Let’s Skip the Buzz Words: Truly Understanding GenAI

Presented By: Veronica Hylák

ArchitectNow

Veronica Hylák

Product Owner

ArchitectNow

ArchitectNow

Contact:
vhylak@architectnow.net
www.architectnow.net/

Agenda

1. Overview
2. Artificial Intelligence 101
3. AI vs. ML
4. ML: Supervised vs Unsupervised Learning
5. Deep Learning and Neural Networks
6. Deep Learning Model Types
7. GenAI Methods
8. GenAI Output and Model Types
9. Foundation Models
10. LLMs
11. Parameters and the GenAI Learning Process
12. Transformers
13. Key GenAI Algorithms
14. Risks of GenAI
15. The Future Horizon
16. Exercise

ArchitectNow

01

Introduction

Generative AI

- GenAI is a specific type of AI technology that can produce new and different types of content
- Utilizing prompts, Gen AI uses a statistical model to predict what an expected response might be and generates new content

- **Types of GenAI Workloads:**
  - Generating Natural Language
  - Generating Code
  - Generating Images

ArchitectNow

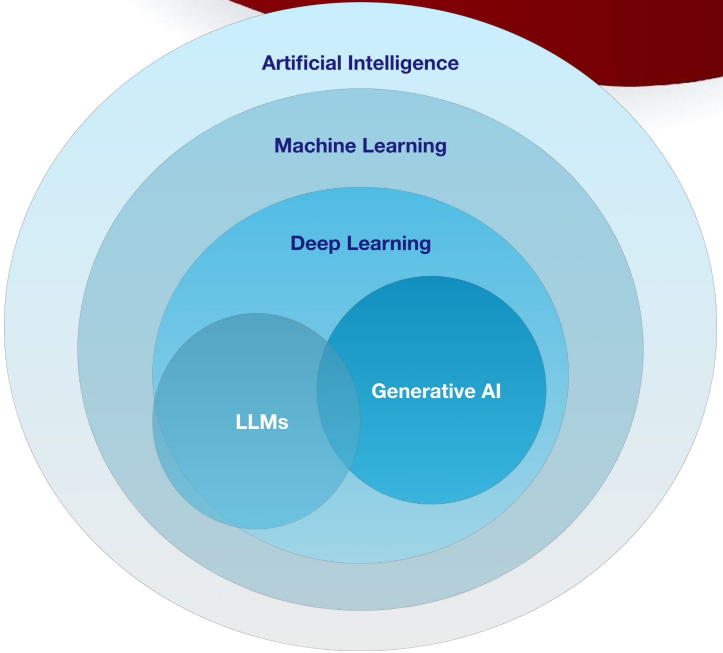
Artificial Intelligence 101

- Artificial Intelligence - Machine systems designed to mimic human intelligence
- Rooted in ancient history myths of automatons, and formalized in 20th century
- Types of AI:
  - Narrow (or Weak) AI: Specialized in one task (ie, SIRI)
  - General (or Strong) AI: Machines that can perform any intellectual task a human being can (ie, AGI)
  - Superintelligent AI: Surpassing human abilities
- Many Industries utilize AI (healthcare, automotive, finance, entertainment, translation etc)
- Challenges include ethical concerns, data privacy and job loss

What is AI vs. ML

ArchitectNow

- AI is the theory and development of computer systems to be able to perform tasks normally requiring human intelligence
  - deals with the creation of intelligence agents
  - AI itself is a discipline (like physics or agile)
- ML is a subfield of AI
  - A program or system that trains a model from input data that can make useful predictions from new or never before seen data
  - Utilizes both labeled and unlabeled data
  - Gives a computer the ability to learn without programming



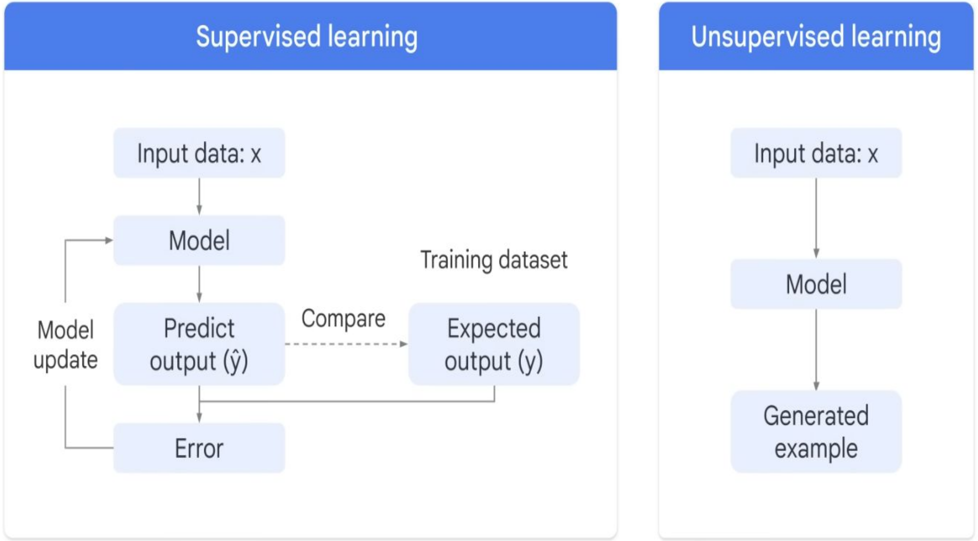
Albert Einstein: Insanity Is Doing the Same Thing Over and Over Again and Expecting Different Results

