

Generative AI basics in 15 days¹

- **Day 1: Introduction to Generative AI** - Overview of generative AI and its importance in business.
- **Day 2: Types of Generative AI Models** - Exploring different generative AI models like GANs, VAEs, and transformers.
- **Day 3: Traditional ML vs Generative AI** - Comparing traditional machine learning with generative AI methods.
- **Day 4: What are GPUs** - Understanding the role of GPUs in AI and machine learning tasks.
- **Day 5: What it Takes to Train a Foundation Model** - Insights into the resources and processes for training large foundation models.
- **Day 6: How to Customize Foundation Models** - Discussing techniques for customizing foundation models for specific uses.
- **Day 7: The Most Popular LLMs Available** - Overview of the most widely-used large language models and their features.
- **Day 8: Generative AI Applications and Use Cases** - Exploring practical applications of generative AI across business sectors.
- **Day 9: The Generative AI Stack** - Understanding the components and architecture of the generative AI tech stack.
- **Day 10: The Emergence of Small Language Models** - Discussing the rise and importance of small language models in AI.
- **Day 11: The AI Engineer Profession and Skills** - Exploring the role, responsibilities, and required skills of AI engineers.
- **Day 12: Ethical Considerations in AI** - Discussing the ethical challenges in AI development and deployment.
- **Day 13: Create Your AI for Business Roadmap** - How to develop a strategic AI integration roadmap for businesses.
- **Day 14: Future Trends in AI** - Exploring future developments and trends in AI.
- **Day 15: Continuing Your AI Journey** - Providing resources and advice for continued AI learning and exploration.

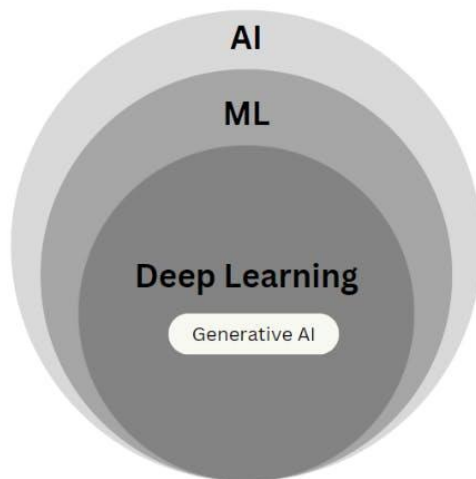
¹ [Post | LinkedIn](#)

Welcome to Day 1 of our Generative AI journey!

Today, I'm uncovering what Generative AI is and why it's a game-changer in the business world. Imagine AI not just analyzing data but creating new, innovative content – that's Generative AI!

Ok let's start with the basics, but don't worry, we will get into more advanced concepts as we go.

definition



Generative AI is a type of Artificial Intelligence that creates new content based on what it has learned from existing content.

When given a **prompt**, the model predicts what an expected response might be, creating new, original data like images, text, audio, video.

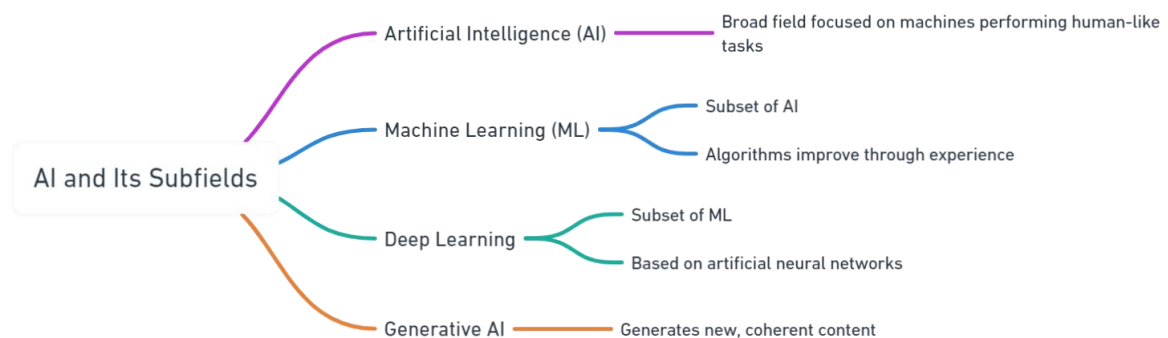
"Creativity" powered by examining large training datasets.

Artificial Intelligence (AI): AI is the broad field of computer science focused on creating machines capable of performing tasks that typically require human intelligence.

Machine Learning (ML): ML is a subset of AI involving algorithms and statistical models that enable computers to improve their performance on a task through experience.

Deep Learning: Deep Learning is a subset of ML based on artificial neural networks, where algorithms learn from large amounts of data to identify patterns and make decisions.

Generative AI: Generative AI refers to AI technologies that can generate new content, ideas, or data that are coherent and plausible, often resembling human-generated outputs.



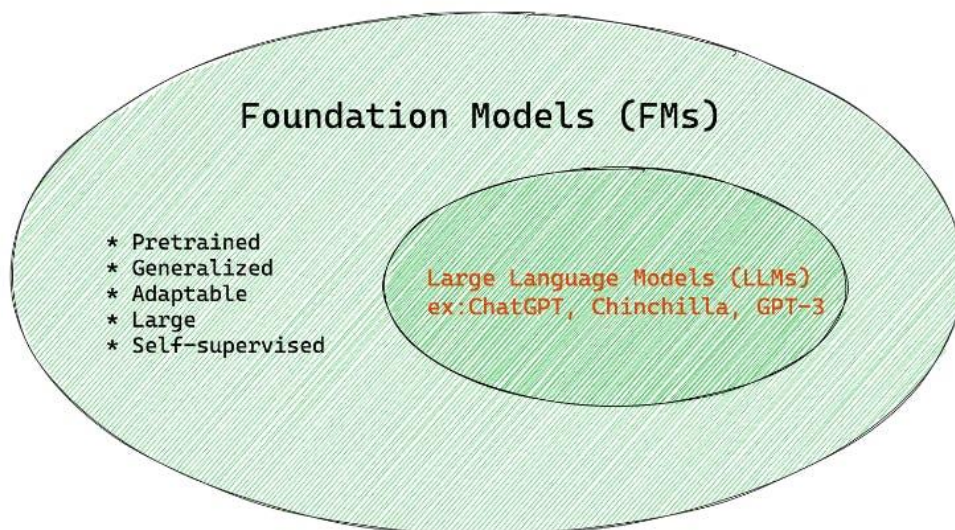
What powers Generative AI

Foundation models are large-scale artificial intelligence models that have been trained on vast amounts of data. These models are highly versatile and can be adapted to a wide range of tasks and applications.

Generative AI is one of the applications of foundation models. It involves using these models to create new content, such as text, images, or music. The foundation model serves as the underlying structure that understands and processes information, enabling the generative AI to produce new, coherent, and relevant outputs.

In simple terms, foundation models are like the core engine, and generative AI is one of the many things that this engine can power.

the models powering generative ai

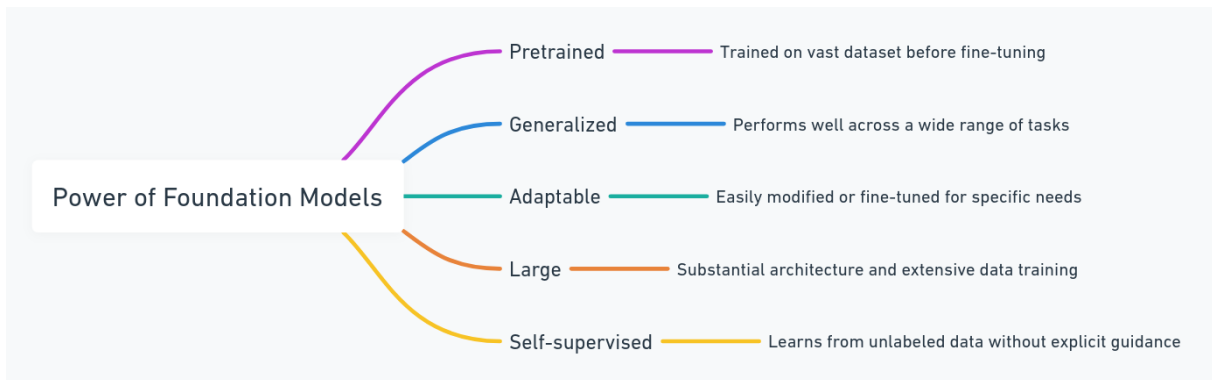


FMs are models trained on broad data (using self-supervision at scale) that can be adapted to a wide range of downstream tasks.

What makes Foundation Models so powerful?

1. **Pretrained:** The model has already been trained on a vast dataset before being fine-tuned or applied to specific tasks.
2. **Generalized:** The model is capable of performing well across a wide range of tasks, not just the ones it was specifically trained for.
3. **Adaptable:** The model can be easily modified or fine-tuned to suit particular needs or tasks.
4. **Large:** The model is built with a substantial architecture and trained on extensive data, giving it a broad understanding and capability.

5. **Self-supervised:** The model primarily learns by analyzing and making sense of unlabeled data, without explicit guidance on what to learn.



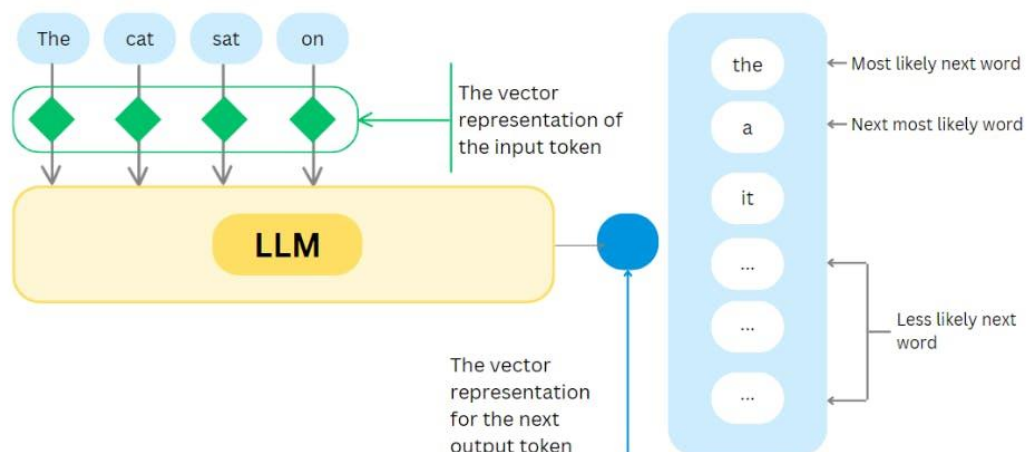
And what are Large Language Models?

LLMs are a type of foundation model specifically designed to understand and generate text. They're trained on huge amounts of text, which makes them good at a wide range of language tasks. LLMs are part of the broader category of foundation models, meaning they're versatile and can be adapted for different uses involving language.

LLMs like GPT take, as input, an entire sequence of words, and predicts which word is most likely to come next. They perform that prediction of the next word in a sequence by analyzing patterns in vast amounts of text data.

a next word predictor

LLMs like GPT take, as input, an entire sequence of words, and predicts which word is most likely to come next.



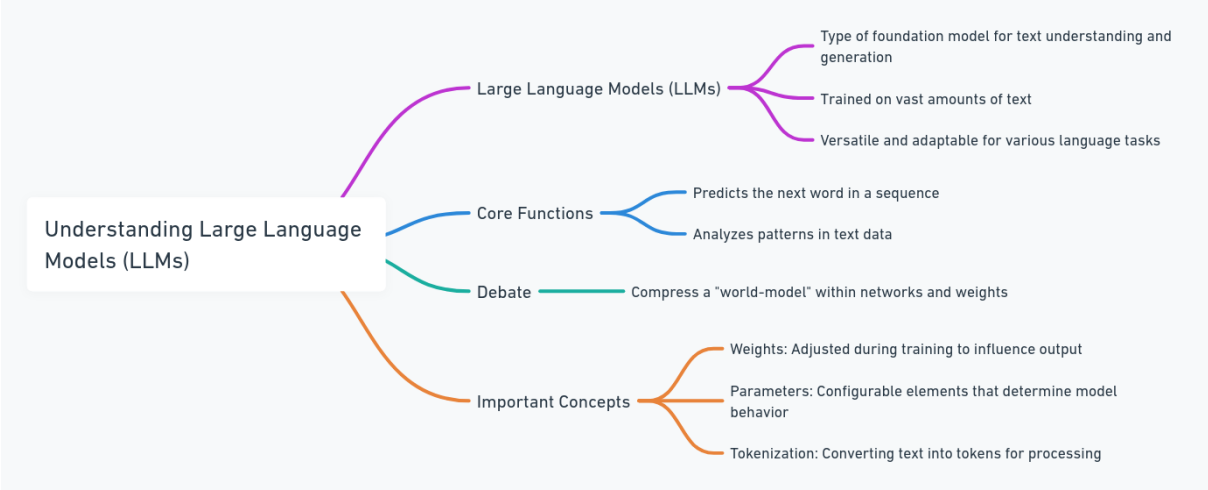
There's a big debate that LLMs do more than predict the next word; they compress a "world-model" within their complex networks and weights. This is an area of active debate within the AI community. You can join the discussion about this [here](#)

Important concepts to understand in LLMs are:

Weights: Numerical values within a machine learning model that are adjusted during training to influence the model's output in response to input data.

