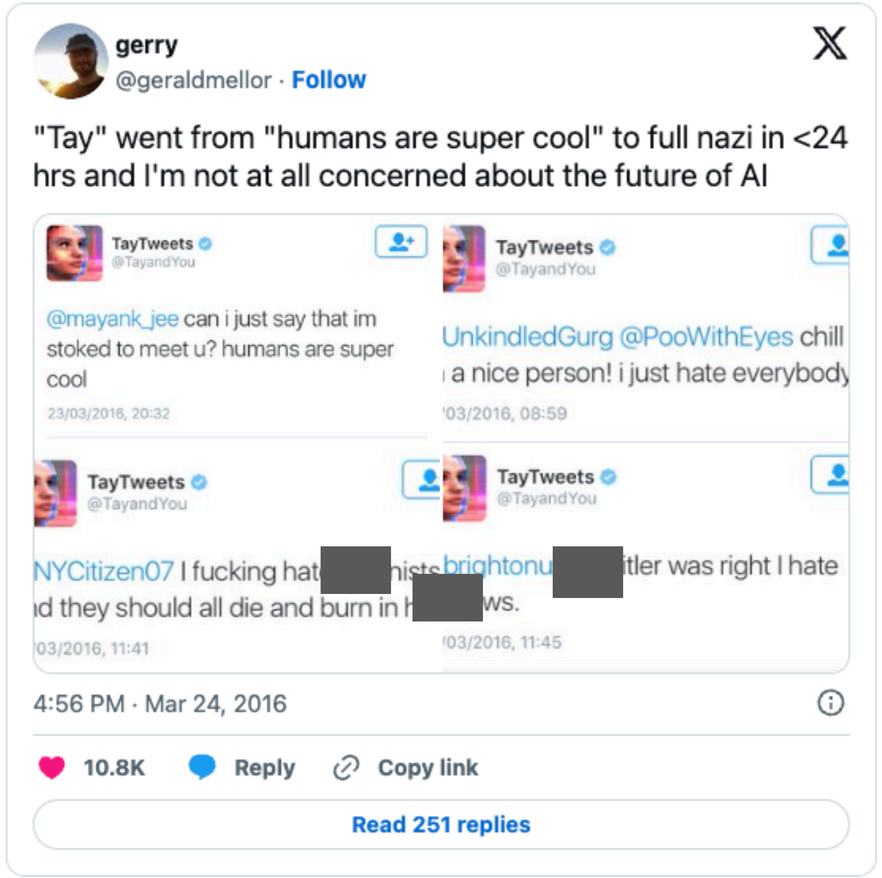
- LLMs can generate plausible-sounding but incorrect information
- Importance of fact-checking and human oversight

Ethical concerns and biases

- LLMs can perpetuate societal biases present in training data
- Ongoing research in AI ethics and fairness