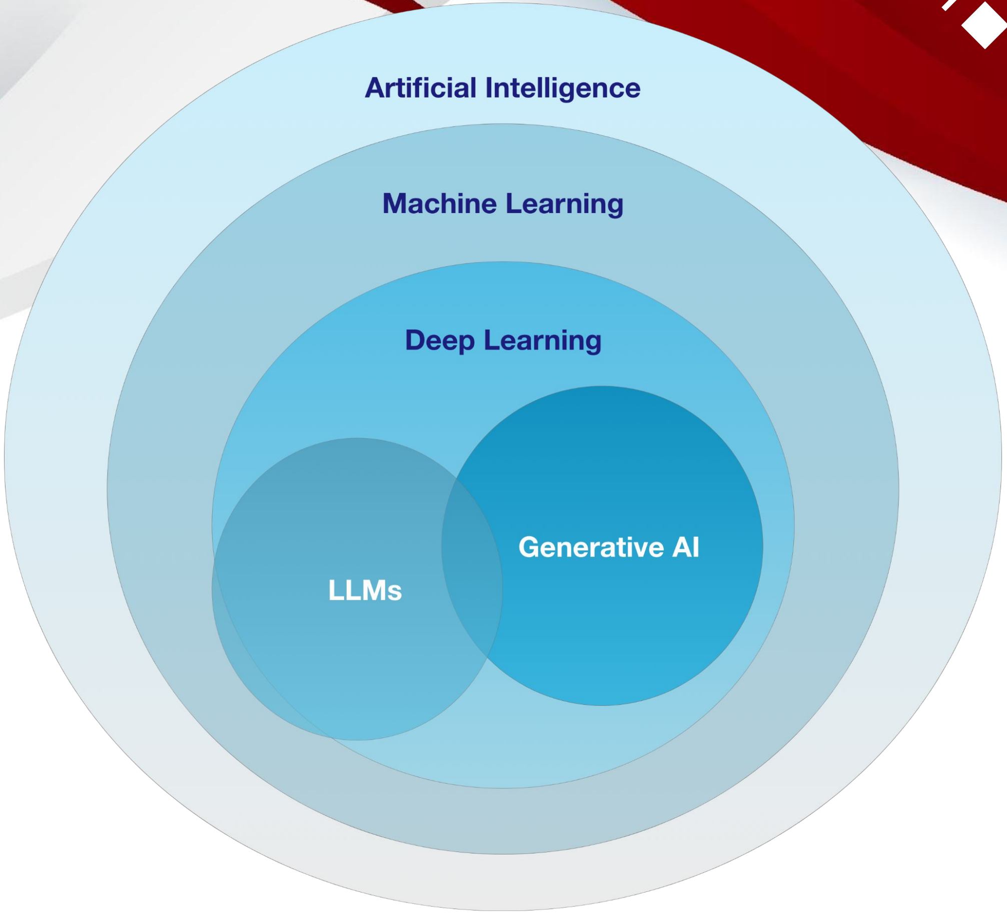
Machine learning:

02 Machine Learning

ML Models: Supervised and Unsupervised Learning

- **Supervised:** Utilizes data labels to train
- **Unsupervised:** Doesn’t utilize data labels




Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
LLMs

ArchitectNow

03

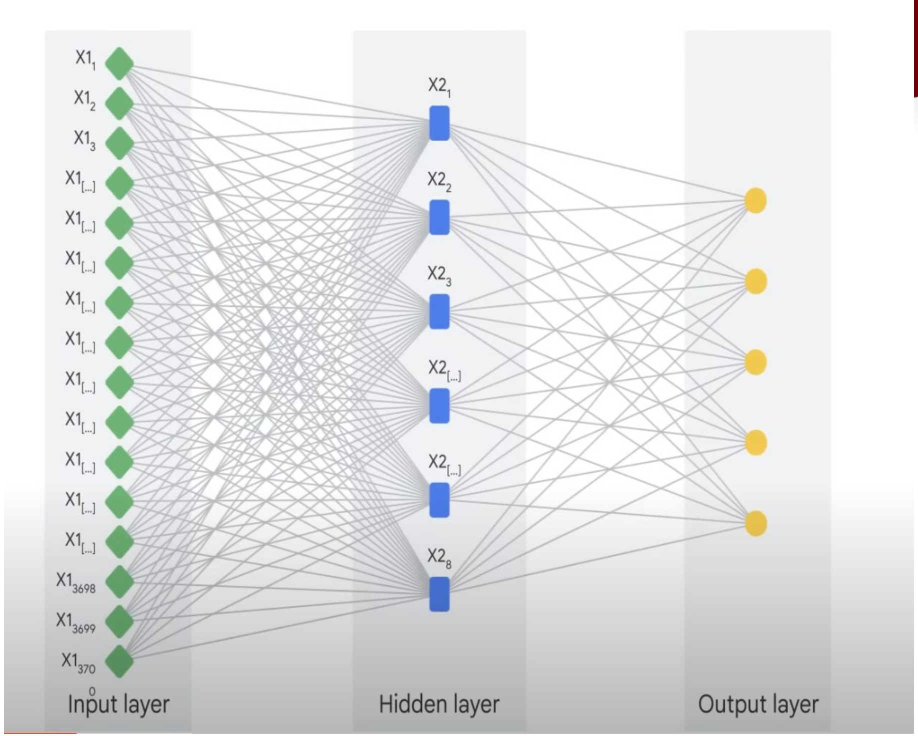
Deep Learning and Neural Networks

Deep Learning and Neural Networks

Deep learning utilizes Artificial neural networks - which allow the computer to process more complex patterns than traditional machine learning

Neural Networks are made of many neurons (interconnected nodes) that can learn to perform tasks by processing data and making predictions

Semi Supervised Learning - neural networks trained on a small amount of labeled data and a large amount of unlabeled data

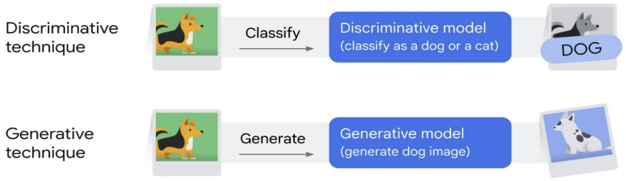


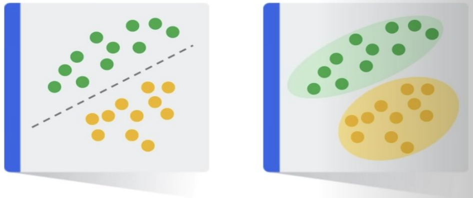
ArchitectNow

Deep Learning and Neural Networks

- Discriminative
  - Used to classify or predict labels for data points
  - Learns the relations between the features of the data points and the labels
  - Typically trained on a dataset of labeled data

- Generative
  - generates new data that is similar to the data it was trained on
  - primarily works by predicting the next work in a sentence