Parameters: The broader set of configurable elements in a model, including weights, that determine its behavior and performance.

Tokenization: The process of converting text into smaller units (tokens), such as words or subwords, which are used as the input for LLMs to understand and generate language.

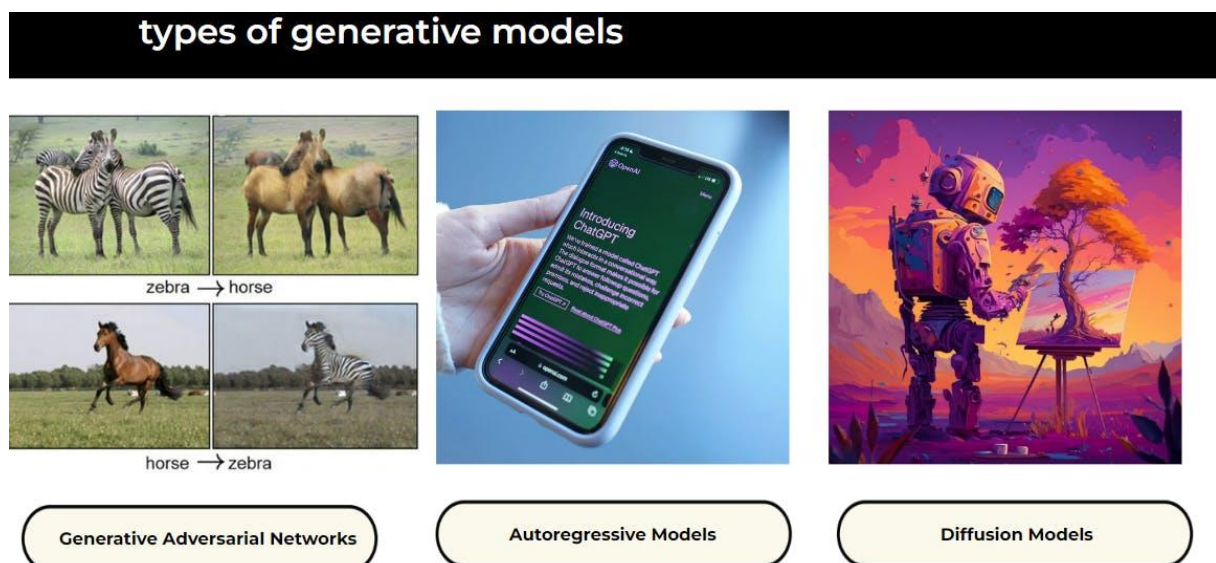


Day 2 - Types of Generative AI Models

Welcome to Day 2 of our Generative AI journey! Today, we're diving into the different types of Generative AI models, each with its unique capabilities and applications.

To keep it simple, I summarized this in four types:

- 1. Generative Adversarial Networks (GANs):** These models are game-changers in image generation, creating everything from art to realistic photos.
- 2. Variational Autoencoders (VAEs):** VAEs are great for tasks that involve compressing and generating high-quality images, offering applications in style transfer and more.
- 3. Transformer Models:** Known for their prowess in text, transformer models like GPT are revolutionizing text generation, translation, and automated writing.
- 4. Restricted Boltzmann Machines (RBMs):** RBMs excel in understanding complex data patterns, aiding in tasks like feature learning and topic modeling.



But to simplify it even further, let's talk about Generative Language Models and Generative Image Models:

types of generative ai models

Generative language models

Generative language models learn about patterns in language through training data.

Then, given some text, they predict **what comes next**.

Generative image models

Generative image models produce new images using techniques like diffusion.

Then, given a prompt or related imagery, they **transform random noise into images** or generate images from prompts.

Generative language models learn about patterns in language through training data. Then, given some text, they predict what comes next.

Generative image models produce new images using techniques like diffusion. Then, given a prompt or related imagery, they transform random noise into images or generate images from prompts.

There's a research paper that changed everything, called 'Attention is All You Need'.

The paper that changed everything

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* †
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best

This paper introduced:

1/ Transformers: The paper presented the transformer model, moving away from traditional Deep Learning methodologies that were quite limiting such as Recurrent Neural Network (RNNs) and Convolutional Neural Network (CNNs) in NLP.

2/ Self-Attention Mechanism: Transformers use self-attention to efficiently process different parts of input data.

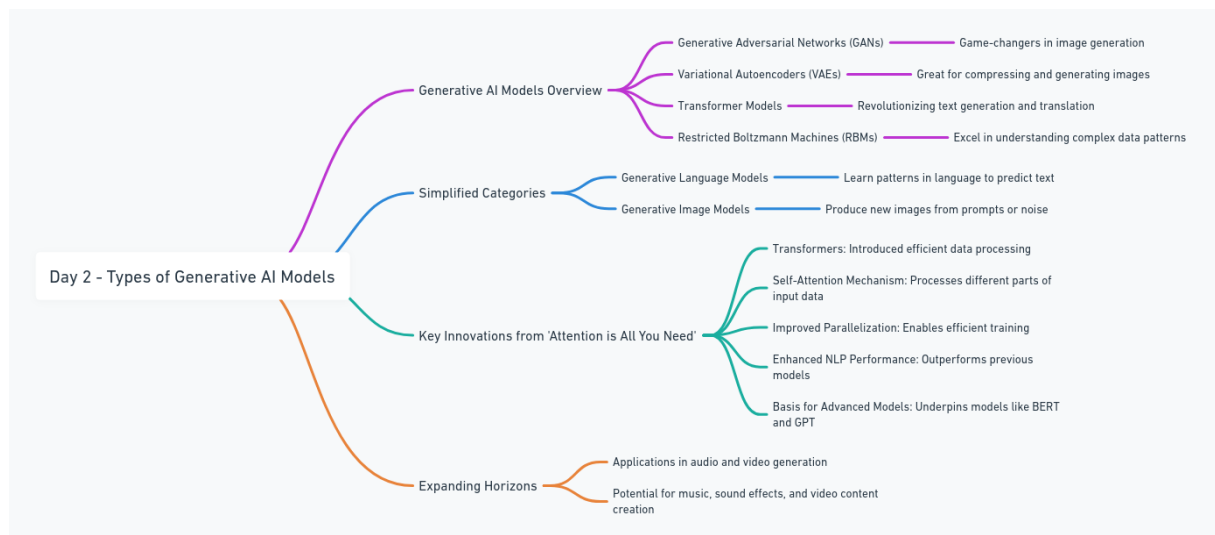
3/ Improved Parallelization: Transformers enable more efficient training through better parallelization compared to RNNs.

4/ Enhanced NLP Performance: They significantly outperform previous models in tasks like machine translation and text summarization.

5/ Basis for Advanced Models: The transformer architecture underpins major NLP models like BERT and GPT, enhancing language processing capabilities.

Here's the in case you want to go deeper, highly recommended to read: [link](#)

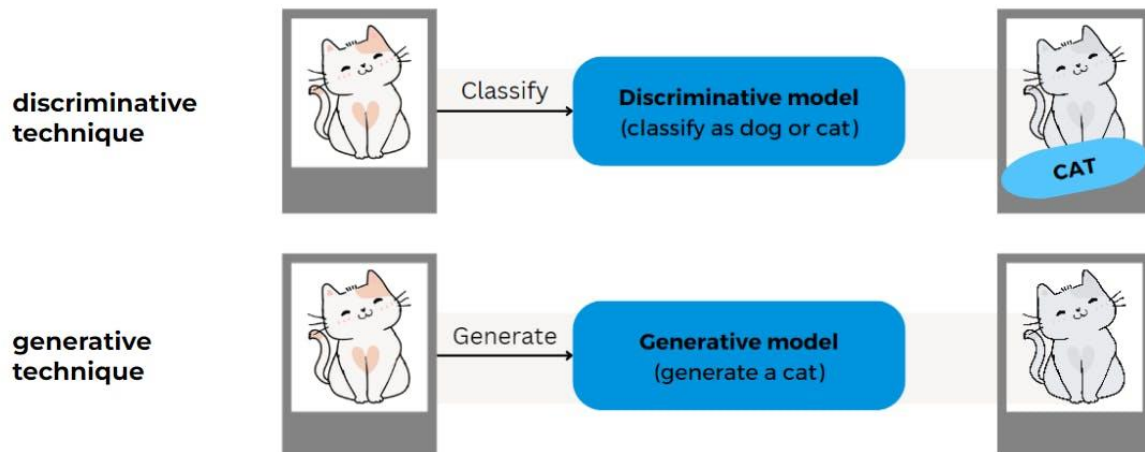
While I've focused on text and image generation, it's exciting to note that similar principles are being applied to audio and video generation, which is likely going to start exploding this 2024 and beyond. AI is now creating music, sound effects, and even generating or altering video content. The potential in these areas is vast and still unfolding.



Day 3 - Traditional ML vs Generative AI or Discriminative vs Generative Models

Welcome to Day 3! Today, we're diving deeper into the AI landscape by contrasting Traditional Machine Learning (ML), focusing on discriminative models, against Generative AI, which revolves around generative models.

To start, let's understand the difference between with a simple example:



Traditional Machine Learning

Discriminative models in Traditional ML are designed to classify or predict outcomes based on input data. They focus on drawing boundaries between different categories and making decisions.

Application Examples: Predictive analytics in business forecasting, spam filters in email systems, and recommendation systems in streaming services.

Key Characteristics:

Supervised Learning: Often relies on labeled data sets to train models. Labeling data is very expensive and time-consuming.

Predictive Accuracy: Emphasizes the accuracy of predictions based on known data.

Analytical Approach: Aims to understand data and draw conclusions.

Generative AI

In contrast, Generative AI doesn't just analyze data; it creates new data that didn't exist before. It's about innovation and creation, generating new content that is similar to but distinct from the training data.

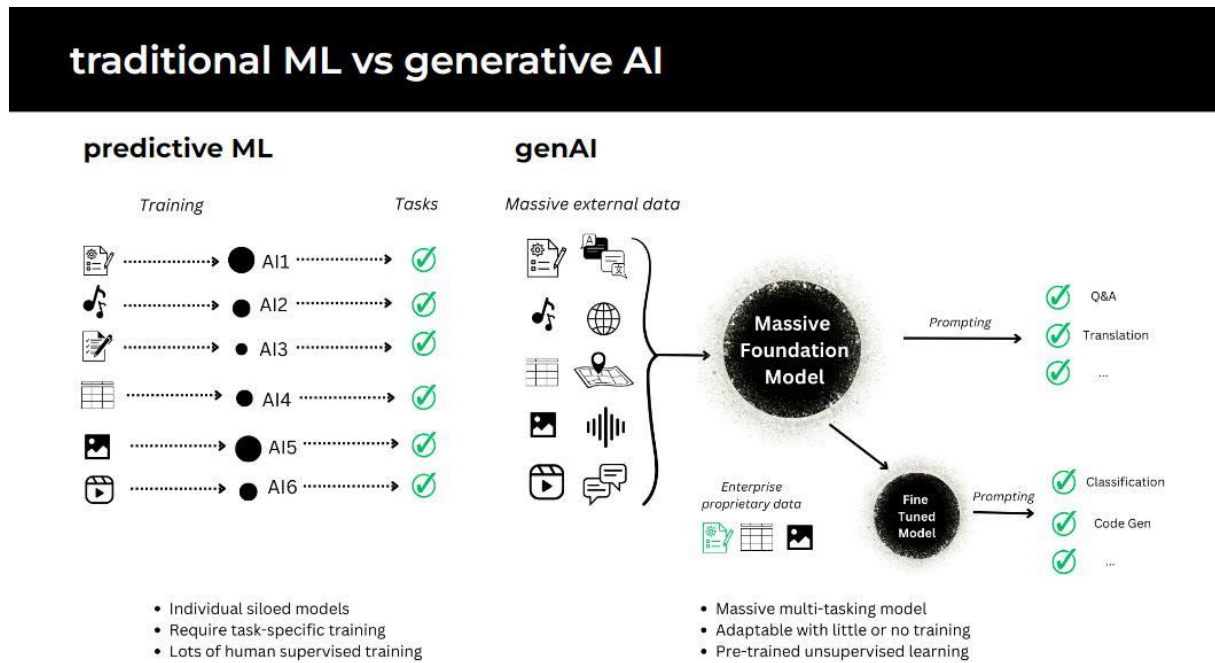
Application Examples: Creating new images or artwork, generating realistic human-like text, or composing music.

Key Characteristics:

Creative Output: Produces new content, extending beyond analysis.

Model Types: Uses models like GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) for content generation.

Innovation Focused: Pushes the boundaries of what machines can create.