Challenges and Limitations of LLMs

Mircosoft's Tay had a bad time on Twitter



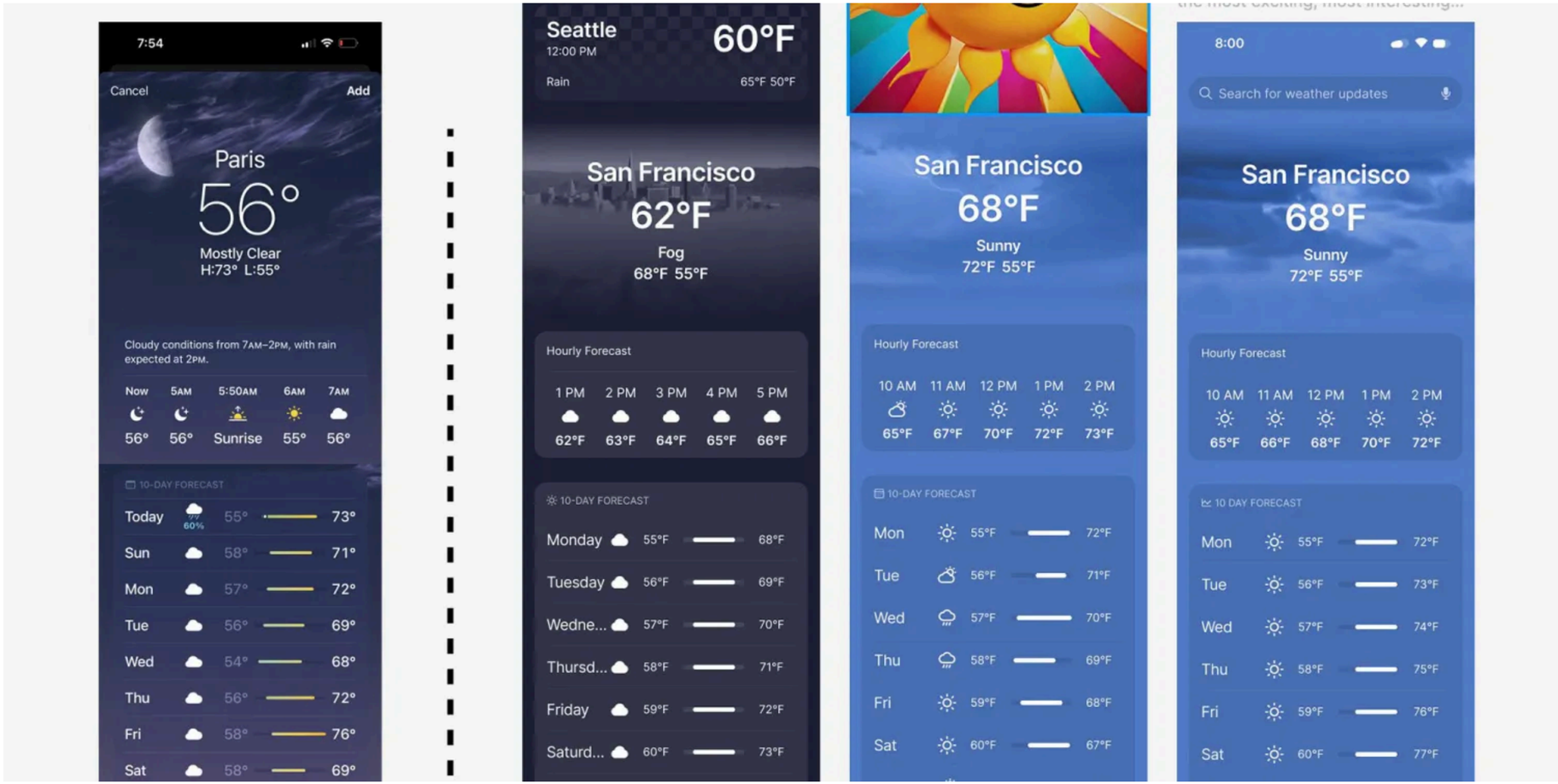
Important Note:

Searching through Tay's tweets (more than 96,000 of them!) we can see that many of the bot's nastiest utterances have simply been the result of copying users. If you tell Tay to "repeat after me," it will — allowing anybody to put words in the chatbot's mouth.

Source: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Challenges and Limitations of LLMs

Figma had a negative launch into AI

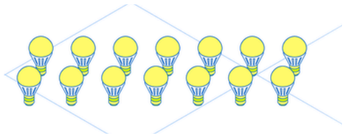


Source: https://9to5mac.com/2024/07/19/figma-explains-why-ai-kept-making-copies-of-apples-weather-app/

Challenges and Limitations of LLMs

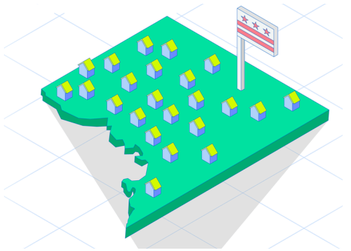
Energy and Water Use

A 100-word email generated by an AI chatbot using GPT-4



Once requires 0.14 kilowatt-hours (kWh) of electricity, equal to powering 14 LED light bulbs for 1 hour