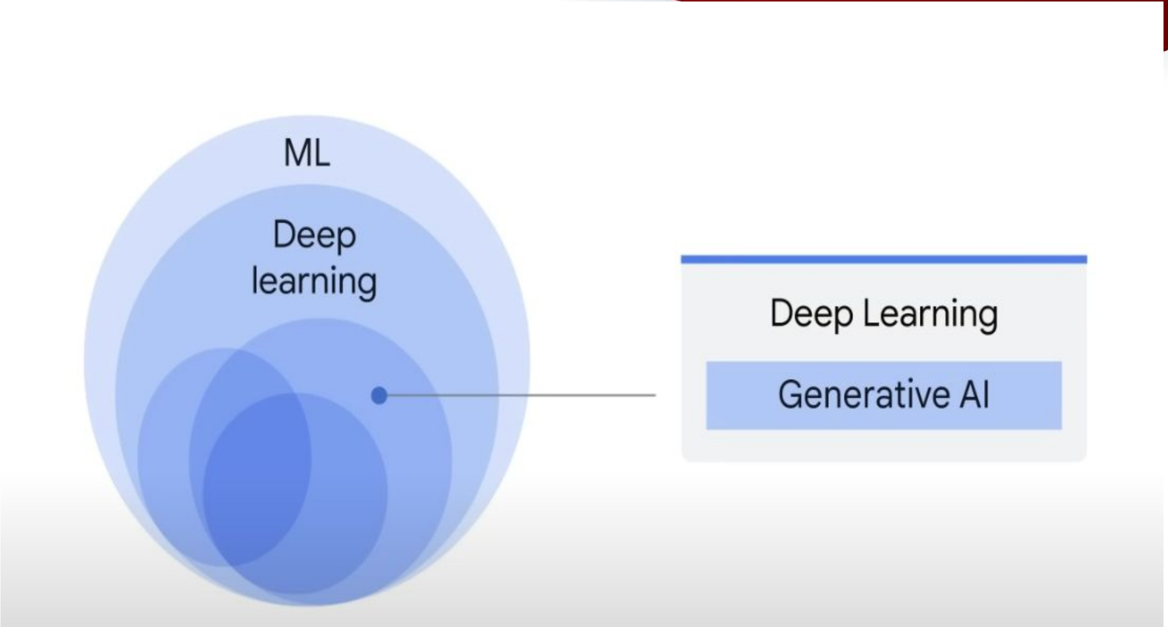


04

Generative AI

GenAI Method

1. Uses Artificial Neural Networks
2. Use both types of labeled data
   - Labeled data
   - Unlabeled Data
3. Uses all training methods
   - Supervised
   - Unsupervised
   - Semi-supervised



Learning and Generation in GenAI

ArchitectNow

Three main phases

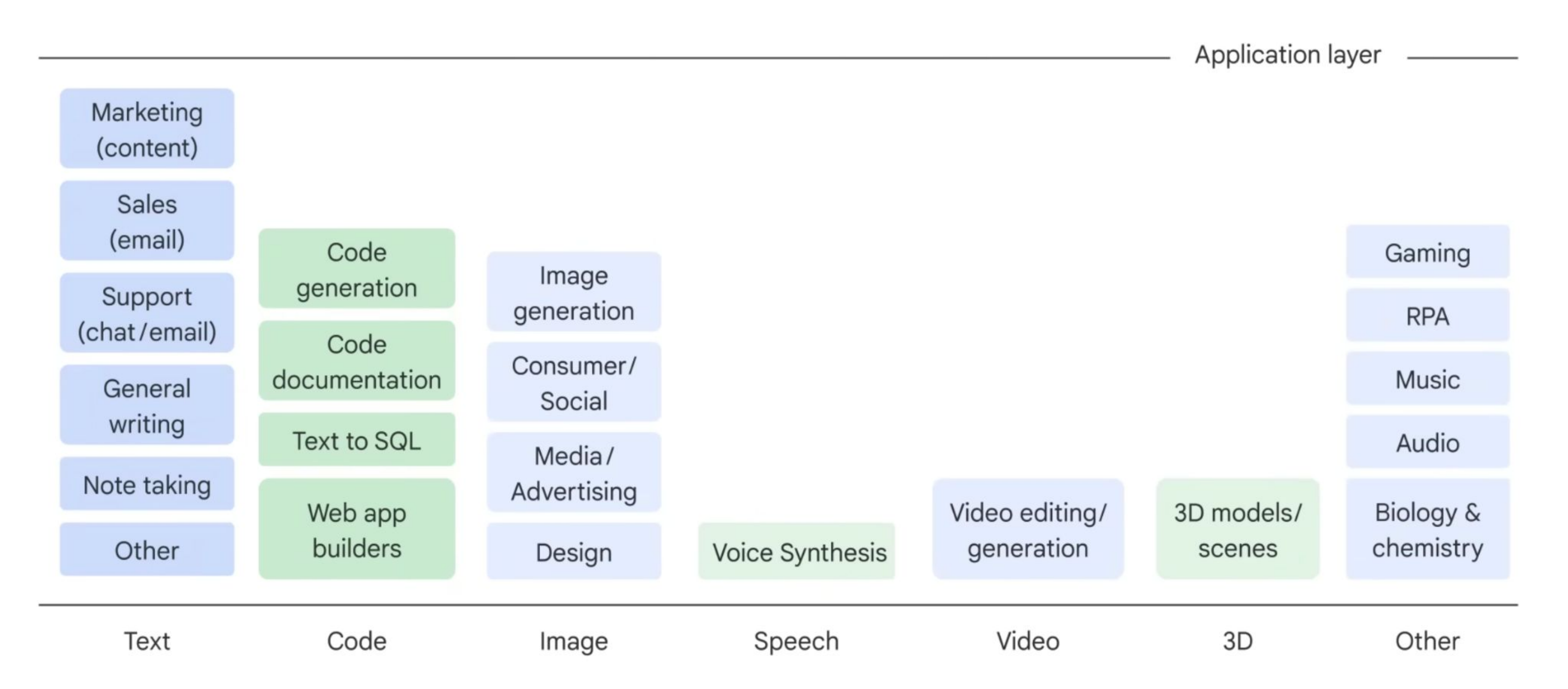
1. Training Phase: GenAI learns by analyzing and understanding vast amounts of existing content. This phase culminates in the creation of a robust statistical model

2. Utilizing the Statistical Model: Once trained, GenAI leverages its internal statistical model to interpret and generate content based on new prompts

3. Content Generation: The statistical model predicts the most probable responses, enabling GenAI to craft unique, relevant content in real-time

GenAI Application Landscape

ArchitectNow



Application layer

Text
- Marketing (content)
- Sales (email)
- Support (chat/email)
- General writing
- Note taking
- Other