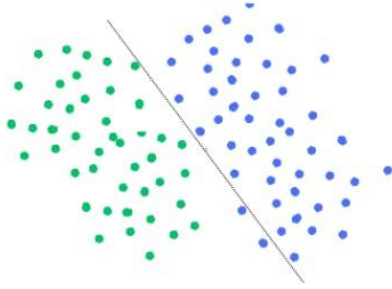
The Role of Discriminative vs Generative in AI

The distinction between discriminative and generative models is vital. Discriminative models excel in classification and prediction tasks, making them suitable for analytical applications. In contrast, generative models are unparalleled in their ability to create and innovate, making them ideal for tasks requiring new content generation.

deep learning model types

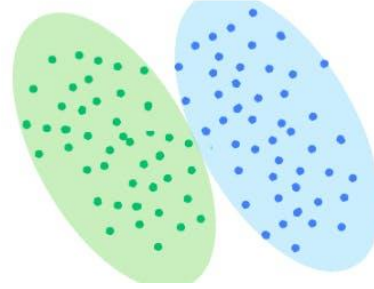
Discriminative

- Used to classify or predict
- Typically trained on a labeled dataset
- Learns the relationship between the features of the labeled data points

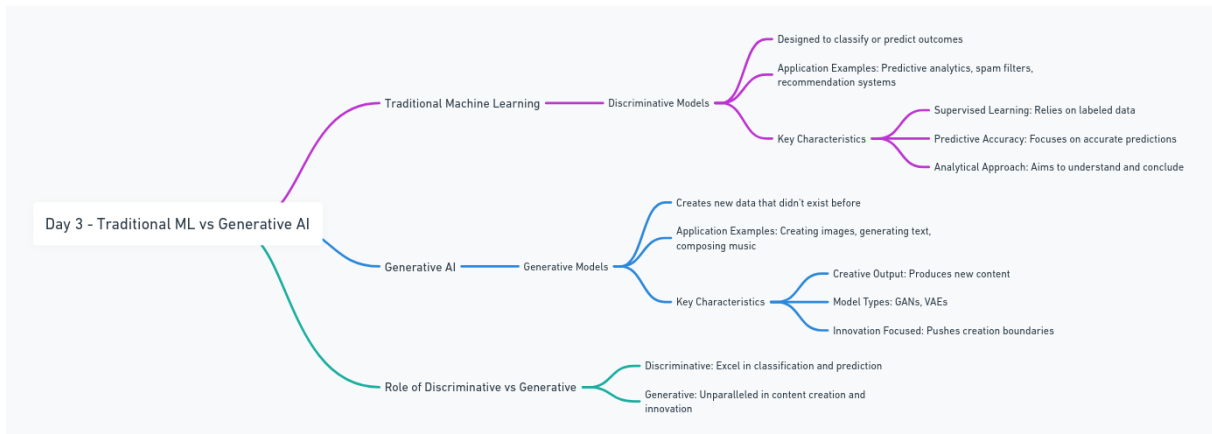


Generative

- Generated new data that is similar to data it was trained on
- Understands distribution of data and how likely a given examples is
- Predict next word in a sequence



Understanding whether your business needs to analyze and classify existing data or generate new, unseen content will guide you in choosing the right AI approach.



Day 4 - Demystifying GPUs in AI

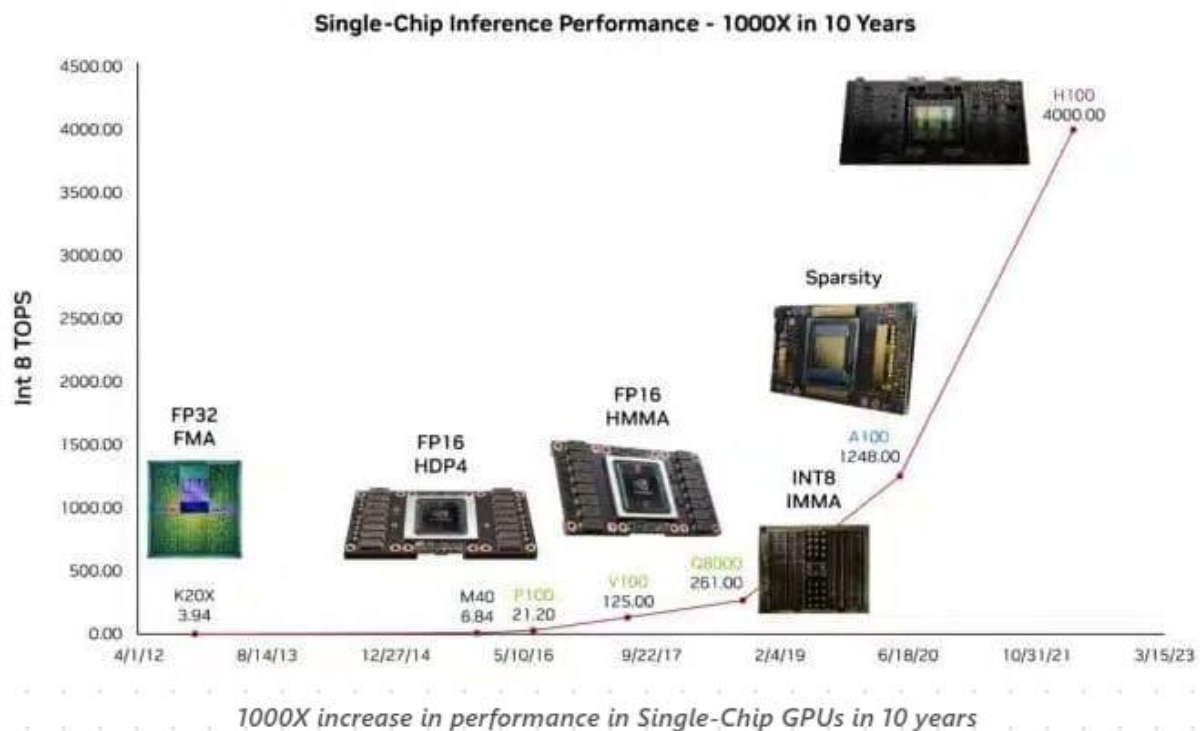
Welcome to Day 4! Today, we're focusing on GPUs - the driving force behind the AI revolution, especially in training and inference tasks.

What are GPUs?

GPUs, or Graphics Processing Units, were initially designed for rendering graphics and video tasks. Remember the excitement of upgrading your PC with a new GPU for better gaming? That same technology has become pivotal in AI, thanks to its architecture and parallel processing capabilities.

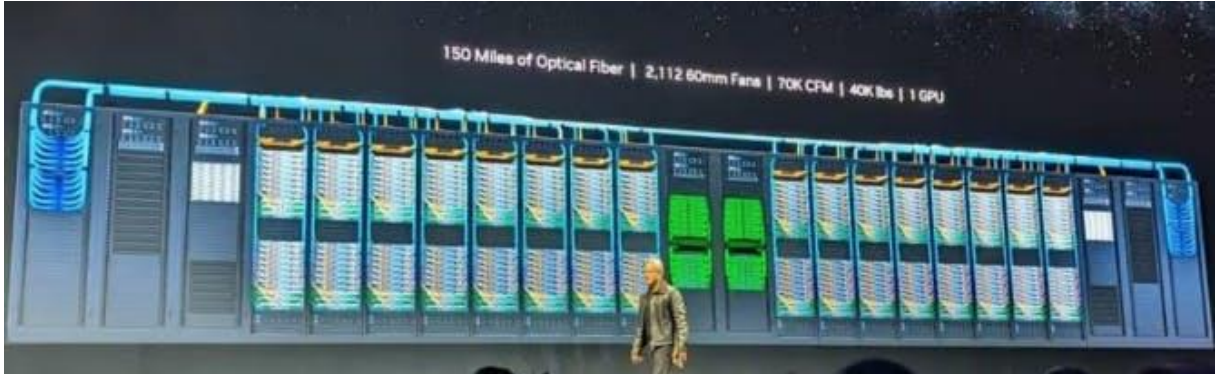
Why are GPUs Crucial for AI?

- **GPU Architecture:** GPUs have thousands of small cores (organized into Streaming Multiprocessors) designed for parallel computing. This setup is perfect for AI workloads, which often require simultaneous processing of large data sets.
- **Parallel Processing Power:** GPUs excel at performing multiple calculations at once, making them ideal for handling the complex mathematical operations needed in AI.
- **Speed and Efficiency:** With rapid thread switching and high memory latency tolerance, GPUs significantly reduce the time needed for training neural networks, like GPT-3, which requires 300 zettaflops of computing power.
- **AI Framework Support:** Manufacturers like Nvidia, AMD, and Intel have optimized GPUs for AI frameworks such as TensorFlow and PyTorch.



The Rise of AI Supercomputers

AI research has skyrocketed with the advent of AI supercomputers, clusters of GPUs working together. These supercomputers, like Summit, Sierra, and Fugaku, are pushing the boundaries in fields like scientific research and climate modeling.



CEO of Nvidia presenting DGX GH200 AI Supercomputer

Selecting the Right GPU

Choosing the appropriate GPU involves understanding your needs - whether it's for AI training, inference, or both, and balancing factors like budget, performance, compatibility, and scalability.

The Future of GPUs in AI

We're witnessing a 1000X increase in single-chip GPU performance over the last decade. The future holds more specialized AI chips, quantum computing advancements, and edge AI integrations, further transforming the AI landscape.

In Conclusion, GPUs have transitioned from enhancing our gaming experiences to becoming the backbone of AI, fueling advancements in machine learning and beyond. Their role in accelerating AI workloads is indisputable and will continue to shape the future of technology.

Demystifying GPUs in AI

What are GPUs?

- Initially designed for graphics and video
- Pivotal in AI for architecture and parallel processing

Why are GPUs Crucial for AI?

- GPU Architecture: Thousands of small cores for parallel computing
- Parallel Processing Power: Ideal for AI's complex calculations
- Speed and Efficiency: Reduces training time for neural networks
- AI Framework Support: Optimized by manufacturers like Nvidia, AMD, and Intel

The Rise of AI Supercomputers

- Clusters of GPUs working together
- Examples: Summit, Sierra, Fugaku
- CEO of Nvidia presents DGX GH200 AI Supercomputer

Selecting the Right GPU

- Considerations: Needs, budget, performance, compatibility, scalability

The Future of GPUs in AI

- 1000X increase in performance over the last decade
- Trends: Specialized AI chips, quantum computing, edge AI

Conclusion

- Transition from gaming to AI backbone
- Accelerating AI workloads, shaping future technology

Day 5 - What it Takes to Train a Foundation Model

Welcome to Day 5 of our AI journey! Today, I will focus on the reasons why training your foundation model can be a pivotal step for your business but it is a decision that has to be taken very wisely.

Control Over the Model

Tailored Approach: When you train your model, you control the data and parameters, allowing you to tailor it to specific styles or domains. This customization ensures the model aligns perfectly with your business needs.

Improved Performance

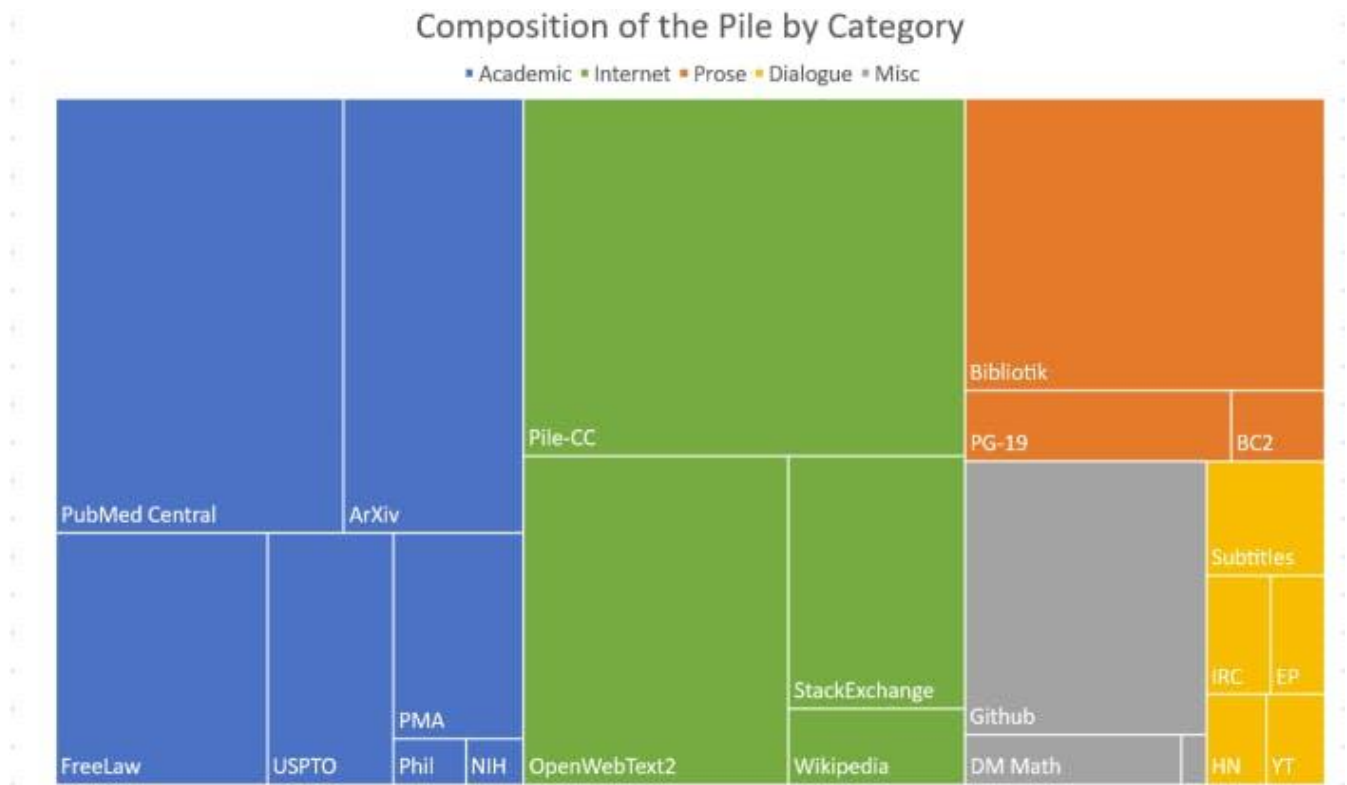
State-of-the-Art Results: Foundation models trained on large, diverse datasets can outperform pre-trained models, especially if your dataset is domain-specific.