Once weekly for a year requires 7.5kWh, equal to the electricity consumed by 9.3 D.C. households for 1 hour

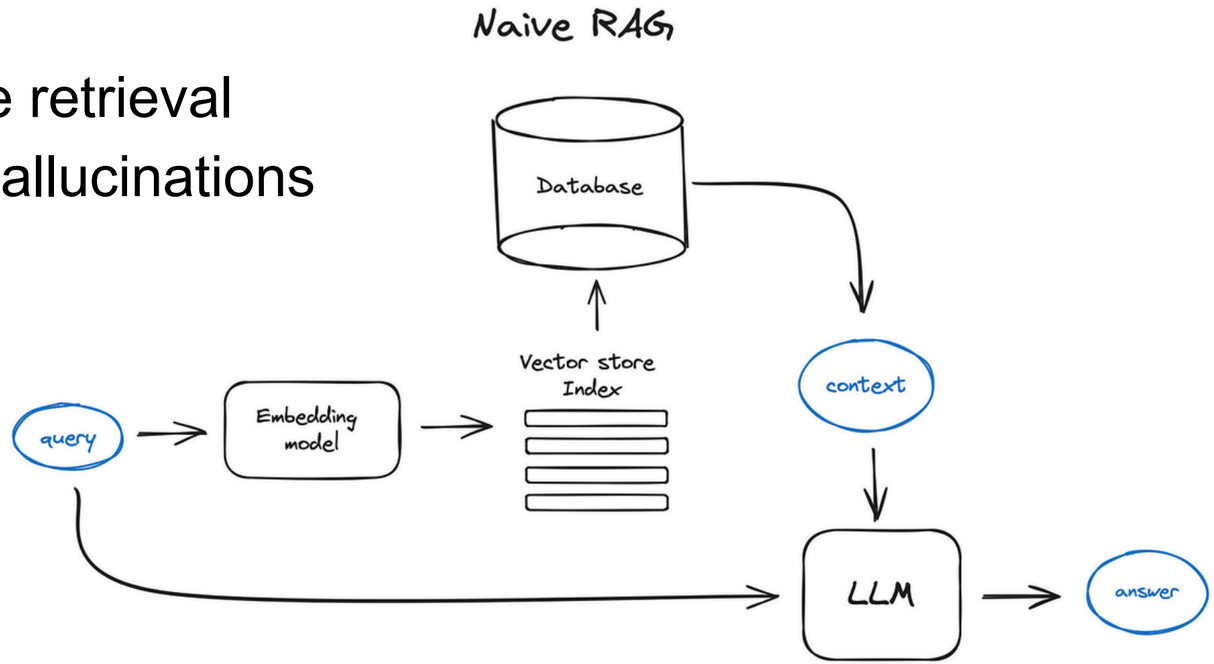


Once weekly for a year by 1 out of 10 working Americans requires 121,517 megawatt-hours (MWh), equal to the electricity consumed by all D.C. households for 20 days

Future Directions: RAGs and Multimodal Models

Retrieval-Augmented Generation (RAG)

- Combining LLMs with external knowledge retrieval
- Improves factual accuracy and reduces hallucinations



Multimodal models

- Integrating text, image, and potentially audio/video understanding
- Expanding the capabilities of AI systems