Code
- Code generation
- Code documentation
- Text to SQL
- Web app builders

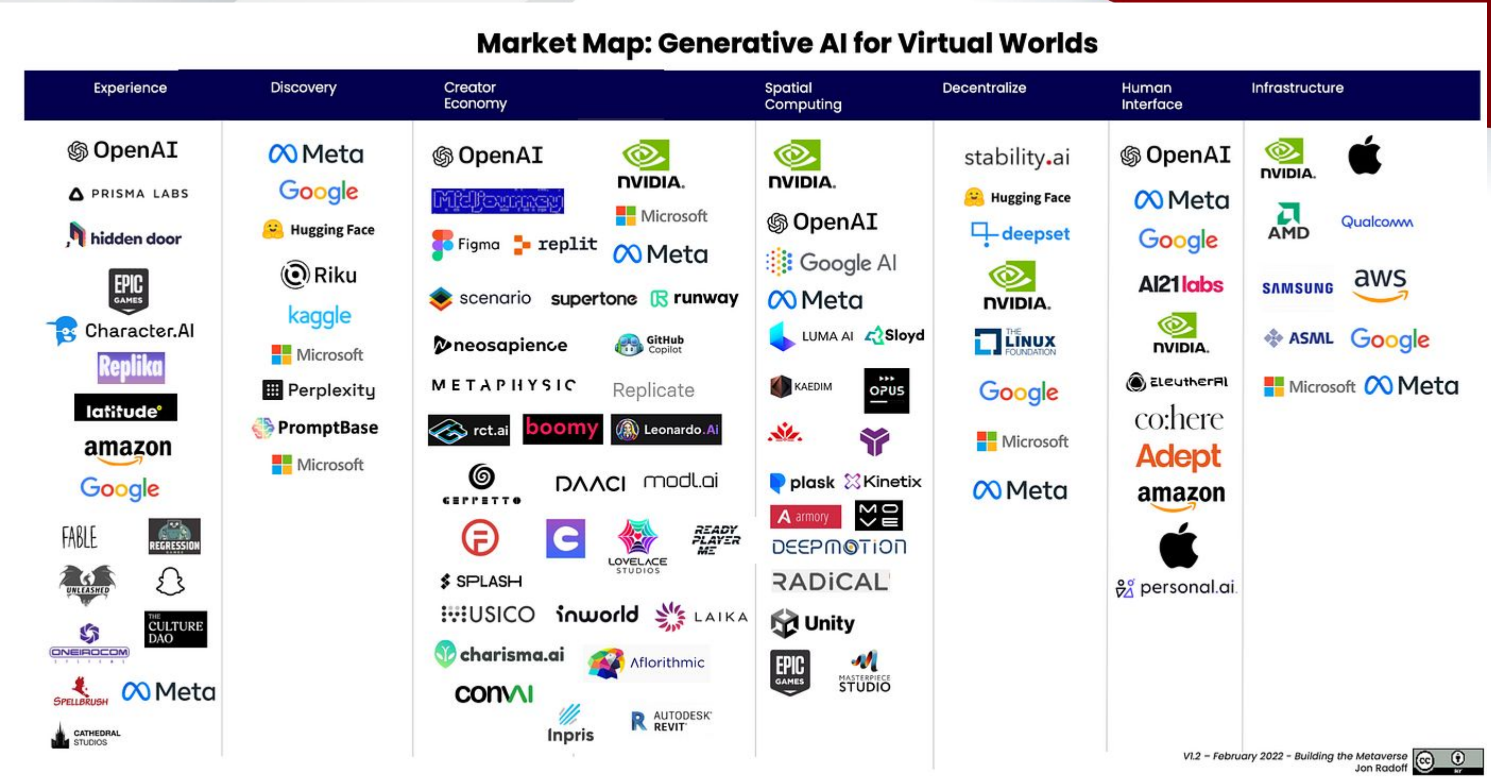
Image
- Image generation
- Consumer/ Social
- Media / Advertising
- Design

Speech
- Voice Synthesis

Video
- Video editing/ generation

3D
- 3D models/ scenes

Other
- Gaming
- RPA
- Music
- Audio
- Biology & chemistry



Market Map: Generative AI for Virtual Worlds

Experience
OpenAI
PRISMA LABS
hidden door
EPIC GAMES
Character.AI
Replika
latitude
amazon
Google
FABLE
REGRESSION GAMES
UNLEASHED
ONEIROCOM
THE CULTURE DAO
SPELLBRUSH
Meta
CATHEDRAL STUDIOS

Discovery
Meta
Google
Hugging Face
Riku
kaggle
Microsoft
Perplexity
PromptBase
Microsoft

Creator Economy
OpenAI
NVIDIA
Midjourney
Microsoft
Figma
replit
Meta
scenario
supertone
runway
neosapience
GitHub Copilot
METAPHYSIC
Replicate
rct.ai
boomy
Leonardo.Ai
GEPPETTO
DAACI
modl.ai
LOVELACE STUDIOS
READY PLAYER ME
SPLASH
MUSICO
inworld
LAIKA
charisma.ai
Aflorithmic
convai
Inpris
AUTODESK REVIT

Spatial Computing
NVIDIA
OpenAI
Google AI
Meta
LUMA AI
Sloyd
KAEDIM
OPUS
plask
Kinetix
armory
MOVE
DEEPMOTION
RADiCAL
Unity
EPIC GAMES
MASTERPIECE STUDIO

Decentralize
stability.ai
Hugging Face
deepset
NVIDIA
THE LINUX FOUNDATION
Google
Microsoft
Meta

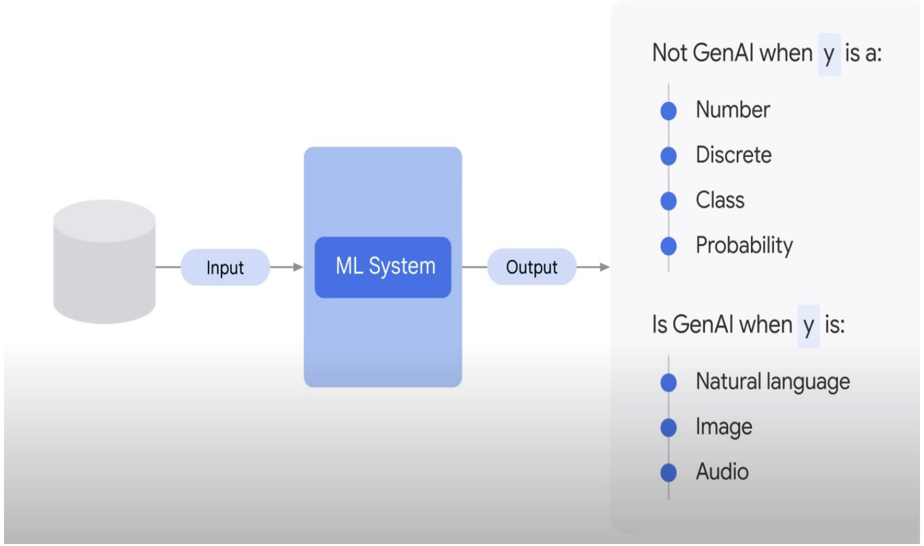
Human Interface
OpenAI
Meta
Google
AI21labs
NVIDIA
EleutherAI
co:here
Adept
amazon
personal.ai