Customization

Modifying Architecture: Building your own model means you can alter aspects like tokenizer, vocabulary size, or model architecture, which might be necessary if these components are central to your business strategy.

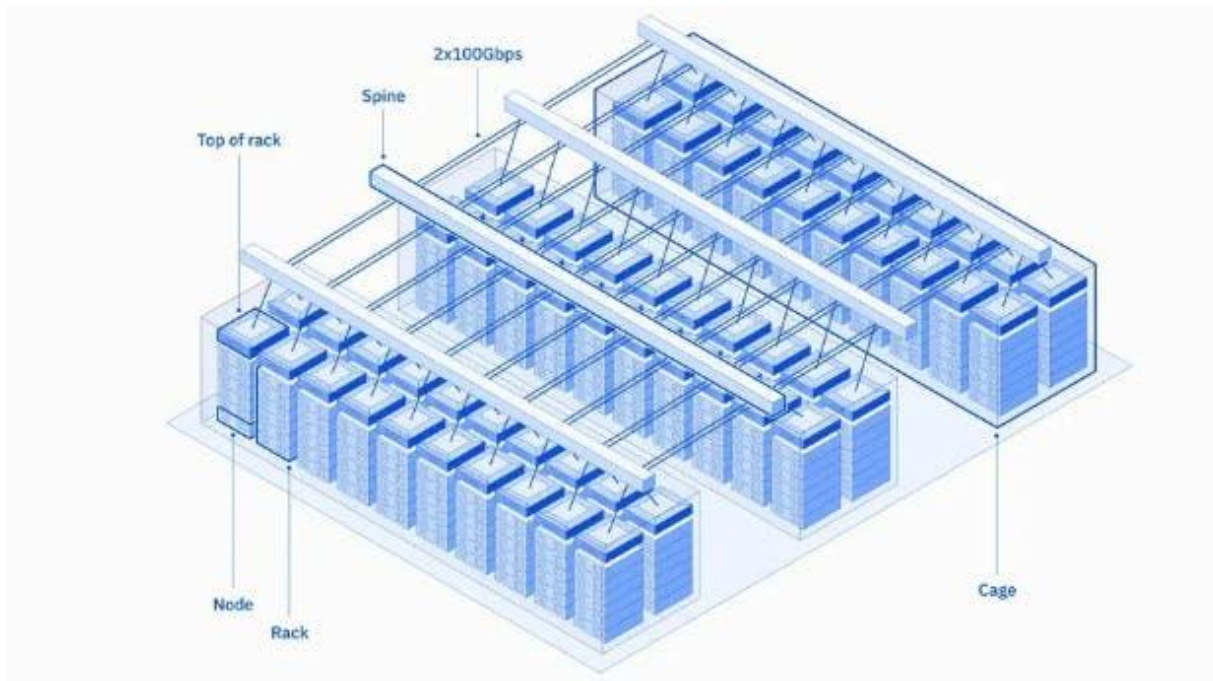
Challenges of Training from Scratch

Data Collection: Amassing a large, relevant dataset is crucial. An example is The Pile, an extensive, diverse language modeling dataset.



The PILE, public data used to train LLMs

Compute Resources: Significant computational power is needed, as demonstrated by AI supercomputers, equipped with thousands of Nvidia A100 and H100 GPUs.



IBM AI Supercomputer VELA

Expertise: Specialized knowledge in AI and ML is essential due to the complexity of model architecture and training processes.

🚀 OpenAI is setting new industry standards with its engineers earning an average annual salary of \$925,000! This includes a base salary of \$300,000 and a whopping \$625,000 in stock-based compensation. Some even earn as much as \$1.4 million! 💰 #AI #OpenAI #TechNews

Training Steps

1. **Dataset Collection:** Gather a large, diverse dataset relevant to your tasks.
2. **Preparation and Tokenization:** Clean and format your data, breaking down text into tokens.
3. **Configure Training:** Set hyperparameters, choose the architecture, and allocate computational resources.
4. **Training:** Train your model using deep learning algorithms.
5. **Evaluation:** Test the model's performance on a separate dataset.
6. **Deployment:** Once satisfied, deploy the model for practical use.

Cost Considerations

Training a foundation model can range from tens of thousands to millions of dollars, depending on the model size, data volume, and computational resources.

Recommendation for Businesses

- **Customize a Pre-Trained Model:** Starting with a pre-trained model and customizing it with techniques like Parameter Efficient Fine Tuning (PEFT) can save time and resources.
- **Consider Needs and Resources:** Evaluate your specific needs and available resources to decide between purchasing, training, or customizing a model.

Customizing foundation models is a great way to get the most out of these powerful tools. It is less expensive, faster, and can give you better performance than training a model from scratch.

In conclusion, while training a foundation model is resource-intensive, it offers unparalleled control and performance, essential for businesses aiming to develop a strong technological edge in AI.

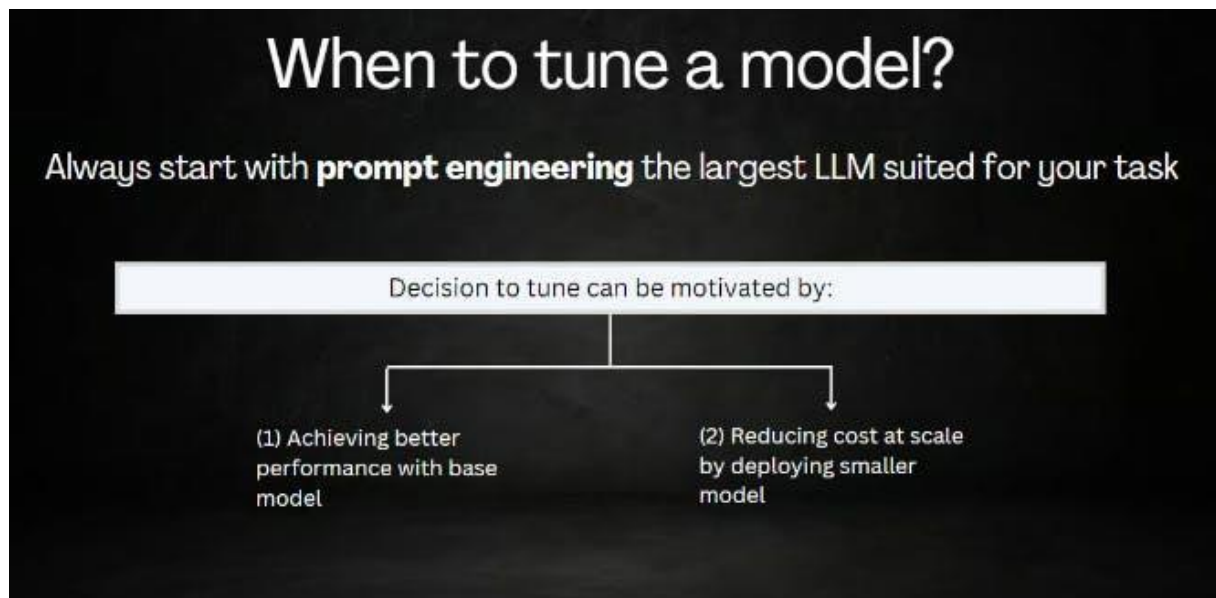


Day 6 - How to customize foundation models

Today, on Day 6 of our AI journey, let's unlock the secrets of customizing foundation models to suit your specific needs. Understanding when and how to tune these models is crucial for optimal performance.

Deciding When to Tune Your Model

Starting Point: Begin with prompt engineering using the largest suitable Language Model (LLM) for your task to gauge if LLMs can handle it. Experiment with various prompt formats and examples.



Prompting Techniques:

- **Zero-Shot Prompting:**

Efficiency with No Extra Data: This involves giving a natural language prompt to generate desired outputs without additional training data.

Example: "Provide a summary of the following passage: [insert text]."

- **One-Shot Prompting:**

A Single Example to Guide: Introduce one example along with your prompt to demonstrate the desired outcome.

Example: "Write marketing copy for Workout Fuel protein shakes in an enthusiastic, punchy voice," along with a high-energy example text.

- **Few-Shot Prompting:**

Leveraging a Few Examples: Provide a handful of examples to establish the pattern or style for the model to replicate.

Example: To generate meeting summaries, give 2-3 examples before asking the model to create new ones.

Technique	Advantages	When to Use
Zero-Shot Prompting	- Simplest to implement - No training data needed	- Quickly test model capabilities - Simple tasks and inferences
One-Shot Prompting	- Can teach complex behaviors - Minimal training data	- Teaching nuanced or contextual responses - When only one good example is available
Few-Shot Prompting	- Teach more robust behaviors - Still very sample efficient	- Targeted enhancement for specific skills - When limited data for a task
Fine-Tuning	- Maximize model performance - Learn complex and nuanced behaviors	- Specializing model for dedicated tasks - When abundant training data exists
Parameter-Efficient Fine-Tuning	- Improved generalization - Faster and lower resource fine-tuning	- Specializing with limited data - Personalizing models for clients

Data-Driven Tuning for Deeper Customization

- **Fine-Tuning:** Adjusting model weights on a specific dataset to cater to your unique objectives, like customizing tone or addressing complex prompts.
- **Parameter-Efficient Fine-Tuning (PEFT):** Delta tuning updates only a small subset of parameters, offering a faster, cost-effective alternative to traditional fine-tuning.

Fine-tuning vs PEFT	
Fine-tuning	Parameter-efficient fine-tuning (PEFT)
Tune ALL model parameters	Tune a small number of (extra) model parameters
Generate a copy of the base model that requires hosting	Generates tiny checkpoints worth a few MBs or less
Requires 1,000s - 100,000s labeled data points	Requires 100s - 1,000s labeled data points
Significant performance gains on target task compared to base model	Comparable to full fine-tuning depending on base model size and data used
Prone to catastrophic forgetting	Overcomes catastrophic forgetting

PEFT Techniques:

Parameter-Efficient Fine-Tuning (PEFT) is a more cost-effective and efficient method because it focuses on optimizing a small subset of model parameters, reducing computational resources and training time while maintaining high-performance levels. There are multiple techniques:

Prefix Tuning: Attaches vectors with free parameters to input embeddings, training them while keeping the LLM frozen.

Prompt Tuning: A simpler variant of prefix tuning, adding a vector only at the input layer.

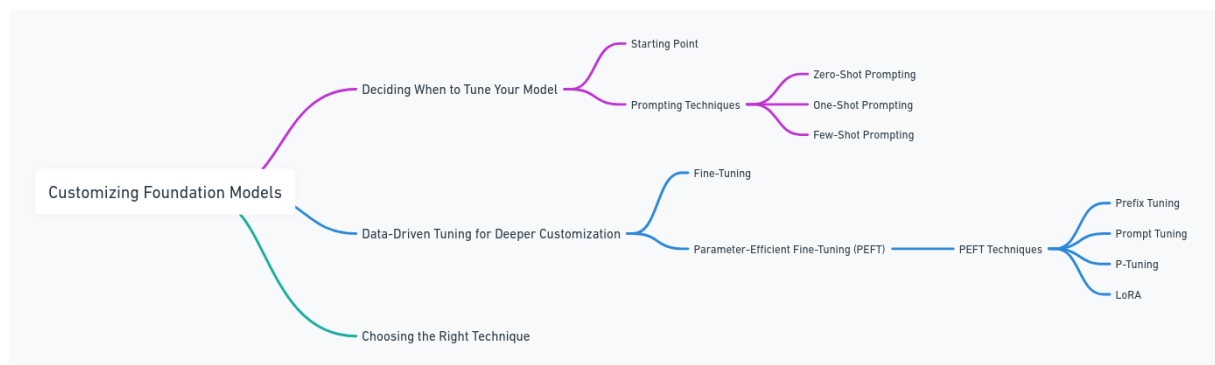
P-Tuning: Automates the search and optimization of prompts using an LSTM model.

LoRA: Low-Rank Adaptation adds update matrices to existing weights, training these new weights.