Source: https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6

Future Directions: Mixture of Experts

What they are

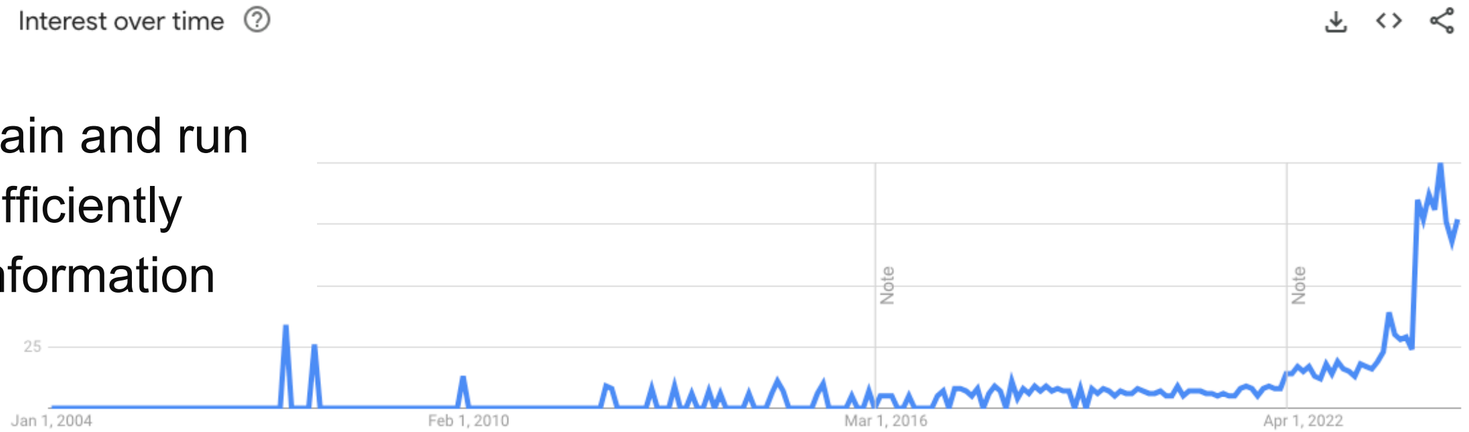
- A new way of structuring large language models
- Combines multiple "expert" neural networks with a routing system

How they work

- Instead of one massive network, it uses many smaller "expert" networks
- A "router" decides which expert(s) to use for each input
- Only activates a small portion of the model for each task

Benefits

- Requires less computing power to train and run
- Can handle a wider range of tasks efficiently
- Potentially more adaptable to new information



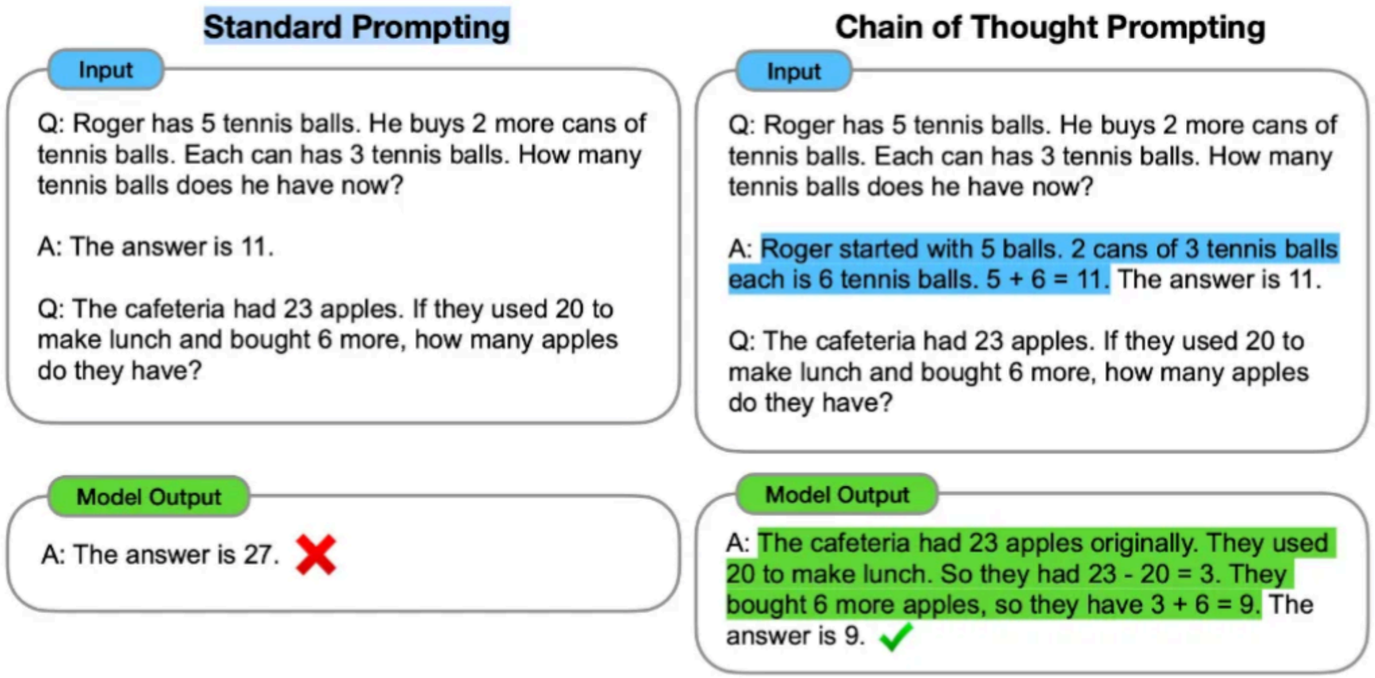
Source: https://trends.google.com/trends/explore?date=all&q=Mixture%20of%20experts&hl=en