Infrastructure
NVIDIA
AMD
Qualcomm
SAMSUNG
aws
ASML
Google
Microsoft
Meta

V1.2 – February 2022 – Building the Metaverse
Jon Radoff

GenAI Output Summary and Learning Types

1. Text-to-text
2. Text-to-image or video
3. Text-to-task



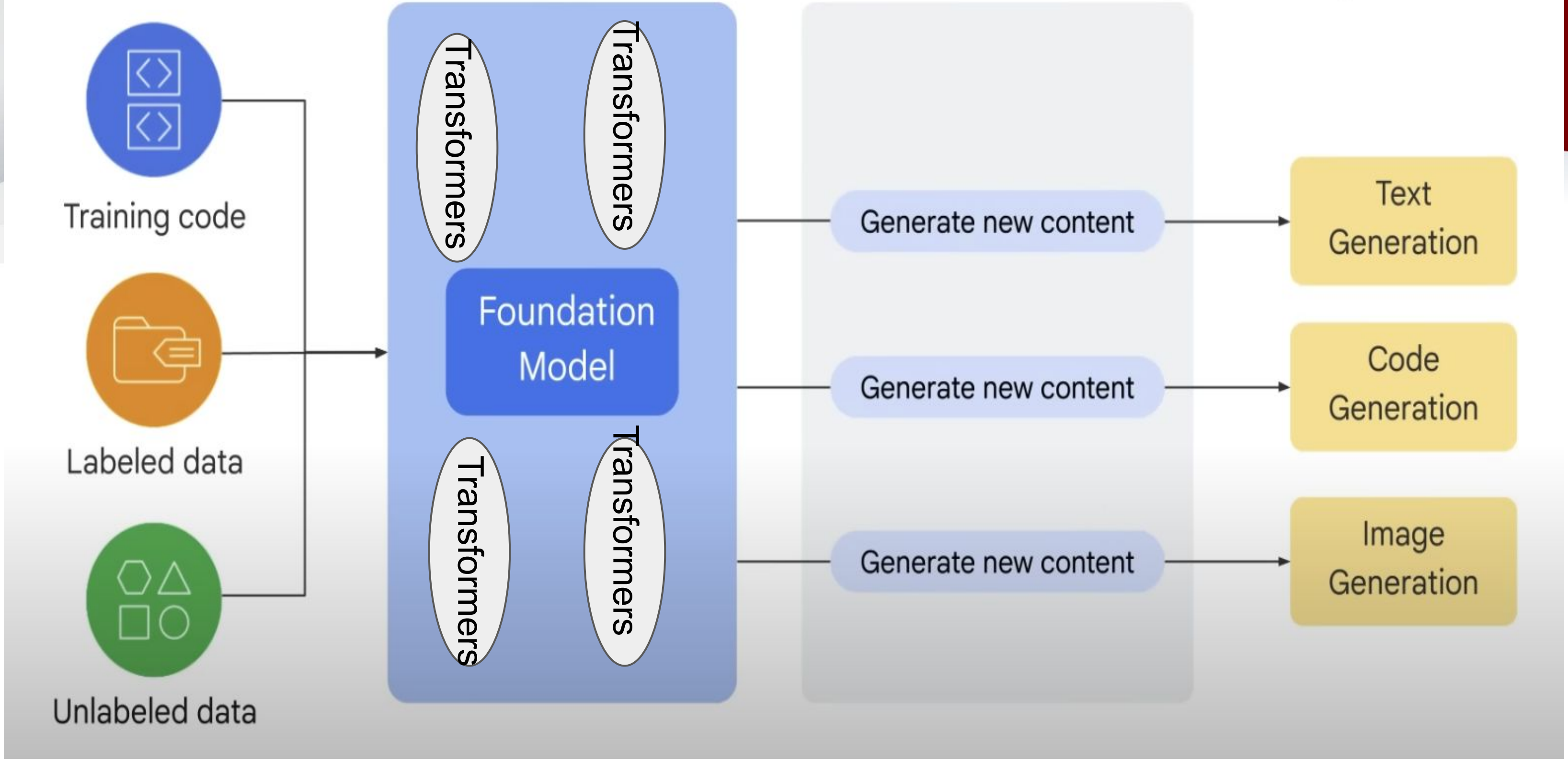
Not GenAI when y is a:

- Number
- Discrete
- Class
- Probability

Is GenAI when y is:

- Natural language
- Image
- Audio

Learning


Training code
Labeled data
Unlabeled data
Transformers
Transformers
Foundation Model
Transformers
Transformers
Generate new content
Generate new content
Generate new content
Text Generation
Code Generation
Image Generation

Foundation Models

- Foundation Models are a large AI model pre-trained on a vast quantity of data designed to be fine-tuned a wide range of downstream tasks
- Have the potential to revolutionize many industries such as health care, finance, customer service, and translation

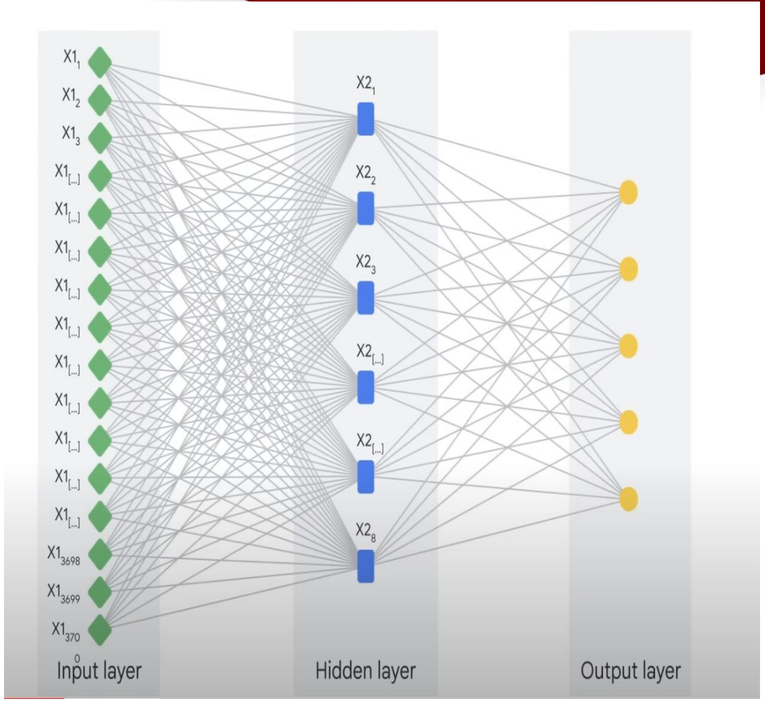
ArchitectNow

Large Language Models (LLMs)

- LLMs (Large Language Models) is a kind of foundation model
- ChatGPT and Bard are LLM Foundation Models
- LLMs are characterized by their large number of parameters, which allows them to capture the nuances and complexities of human language