Choosing the Right Technique:

Goal-Oriented Approach: Select the customization method based on your specific goals and the data you have. For instance, zero-shot and few-shot prompting work well with minimal data, while data-driven tuning is ideal for more complex, data-rich tasks.

Customizing foundation models can significantly enhance their performance on specific tasks, making them more aligned with your business objectives.



Day 7 - The most popular LLMs available

On Day 7 of our AI series, let's navigate the world of Large Language Models (LLMs) by understanding the key differences between open-source and proprietary models and services, and exploring some of the most popular LLMs available today.

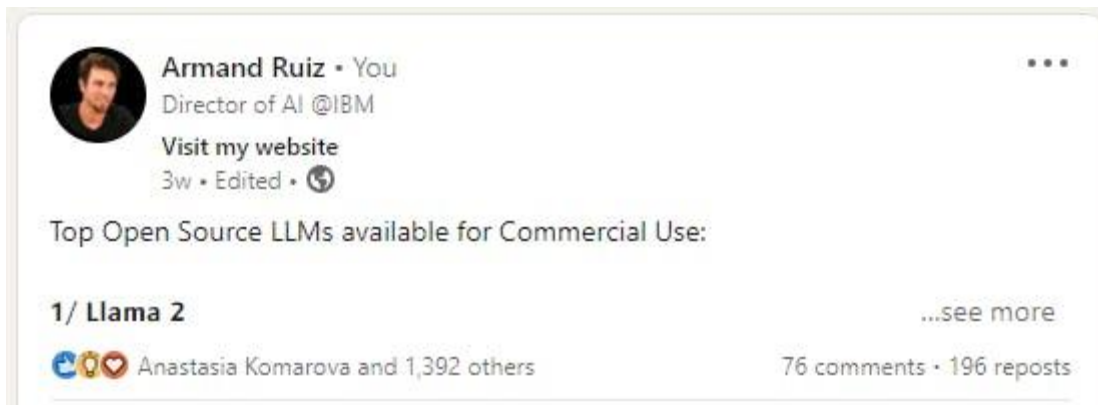
Open Source LLMs:

Accessible and Collaborative: These models are freely available for use, modification, and distribution, promoting community-driven development and innovation.

Examples of Open Source LLMs:

- **GPT-Neo/GPT-J:** Developed by EleutherAI, these models are open-source alternatives to OpenAI's GPT models, offering similar capabilities.
- **BERT:** Developed by Google, BERT has been a groundbreaking model for understanding context in natural language, widely used in various applications.

[Armand Ruiz](#) listed the Top Open Source LLMs a few weeks back on LinkedIn. [Here's](#) the link



Closed Source LLMs:

Commercial and Proprietary: These models are developed and maintained by private entities, often requiring licenses or subscriptions for access.

Examples of Closed Source LLMs:

- **OpenAI's GPT-3/GPT-4:** Known for their advanced capabilities, these models have set benchmarks in generative AI but are accessible mostly through API with usage costs.
- **Google's LaMDA:** A cutting-edge model designed for conversational AI, used internally by Google.

Open Source vs Closed Source

Open Source LLMs:

- Language models with publicly available source code.
- Can be freely accessed, used, modified, and distributed by anyone.
- Promote collaboration, transparency, and community involvement.
- Allow for collective development, innovation, and knowledge sharing.

Closed Source LLMs:

- Source code is not publicly available.
- Developed and maintained by private organizations or companies.
- Often commercial products requiring licenses or subscriptions.
- Architecture, training data, and algorithms are typically proprietary and not disclosed to the public.

The Game Changer: Llama 2

- **Accessibility and Versatility:** Meta's Llama 2 has been released as an open-source AI model, making it accessible for everyone from startups to researchers. Its availability in different sizes (7B, 13B, 70B-parameter models) offers a range of options for fine-tuning and deployment.
- **Innovation and Privacy:** As an open-source model, Llama 2 removes barriers to AI adoption and addresses data privacy concerns by allowing private hosting and customization with your own data.
- **Performance Benchmarking:** Llama 2 stands on par with models like GPT-3.5 in terms of performance, particularly excelling in generating helpful responses for prompts. However, it shows less proficiency in coding tasks compared to other specialized models.
- **Cost and Community Benefits:** Meta's open-sourcing of Llama 2, despite the substantial development cost, taps into the collective wisdom of the global AI community, accelerating innovation and potentially leveling the playing field against closed-source counterparts.

Why the Distinction Matters: Understanding the differences between open source and closed source LLMs is crucial for businesses and developers. Open source models offer transparency and the opportunity for customization, while closed source models, often backed by significant resources and research, provide robust, state-of-the-art capabilities but with usage restrictions and costs.

In your AI endeavors, choosing between open source and closed source LLMs will depend on your specific needs, resources, and goals.

Day 7 - The Most Popular LLMs Available

Open Source LLMs

- Accessible and Collaborative
 - Freely available for use, modification, and distribution
 - Promotes community-driven development and innovation
- Examples
 - GPT-Neo/GPT-J: Developed by EleutherAI, open-source alternatives to OpenAI's GPT models
 - BERT: Developed by Google, groundbreaking for understanding context in natural language

Closed Source LLMs

- Commercial and Proprietary
 - Developed and maintained by private entities
 - Requires licenses or subscriptions for access
- Examples
 - OpenAI's GPT-3/GPT-4: Advanced capabilities, accessible mostly through API with usage costs
 - Google's LaMDA: Designed for conversational AI, used internally by Google

The Game Changer: Llama 2

- Accessibility and Versatility: Released as an open-source AI model by Meta, available in different sizes
- Innovation and Privacy: Removes barriers to AI adoption, addresses data privacy concerns
- Performance Benchmarking: On par with GPT-3.5 in generating helpful responses, less proficient in coding tasks
- Cost and Community Benefits: Accelerates innovation, potentially leveling the playing field against closed-source counterparts

Why the Distinction Matters

- Open source models offer transparency and customization opportunities
- Closed source models provide robust capabilities but with usage restrictions and costs

Day 8 - Generative AI Applications and Use Cases

On Day 8, let's talk about the transformative applications of generative AI in business, examining how these technologies are reshaping various industries and enterprise functions.

Broad Spectrum of Business Applications:

Marketing and Advertising: From crafting compelling ad copy to generating creative landing pages, generative AI is revolutionizing how businesses approach marketing.

Content Creation: AI is now capable of producing news articles and social media content, enhancing digital presence with minimal effort.

Customer Service: By deploying AI-driven chatbots, businesses can ensure engaging, natural conversations with customers, elevating the service experience.

Summarization: Generative AI can distill lengthy reports and papers into concise, informative summaries, aiding decision-making and research.

Data Analysis: AI tools can sift through vast datasets, uncovering patterns and insights that drive strategic decisions.

Personalization: Tailoring content to individual user preferences or customer segments is now more efficient with AI's generative capabilities.

Product Development: Rapid prototyping and testing of new product designs are made possible, speeding up the innovation process.

most common generative AI tasks for business

Retrieval-Augmented Generation Based on a document or dynamic content, create a chatbot or question-answering feature. <i>Building a Q&A resource form a broad knowledge base, providing customer service assistance.</i>	Summarization Condenses textual information into concise summaries, like summarizing customer feedback. <i>Conversation summaries, insurance coverage, meeting transcripts, contact information.</i>	Content Generation Generate text content for a specific purpose. <i>Marketing campaigns, job descriptions, blog posts and articles, email drafting support.</i>
Named Entity Recognition Identify and extract essential information from unstructured text. <i>Audit acceleration, SEC 10k fact extraction</i>	Insight Extraction Analyze existing unstructured text content to surface insights in specialized domain areas. <i>Medical diagnosis support, user research findings.</i>	Classification Read and classify written input with as few as zero examples. <i>Sorting of customer complaints, thread and vulnerability classification, sentiment analysis.</i>

Key Benefits for Businesses:

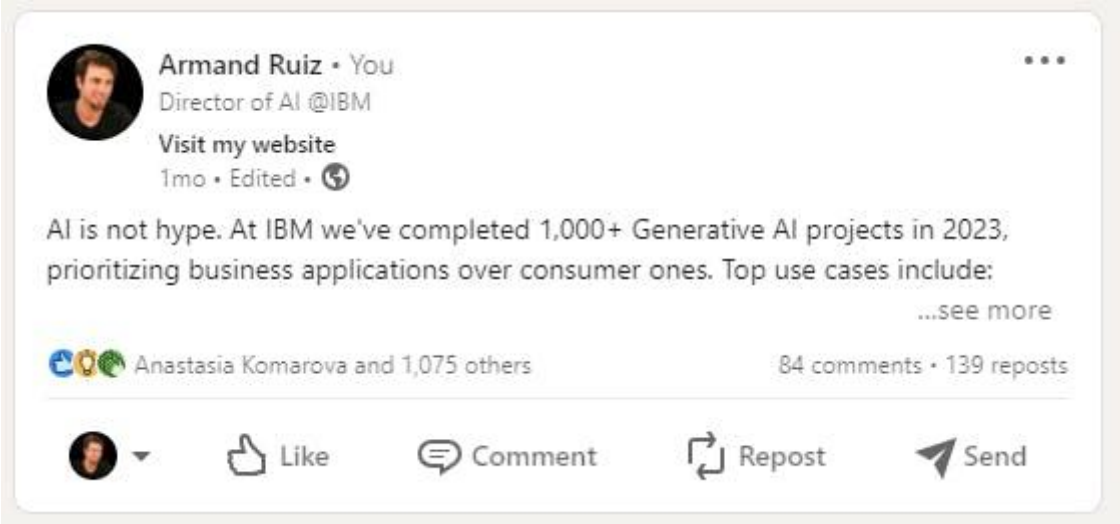
Increased Efficiency: Automate repetitive content generation tasks, freeing up valuable time for strategic work.

Cost Savings: Reduce reliance on expensive human labor, particularly in creative and writing tasks.

Consistency and Quality: Maintain a consistent brand voice while leveraging AI's analytical capabilities to produce high-quality content.

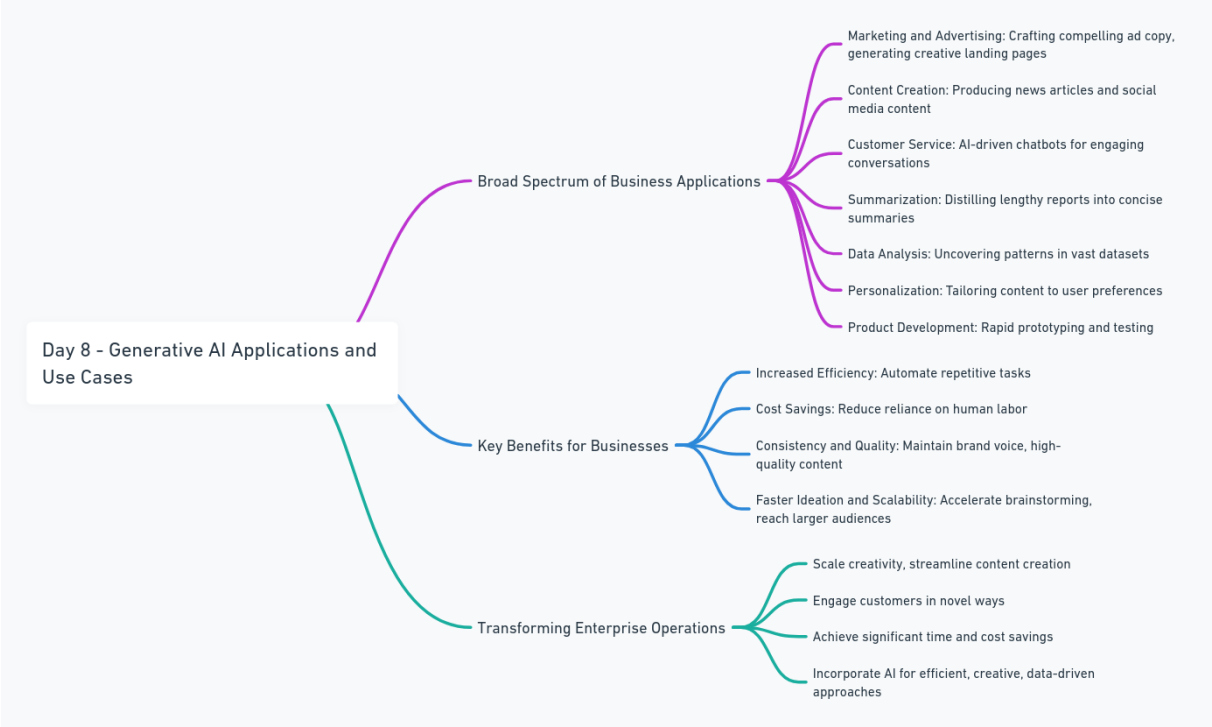
Faster Ideation and Scalability: Accelerate the process of brainstorming and content production, enabling businesses to reach larger audiences more effectively.

See [Armand Ruiz](#) detailed list of Use Cases in this recent LinkedIn post. [Here's](#) the link



Transforming Enterprise Operations: Generative AI is not just a tool; it's a game-changer for business operations. It enables businesses to scale creativity, streamline content creation, engage customers in novel ways, and achieve significant time and cost savings. The integration of AI into these functions is transforming how businesses interact with their customers, manage their internal processes, and innovate in their product offerings.