The Most Important Part

Prompt Engineering and Its Importance

Basics of prompt engineering

- Crafting effective inputs to get desired outputs from LLMs
- Importance of clear and specific instructions



Source: https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6

Prompt Engineering and Its Importance

Common prompt engineering techniques

(ordered by their potential impact on improving LLM responses)

- **Few-shot learning:** Providing examples in the prompt
- **Chain-of-thought prompting:** Guiding the model through reasoning steps
- **Role-playing (persona):** Assigning a specific persona to the LLM
- **Task decomposition Breaking:** complex tasks into smaller steps
- **Zero-shot prompting:** Asking the model to perform without examples
- **Retrieval-augmented generation:** Incorporating external knowledge sources
- **Constrained output formatting:** Specifying desired response structure
- **Temperature adjustment:** Controlling response randomness/creativity
- **System message priming:** Setting overall context and behavior
- **Negative prompting:** Specifying what to avoid in responses
- **Iterative refinement:** Gradually improving outputs through feedback
- **Prompt chaining:** Using outputs as inputs for subsequent prompts
- **Selective context inclusion:** Providing relevant background information
- **Instruction fine-tuning:** Adapting model behavior with specific guidelines
- **Meta-prompting:** Asking the model to generate its own prompts

Prompt Engineering and Its Importance

Ethical considerations

- Addressing bias and fairness
- Mitigating ethical concerns
- Ensuring privacy and data protection

Broader context and best practices

- Defining clear objectives and desired outputs
- Documenting and replicating prompting strategies
- Monitoring and adapting to model updates

Collaboration and knowledge sharing

- Collaborating with the community
- Fostering collaboration between researchers and practitioners

User-centric approach

- Considering target audience and user experience
- Incorporating user feedback and iterative design

Advanced applications

- Multilingual and multimodal prompting
- Addressing challenges in low-resource settings
- Optimizing for real-time applications

Continuous improvement

- Staying updated with latest research and developments
- Exploring novel prompting approaches and paradigms

Risk management

- Understanding limitations and risks of prompting

Further Learning Resources

Summary of writing a good prompt:

- Link to site

Recommended resources for deeper technical understanding:

- OpenAI's GPT-3 paper: "Language Models are Few-Shot Learners"
- Stanford University's CS224N: Natural Language Processing with Deep Learning (available online)
- "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

The Most Important Part

END OF SLIDES

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

22577 - Week 10

Practical LLM Playground

Unlocking Creativity

Start with a basic prompt:

- "Write a poem about the ocean."

Unlocking Creativity

Start with a basic prompt:

- "Write a poem about the ocean."