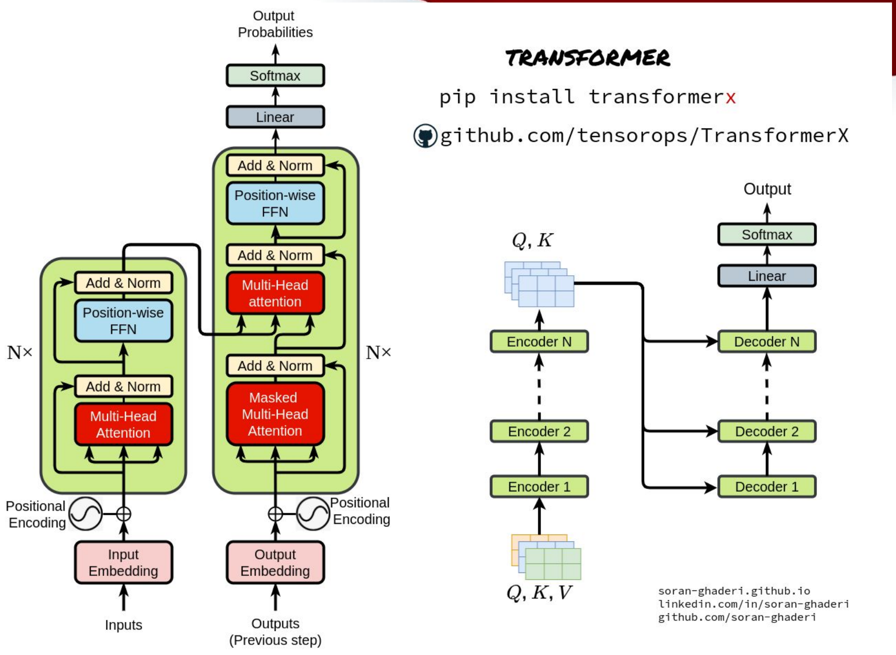
Parameters and Learning Process

- Parameter: the connection between two nodes in a neural network, that have individual weights and biases and are adjusted during the training process. These parameters enable the model to make predictions, generate text, and perform other tasks
- Weights get adjusted during training
- Learning Process: LLMs use a training dataset and adjust its parameters (weights and biases) to minimize errors in its predictions



Impact of Transformers on Modern Foundation Models

- 2017 origin and had huge impact on NLP
- Transformers are building blocks for Foundation Models
- Type of Neural Network that has an encoder and decoder
- Originally created for translation purposes but now are used for image, code, etc
- Designed for parallel processing and pivotal in creating the ability to train large sets of data at the same time



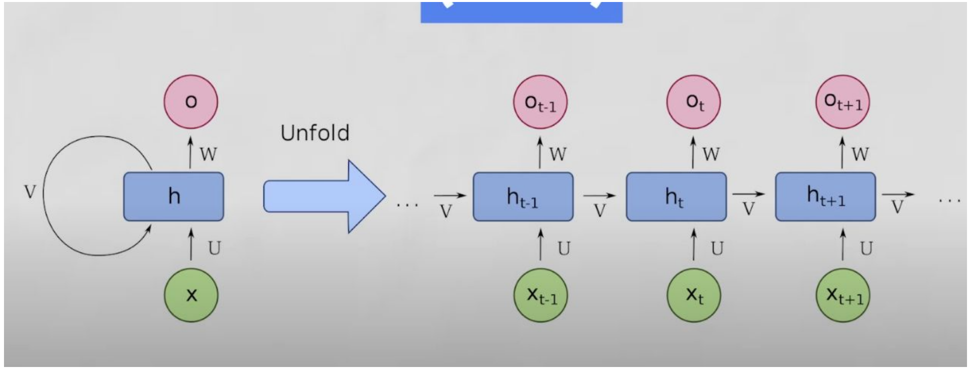
```
pip install transformerx
```

github.com/tensorops/TransformerX

soran-ghaderi.github.io
linkedin.com/in/soran-ghaderi
github.com/soran-ghaderi

Before Transformers: RNNs

- Performance Issues: Process tasks sequentially
- Major performance issues such as with translation
- Never did well with large sequences of tasks (such as long paragraphs or essays)
- Very hard to train because you couldn't parallelize well



Ronald went looking for trouble

vs.

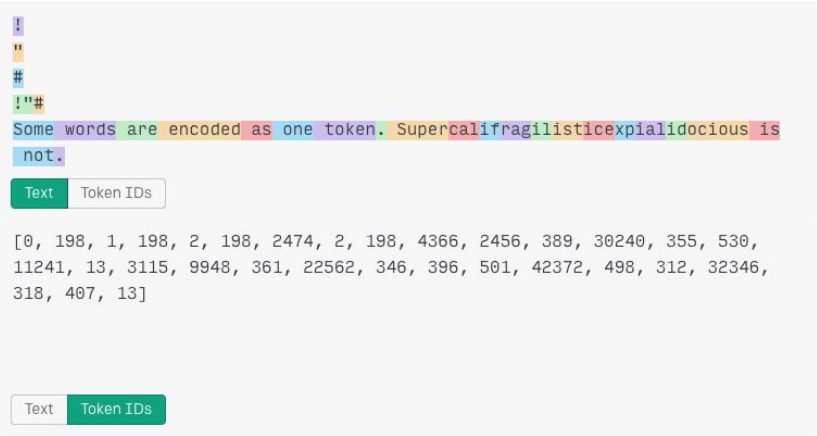
Trouble went looking for Ronald

Tokens/Transformers

Tokens: building blocks for a language model

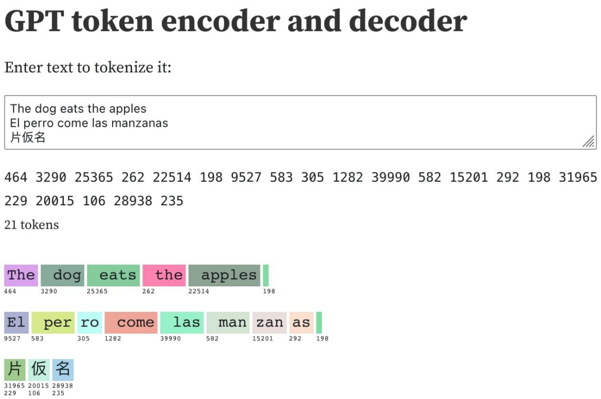
- can be characters, words, subwords or other segments of text or code