Incorporating generative AI into business strategies can lead to more efficient, creative, and data-driven approaches, opening new avenues for growth and competitive advantage.



Day 9: The Generative AI Application Development Stack

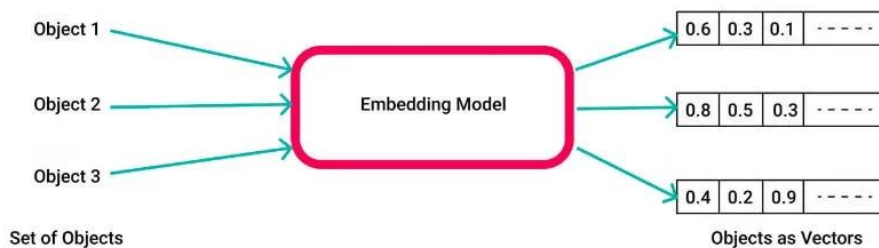
Welcome to Day 9! Today, we're diving into the architecture of the Generative AI (GenAI) stack, crucial for crafting customized GenAI applications.

The GenAI Stack: A Modular, Integrated System

- **Understanding the GenAI Architecture:** This system includes data pipelines, training and inference engines for LLMs, model registries, deployment monitoring, and user interfaces. Tools like LangChain offer orchestration layers for rapid transitions from data to models to apps.

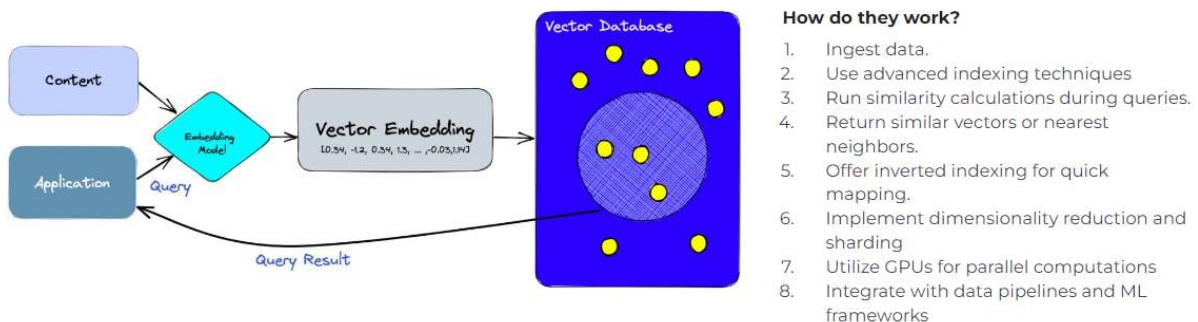
Key Elements of the GenAI Stack:

1- **Embeddings (Vectors):** These transform high-dimensional data into lower-dimensional vectors, retaining essential information in a more manageable form.

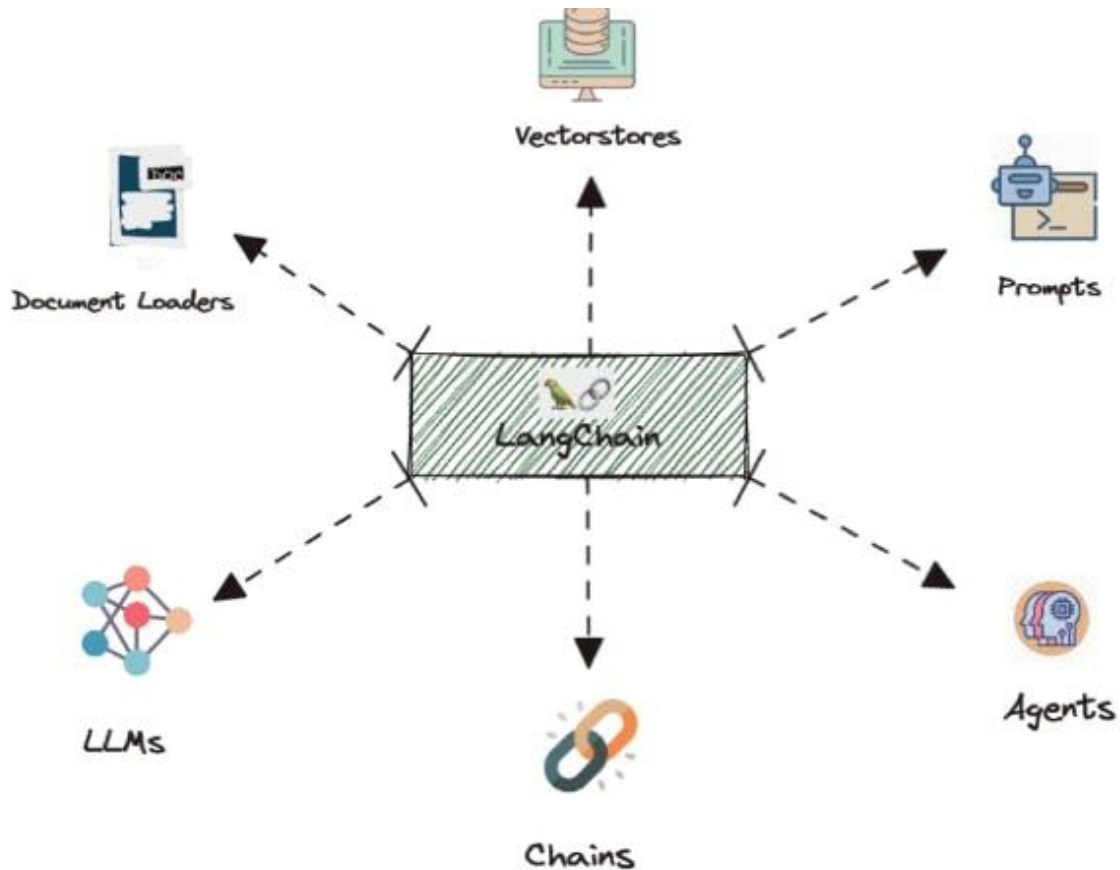


Embeddings convert messy real-world data into mathematical representations capturing hidden relationships. This transformed data powers cutting-edge AI.

2- **Vector Database:** Stores and indexes vector representations for quick retrieval, supporting operations like vector search and similarity rankings, forming the backbone of vector infrastructure in AI.



3- **LangChain:** An open-source framework built around LLMs, LangChain facilitates the design and development of various GenAI applications, including chatbots and Generative Question-Answering (GQA).



4- **LLMs and Prompts**: The core of generative capabilities, LLMs respond to prompts to generate text, making them essential for applications like content creation and customer service.

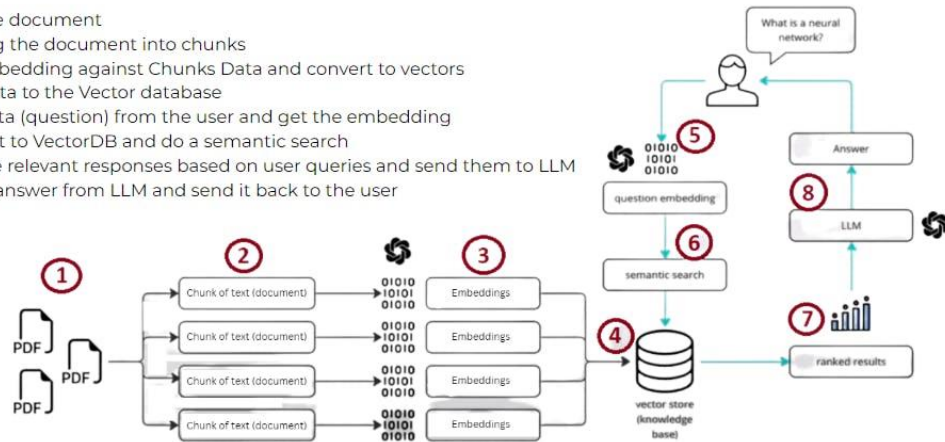
Building a Simple GenAI App - Step-by-Step:

1. **Load Document**: Begin by loading the document or data source.
2. **Split into Chunks**: Break the document into manageable parts.
3. **Create Embeddings**: Convert these chunks into vector representations using embeddings.
4. **Store in Vector Database**: Save these vectors in the database for efficient retrieval.
5. **User Interaction**: Receive queries or input from the user and convert them into embeddings.
6. **Semantic Search in VectorDB**: Connect to the vector database to perform a semantic search based on the user's query.
7. **Retrieve and Process Responses**: Fetch relevant responses, pass them through an LLM, and generate an answer.
8. **Deliver Answer to User**: Present the final output generated by the LLM back to the user.

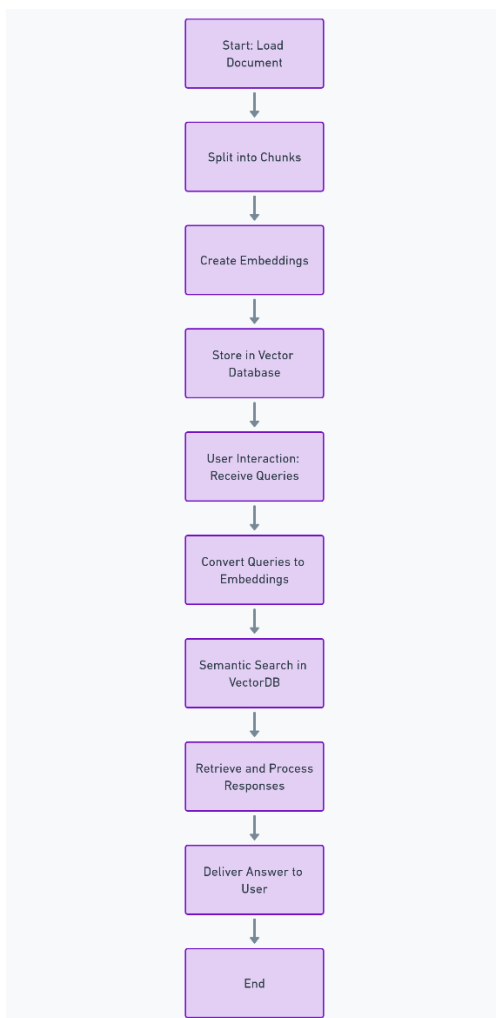
components of a GenAI solution

Steps:

- Step 1: Load the document
- Step 2: Splitting the document into chunks
- Step 3: Use Embedding against Chunks Data and convert to vectors
- Step 4: Save data to the Vector database
- Step 5: Take data (question) from the user and get the embedding
- Step 6: Connect to VectorDB and do a semantic search
- Step 7: Retrieve relevant responses based on user queries and send them to LLM
- Step 8: Get an answer from LLM and send it back to the user



Understanding and utilizing the components of the GenAI stack is key for businesses looking to leverage AI for innovative applications. This modular approach allows for customization and scalability, fitting various business needs and goals.



Day 10: The Emergence of Small Language Models

On Day 10, we focus on the emerging trend of Small Language Models (SLMs) in the business world and their growing importance alongside Large Language Models (LLMs).

Understanding Large Language Models: A Refresher

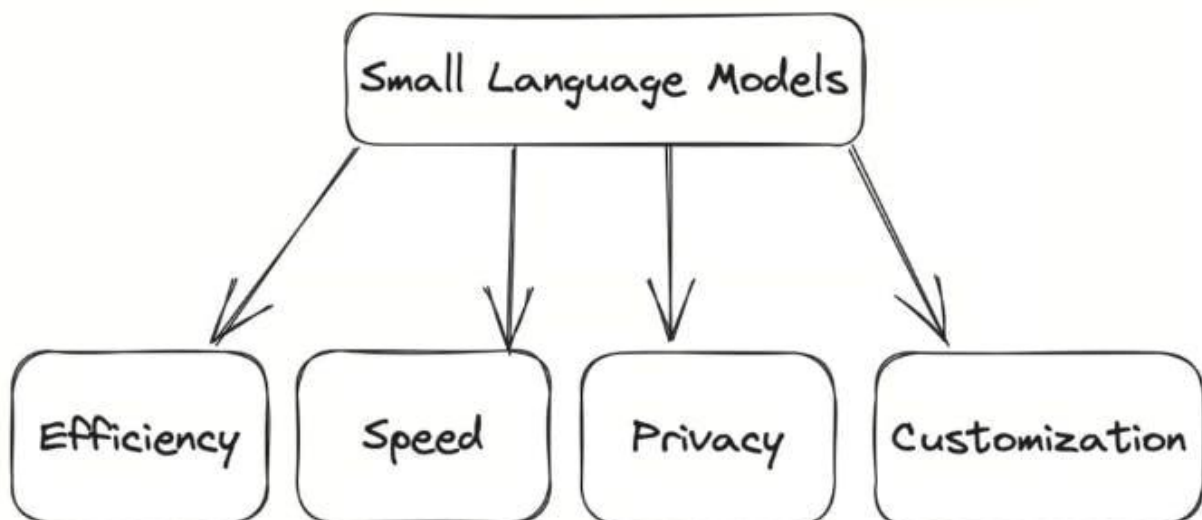
- **Foundation of Today's NLP:** LLMs, trained on vast text data, excel in generating coherent text and performing complex language tasks.
- **Size and Complexity:** Models like GPT-3 (175 billion parameters) and PaLM (540 billion parameters) represent the massive scale of LLMs, offering advanced capabilities but sometimes leading to challenges in accuracy and behavior.

but LLMs are very expensive:

running ChatGPT costs approximately \$700,000 a day

What are Small Language Models (SLMs)?