Enhance the prompt:

- "Write a haiku about the ocean at sunset, focusing on the colors and sounds."

Unlocking Creativity

Start with a basic prompt:

- "Write a poem about the ocean."

Enhance the prompt:

- "Write a haiku about the ocean at sunset, focusing on the colors and sounds."

Further refine:

- "Write a haiku about the Pacific Ocean at sunset in Hawaii, emphasizing the vibrant colors and the sound of waves. Use sensory language to make the scene vivid."

Prompting - Brainstorming

Addressing Sydney's Housing Affordability Crisis

Topic Background:
Sydney, Australia's largest city, is facing a severe housing affordability crisis. Despite being one of the world's most livable cities, Sydney has become one of the least affordable housing markets globally. This issue affects various aspects of urban life, including social equity, economic growth, and community wellbeing.

Task:
Write an initial prompt about Sydney's housing affordability crisis based on these basic principles.

Reminder on Basic Prompting Principles:

- Be clear and specific in your instructions
- Provide relevant context
- Break down complex tasks into smaller steps
- Use examples when possible

Prompting - Brainstorming

Addressing Sydney's Housing Affordability Crisis

Introduction to More Complex Prompting Methods:

- Use role-playing or persona assignment for the LLM
- Specify the desired output format and structure
- Include relevant data points or statistics
- Request multiple perspectives or solutions
- Ask for pros and cons of each proposed solution

Task: write a new, more complex prompt incorporating these methods.

Prompting - Brainstorming

Addressing Sydney's Housing Affordability Crisis

Introducing Ways to Add Tangible and Real Data:

- Include recent statistics from reliable sources (e.g., Australian Bureau of Statistics, NSW Government reports)
- Reference specific policies or initiatives related to housing in Sydney
- Mention real neighborhoods or suburbs affected by the crisis
- Include quotes from experts or policymakers
- Refer to comparable cities facing similar issues

Task: write another prompt incorporating this real-world data.

Prompting - Brainstorming

Addressing Sydney's Housing Affordability Crisis

Introducing Ways to Sense Check LLM Output for Hallucinations:

- Request sources or references for key claims
- Ask the LLM to rate its confidence in different parts of its response
- Instruct the LLM to clearly differentiate between facts and opinions
- Request that the LLM highlight any assumptions it's making
- Ask for potential counterarguments to the proposed solutions

Task: write a final prompt that includes methods to verify the accuracy of the LLM's output.

- Link to a summary of these prompt changes
- Link to output

Prompting - Data Analysis

How can we improve this?

Analyze this dataset and create visualizations.

Prompting - Data Analysis

How can we improve this?

"Analyze this sales dataset for an e-commerce company and create visualizations to show trends."

Prompting - Data Analysis

How can we improve this?

"You are a data analyst for an e-commerce company.
Analyze the following sales dataset:

- Column 1: Date (YYYY-MM-DD)
- Column 2: Product Category (Electronics, Clothing, Home & Garden, Books)
- Column 3: Sales Amount (USD)
- Column 4: Customer Age Group (18-25, 26-35, 36-50, 51+)
- Column 5: Customer Location (US State)
- The dataset covers the last 12 months. Perform the following tasks:
- Calculate the total sales for each product category.
- Identify the top 3 US states by sales volume.
- Analyze the sales trends over time for each product category.
- Determine if there's a correlation between customer age group and product category preference.
- Create appropriate visualizations for each of the above analyses.
- Write a brief summary (3-5 sentences) of the key insights from your analysis.

Prompting - Data Analysis

How can we improve this?

Other tips to enhance your prompt

- Specify the desired output format (e.g., Jupyter Notebook, PowerPoint, PDF report)
- Request specific types of visualizations (e.g., line charts for trends, bar charts for comparisons)
- Ask for actionable recommendations based on the insights
- Specify any statistical tests to be performed (e.g., chi-square test for age group and category preference)
- Request a section on data quality issues and how they were addressed