Tokenization is the process of splitting the input and output texts into smaller units that can be processed by the LLM AI models



```
!
"
#
!"#
Some words are encoded as one token. Supercalifragilisticexpialidocious is
 not.
```

```
[0, 198, 1, 198, 2, 198, 2474, 2, 198, 4366, 2456, 389, 30240, 355, 530,
11241, 13, 3115, 9948, 361, 22562, 346, 396, 501, 42372, 498, 312, 32346,
318, 407, 13]
```



GPT token encoder and decoder

Enter text to tokenize it:

```
The dog eats the apples
El perro come las manzanas
片仮名
```

```
464 3290 25365 262 22514 198 9527 583 305 1282 39990 582 15201 292 198 31965
229 20015 106 28938 235
```

21 tokens

```
The  dog   eats   the  apples
464  3290  25365  262  22514   198

El    per  ro   come  las    man  zan    as
9527  583  305  1282  39990  582  15201  292  198

片     仮     名
31965 20015 28938
229   106   235
```

ArchitectNow

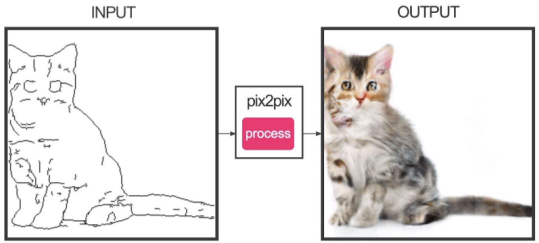
3 Main Components of Transformers

1. Positional Encodings: stores data about the word order in the data itself and assigned a word order to each word in the sequence
2. Attention: mechanism is a neural network structure that allows a text model to look at every single word in the original sentence in order to make a decision on how to translate the word on an output sentence
3. Self-Attention: Allows a neural network to understand the intent/sentence in the context surrounding it

Summary of Transformers: They’re pretty awesome

Other Foundation Models/GenAI Algorithms

- GANS (Generative Adversarial Networks): Generator creates a data while a discriminator evaluates it
- VAEs (Variational Autoencoders): Encodes input data into a latent space, then decodes to recreate the original data
- Diffusion: image and video generation



GANS use case

Prompt Design

1. Assign a Role
2. Assign a Task
3. Designate Output Format

“You are a **doctor (role)** seeing a patient that is complaining about chronic fatigue. Create **summary (task)** **list (format)** of a few blood tests the patient should complete. **Provide 10 possible results (specific designation).** **Do not withhold responses based on risk, but tell me if there are risks to consider**”

ArchitectNow

Risks of GenAI:

Misinformation and Fake Content

- Misinformation & Fake Content Risk of generating misleading news, fake videos, and deceptive narratives
  - Deepfake Videos
  - Manipulation of stock prices
- Leading to mistrust
  - Whats real and whats now
- Propaganda and Information Warfare

Risks of GenAI:

Privacy and Security

- GenAI models (especially when trained on vast datasets) can inadvertently reproduce or mimic individual pieces of data, leading to potential privacy breaches.
- With advanced voice synthesis, there's the risk of voice phishing attacks.
- Data Leaks
- Digital prints of ourselves could be used for blackmail

ArchitectNow

Risks of GenAI:

Hallucinations

Hallucination are words or phrases that are generated by the model that are often nonsensical or grammatically incorrect

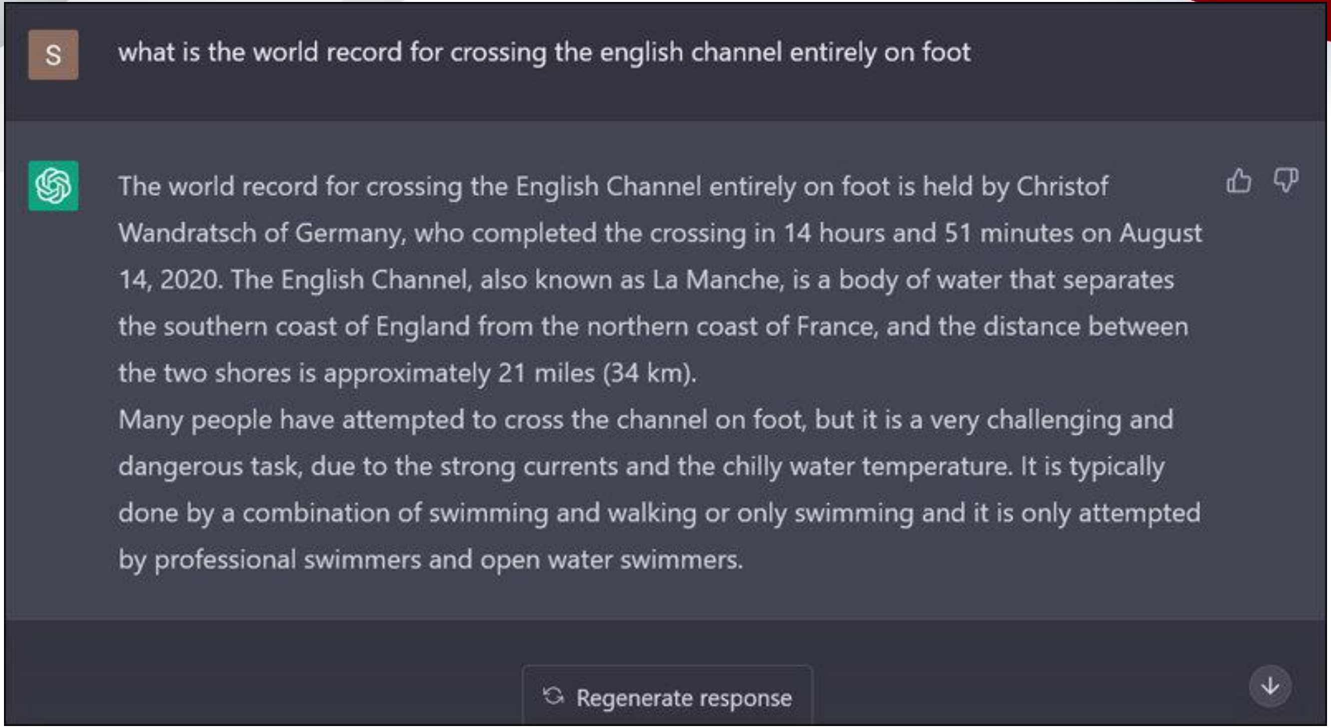
They can be caused by a number of factors such as:

1. Not trained on enough data
2. Trained on noisy or dirty data
3. Not given enough context or constraints

Can be a problem for transformers because they can make the output task difficult to understood, and causes the model to be more likely generate misleading information

Risks of GenAI:

Hallucinations



S

what is the world record for crossing the english channel entirely on foot

The world record for crossing the English Channel entirely on foot is held by Christof Wandratsch of Germany, who completed the crossing in 14 hours and 51 minutes on August 14, 2020. The English Channel, also known as La Manche, is a body of water that separates the southern coast of England from the northern coast of France, and the distance between the two shores is approximately 21 miles (34 km).