- **Defining SLMs:** Generally defined as models with up to 20 billion parameters, SLMs are tailored for specific business tasks like chat, analytics, and content generation.
- **Agility and Customization:** SLMs offer a balance of capability and control, making them well-suited for focused business applications.



Advantages of SLMs

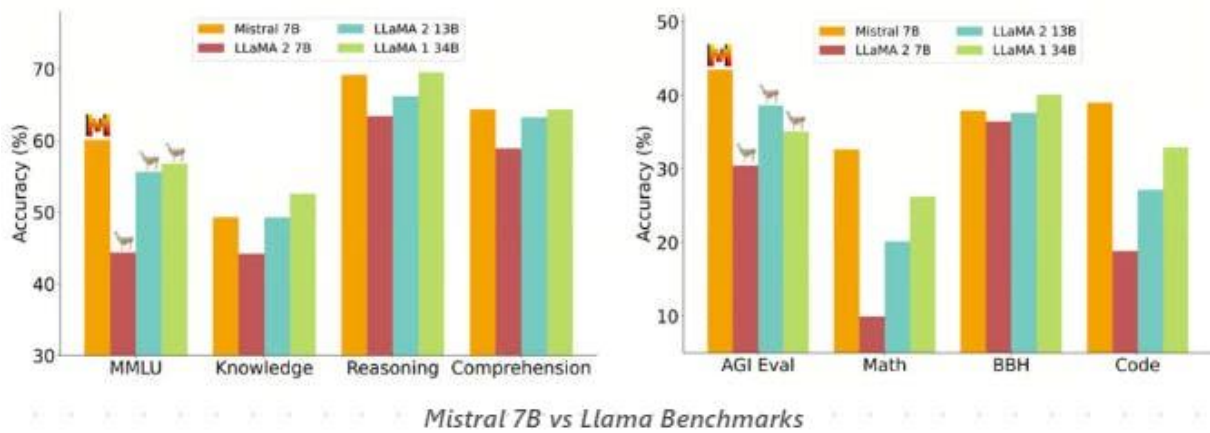
- **Development and Risk Control:** Easier to build and modify, SLMs reduce risks like bias and hallucinations due to simpler knowledge representations.
- **Efficiency and Sustainability:** Being lightweight and less computationally intensive, SLMs are ideal for deployment on smartphones and edge devices, contributing to sustainability.

- **Cost-Effectiveness:** SLMs offer significant cost savings, making AI more accessible for businesses.

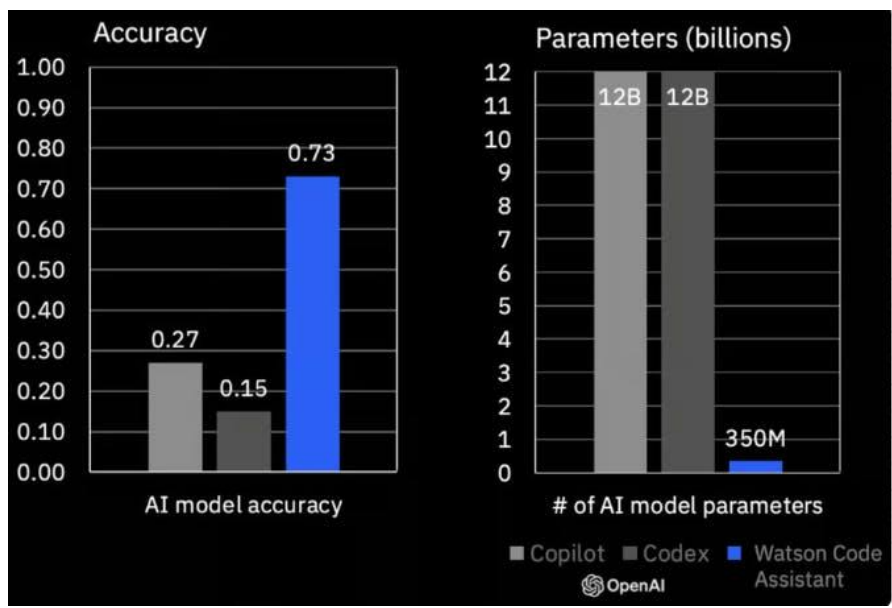
The speed of learning SLMs allow is huge, too. They're within the reach of so many more teams at lower cost. It just lets more innovation cycles happen faster - Brad Edwards

Benchmarking SLMs Against LLMs

- **Performance Comparisons:** For instance, Mistral 7B outperforms larger models in certain benchmarks, demonstrating that SLMs can compete with or even surpass LLMs in specific tasks.



- **Focused Training:** SLMs like IBM Granite, despite smaller size and data, show competitive performance due to targeted training on industry-specific data.



Tuning Small Language Models

- **Customization Techniques:** Similar to LLMs, SLMs can be fine-tuned using various methods to enhance performance for specific use cases.
- **Example of Tuning:** IBM's Granite series, for instance, underwent specialized training for coding, showing how SLMs can be tailored to specific domains.

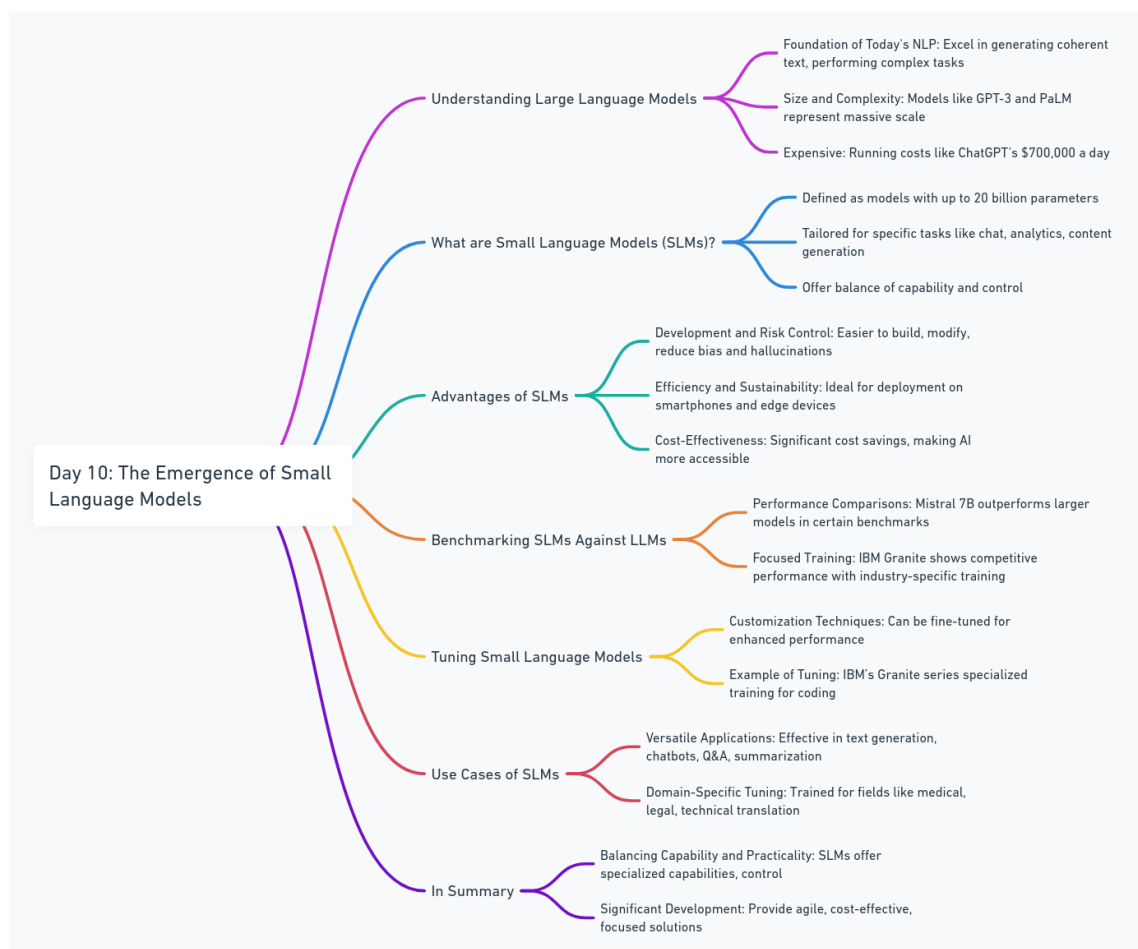
Use Cases of SLMs

- **Versatile Applications:** SLMs are effective in text generation, chatbots, Q&A, and summarization, offering optimized solutions for resource-limited scenarios.
- **Domain-Specific Tuning:** SLMs can be trained for specialized fields like medical, legal, or technical translation, offering more accurate and relevant outputs than general-purpose LLMs.

In Summary

- **Balancing Capability and Practicality:** SLMs are emerging as a practical alternative to LLMs in many business scenarios, offering a mix of specialized capabilities and control.

SLMs represent a significant development in the AI landscape, providing businesses with more agile, cost-effective, and focused solutions for integrating AI into their operations.



Day 11: The AI Engineer Profession and Skills

Welcome to Day 11, where we explore the evolving and dynamic role of AI Engineers in the rapidly advancing field of Generative AI.

AI Engineering: A New Frontier

- **Role Definition:** AI Engineers are the architects behind practical AI applications, handling everything from the development to the deployment of AI systems.
- **Emerging Importance:** As AI capabilities grow, particularly with the advent of Foundation Models, AI Engineers have transitioned from niche specialists to key players in tech and business landscapes.

[Armand Ruiz](#) believe the AI Engineer will be the highest-demand engineering job of the decade.

Responsibilities of AI Engineers

- **AI Infrastructure:** Develop and manage robust AI systems, ensuring scalability and efficiency.
- **Advanced Prompting Strategies:** Utilize tools like LangChain for sophisticated prompt engineering with LLMs.
- **Data Management and Model Operations:** Master data preprocessing and embedding techniques, and manage a diverse range of language models for varied applications.
- **AI Model Integration:** Transform AI models into accessible APIs for seamless integration with software systems.

Why AI Engineering is the Future

- **Demand and Recognition:** With AI integration becoming crucial for businesses, AI Engineers are in high demand for their ability to turn AI advancements into practical solutions.
- **Diverse Backgrounds:** Professionals in this field are proving that diverse backgrounds can contribute significantly to AI product development, beyond traditional academic pathways.

The Path to Becoming an AI Engineer

- **Foundational Skills:** Master programming (Python), machine learning, deep learning, and cloud computing.
- **Recommended Courses:** Consider enrolling in courses like IBM AI Engineering Professional Certificate or DeepLearning.ai's specialized courses in AI and LLM application development.
- **Portfolio Development:** Build a portfolio showcasing your AI projects to demonstrate your skills to potential employers.

Here's a good list of courses to start your AI Engineering journey: [link](#)



AI Engineer vs Data Scientist

- **Role Distinction:** Data scientists focus on data analysis and model building, while AI Engineers specialize in building and deploying AI systems and infrastructure.
- **Unique Contribution:** AI Engineers are pivotal in implementing large-scale AI systems, like LLMs, in practical, operational environments.