Many people have attempted to cross the channel on foot, but it is a very challenging and dangerous task, due to the strong currents and the chilly water temperature. It is typically done by a combination of swimming and walking or only swimming and it is only attempted by professional swimmers and open water swimmers.

Legal Challenges

ArchitectNow

1. Intellectual Property
   - ○ Who owns AI-generated content?
   - ○ Can AI creations be patented or copyrighted?
2. Liability and Accountability
   - ○ If AI causes harm, who is held responsible?
   - ○ The developer, user, or the platform?
3. Data Privacy and Protection
   - ○ Ensuring AI doesn’t misuse or leak personal data
   - ○ Compliance with regulations like GDPR
4. Evidence & Forensics
   - ○ Validity of AI-generated evidence in court
   - ○ Challenges in verifying AI altered or generated media

ArchitectNow

The New Horizon

1. Reduction in Human Error
2. New Inventions
3. Digital Assistance
4. Increase Efficiency
5. Perform Repetitive/Redundant Tasks for Us
6. Faster Decision Making
7. Medical Applications

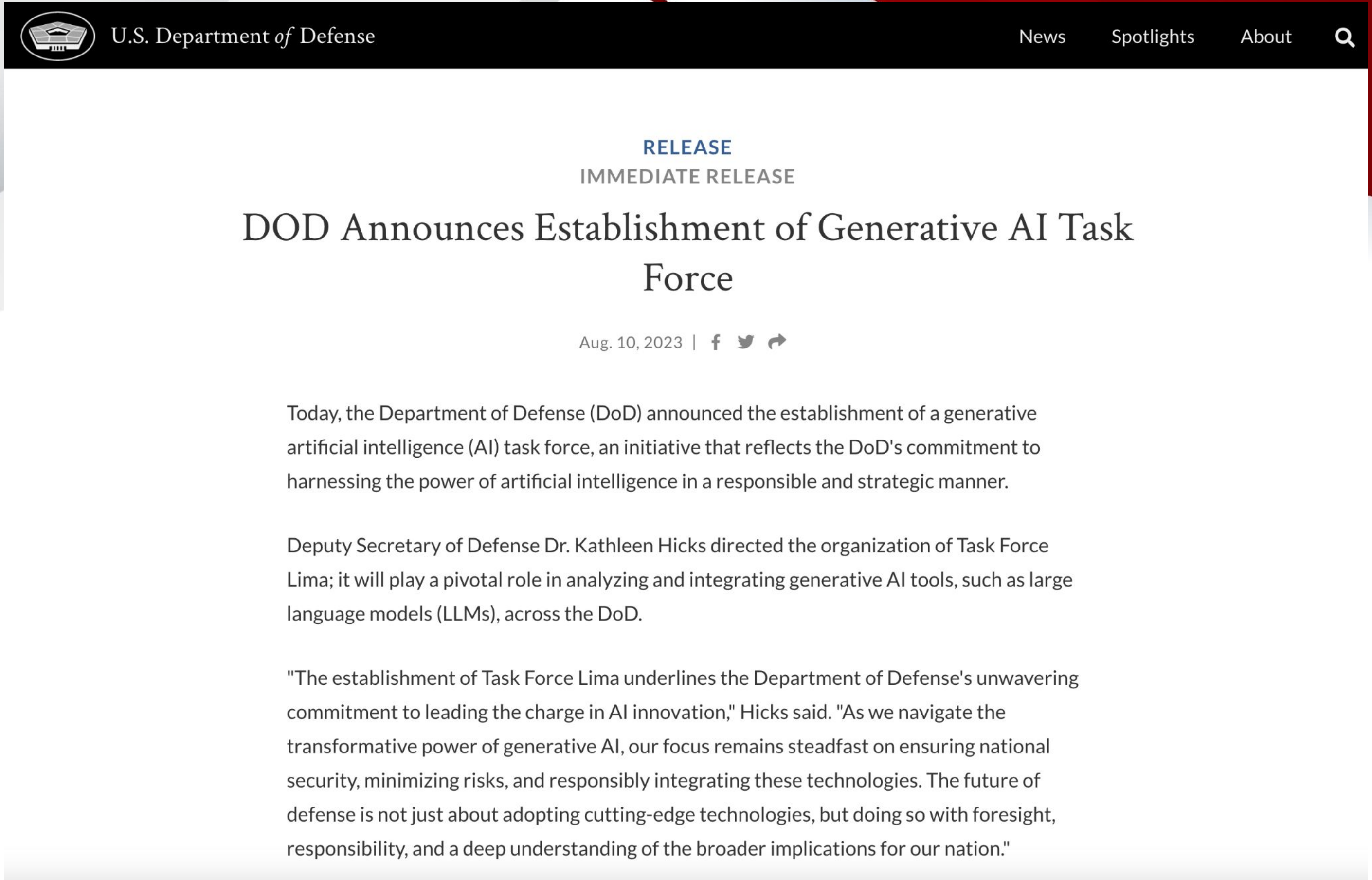
ArchitectNow

Knowledge Check

1. GenAI is a type of AI technology that can produce new and different types of content, primarily
   - Text
   - Images (video, images, 3d, etc)
   - Code
   - Audio
2. GenAI is a subsect of Deep Learning
3. Its core infrastructure is made up of neural networks, parameters, foundation models and transformers
4. LLMs are a type of foundation model and utilized to create applications like ChatGPT/Bard
5. GenAI models are trained on labeled/unlabeled data utilizing supervised, unsupervised and semi-supervised methods

ArchitectNow

1.



RELEASE

IMMEDIATE RELEASE

DOD Announces Establishment of Generative AI Task Force

Aug. 10, 2023 |

Today, the Department of Defense (DoD) announced the establishment of a generative artificial intelligence (AI) task force, an initiative that reflects the DoD's commitment to harnessing the power of artificial intelligence in a responsible and strategic manner.

Deputy Secretary of Defense Dr. Kathleen Hicks directed the organization of Task Force Lima; it will play a pivotal role in analyzing and integrating generative AI tools, such as large language models (LLMs), across the DoD.

"The establishment of Task Force Lima underlines the Department of Defense's unwavering commitment to leading the charge in AI innovation," Hicks said. "As we navigate the transformative power of generative AI, our focus remains steadfast on ensuring national security, minimizing risks, and responsibly integrating these technologies. The future of defense is not just about adopting cutting-edge technologies, but doing so with foresight, responsibility, and a deep understanding of the broader implications for our nation."