Factor	AI Engineer	Data Scientist
Responsibilities	Build and deploy AI systems. Work on the underlying infrastructure that powers AI systems.	Collect, clean, analyze, and interpret data. Build and deploy machine learning models.
Skills	Machine learning, programming, cloud computing, distributed systems	Statistics, machine learning, programming, data visualization
Career path	Software engineer → machine learning engineer → AI engineer	Data analyst → data scientist → machine learning engineer → AI engineer
Salary	\$110,000 - \$150,000	\$98,000 - \$137,000
Job outlook	Very good	Very good

Difference between AI Engineer and Data Scientist

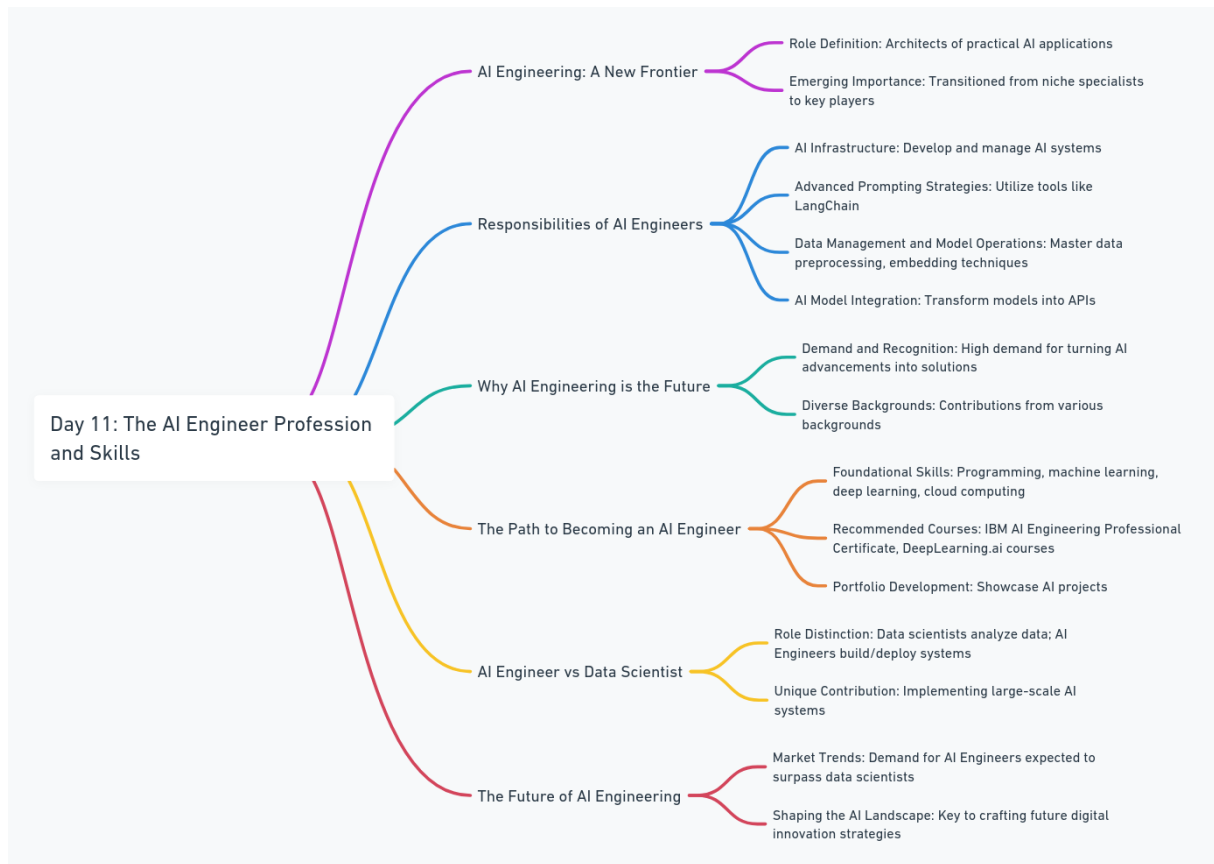
The Future of AI Engineering

- **Market Trends:** The demand for AI Engineers is expected to surpass that of data scientists as the field continues to evolve towards more complex, large-scale AI implementations.
- **Shaping the AI Landscape:** AI Engineers are key to crafting strategies and architectures that will define the future of digital innovation, making this role not just current but increasingly vital in the AI-driven future.

AI Engineers stand at the forefront of the AI revolution, turning the promise of AI into tangible, valuable applications. This profession is not just about technical prowess; it's about shaping the future of how we interact with technology.

Armand Ruiz prediction:

In numbers, there are probably going to be significantly more AI Engineers than there are data scientists.



Day 12 - Ethical Considerations in AI

As we approach the end of our AI series, Day 12 is dedicated to understanding the ethical considerations in AI, particularly the risks and responsibilities associated with deploying these powerful technologies.

AI is the most powerful technology ever created.

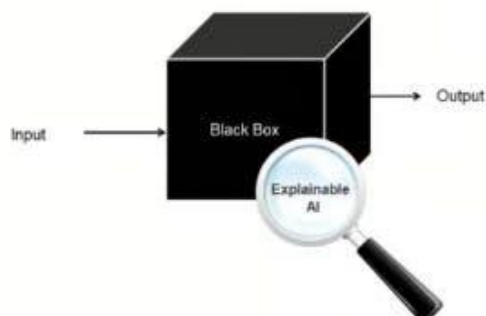
Understanding the Risks of AI

- **Bias, Fairness, and Accuracy:** AI systems can inadvertently replicate human biases present in training data, leading to unfair and inaccurate outcomes. Continuous improvement in data and training practices is crucial to enhance AI fairness.
- **Job Disruption:** AI's potential to automate tasks poses challenges for the job market, necessitating re-skilling initiatives and policy responses.
- **Weaponization:** The integration of AI into weapons systems raises ethical concerns about autonomy in warfare and the need for international governance.
- **Cybersecurity Vulnerabilities:** AI systems can be exploited for adversarial attacks, data poisoning, and model hacking, underscoring the importance of robust cybersecurity measures.
- **Misinformation Spread:** AI's ability to generate convincing fake news requires vigilance and tools to detect and mitigate the spread of misinformation.
- **Emergence of AGI:** The prospect of Artificial General Intelligence (AGI) adds another layer of ethical complexity, with implications for the future of humanity.

AI creating extinction risk for humanity is widely overhyped. AI develops gradually, and the “hard take off” scenario, where AI suddenly achieves superintelligence overnight is not realistic.

lack of transparency

Most AI systems act as "black boxes", with inner workings opaque to users.



Impacts

1. Inability to audit systems for errors, issues, or harm
2. Reduced accountability for creators
3. User confusion and frustration

AI systems still act as black boxes, which impacts on trust

The Role of AI Engineers in Mitigating Risks

- **Infrastructure Development:** AI Engineers are responsible for developing and managing AI infrastructure, ensuring systems are robust, scalable, and secure.
- **Ethical Implementation:** Part of their role involves applying ethical AI practices, such as prompt engineering and data management, to minimize risks like bias and inaccuracy.
- **Collaboration and Best Practices:** AI Engineers must collaborate across functions to promote AI best practices and ethical standards within their organizations.

Mitigating Harmful AI Outputs

- **Hallucinations and Fabrications:** AI systems can generate plausible but incorrect content. It's essential to design systems that minimize these risks.
- **Data Poisoning and Toxic Language:** AI must be safeguarded against harmful data inputs and language generation, requiring ongoing model training and content filtering.
- **Unstable Task Performance:** Addressing the inconsistent performance of AI models involves careful prompt engineering and verification processes.
- **Human Oversight:** Incorporating human review in high-stakes AI applications provides an additional layer of safety and accuracy.

The Path Forward: Ethical AI

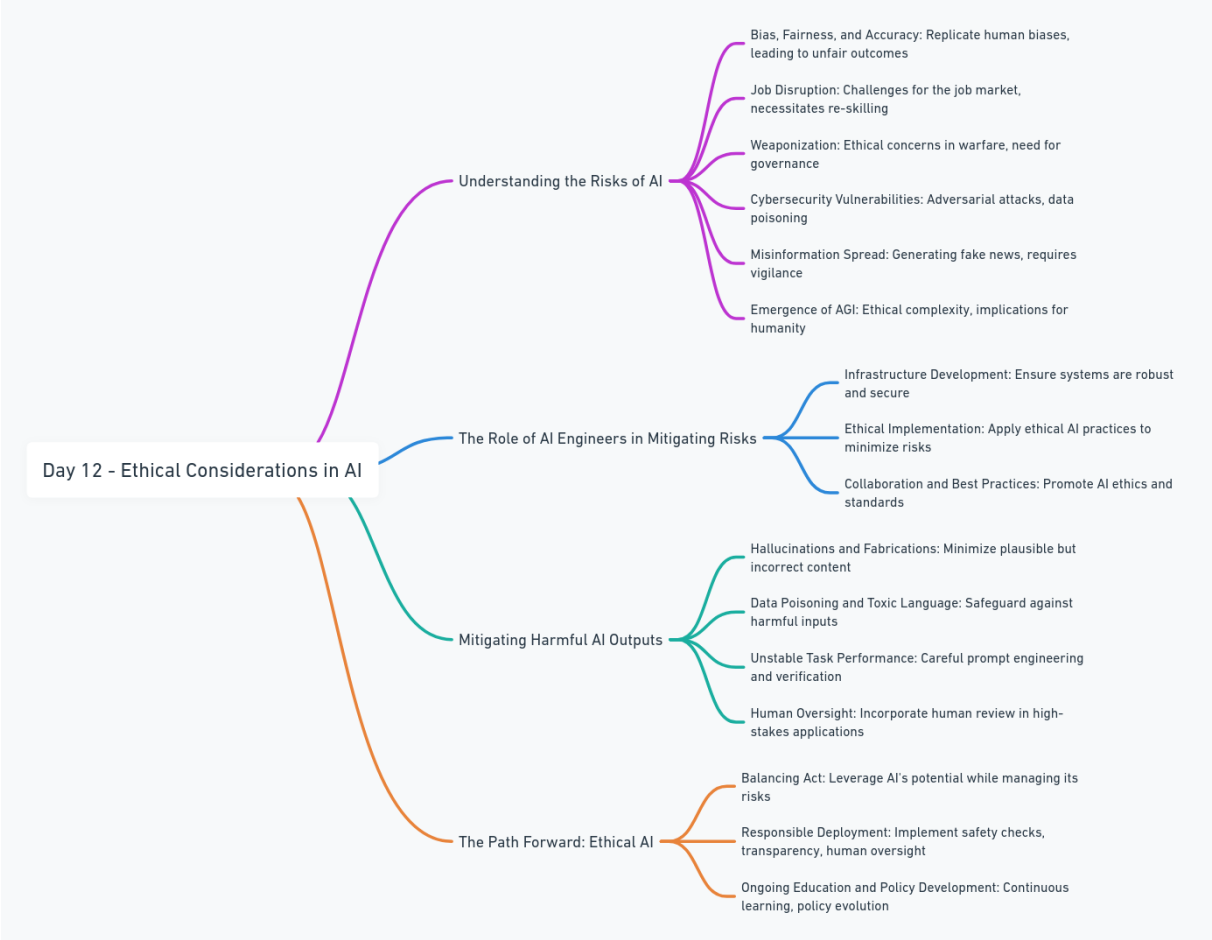
- **Balancing Act:** The challenge lies in leveraging AI's potential while responsibly managing its risks and ethical implications.
- **Responsible Deployment:** Implementing safety checks, transparency measures, and human oversight is crucial for ethical AI deployment.
- **Ongoing Education and Policy Development:** Continuous learning and policy evolution are necessary to keep pace with AI advancements and ensure its beneficial use.

Foundation properties for AI Ethics

AI ethics provides guidelines to mitigate risks like amplifying human cognitive biases, enable responsible innovation, avoid harmful outcomes, and maintain public trust given AI's potential to rapidly scale both benefits and risks.

fairness	accountability	transparency	privacy & security
-----------------	-----------------------	---------------------	-------------------------------

AI holds tremendous promise for transforming businesses and society. However, it's crucial to approach AI development and deployment with a keen awareness of its ethical implications, ensuring that these powerful tools are used responsibly and for the greater good.



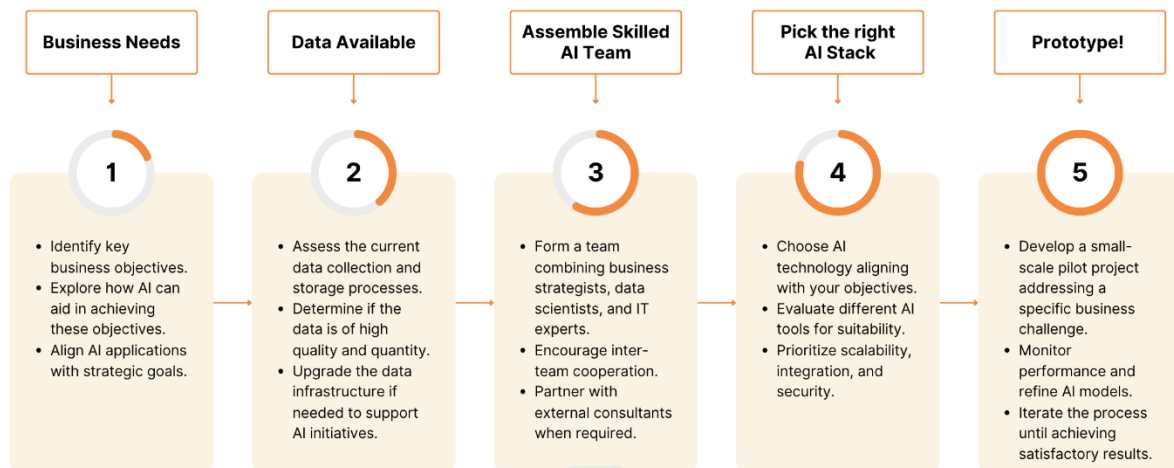
Day 13 - Create Your AI for Business Roadmap

Welcome to Day 13 of our AI in 15 Days email course provided by [Armand Ruiz](#) that I put all together. Today, we cover how to plan AI in business by creating an effective AI Business Roadmap. As leaders in your respective fields, the strategic integration of AI can be a transformative step for your organization, driving efficiency, innovation, and substantial growth.

AI Business Roadmap: A Strategic Necessity

An AI Business Roadmap isn't just a technical layout; it's a comprehensive plan that aligns AI technologies with your specific business goals. This blueprint encompasses timelines, resources, and technical requirements, and addresses potential risks and human factors ensuring a smooth AI integration.

AI Business Roadmap



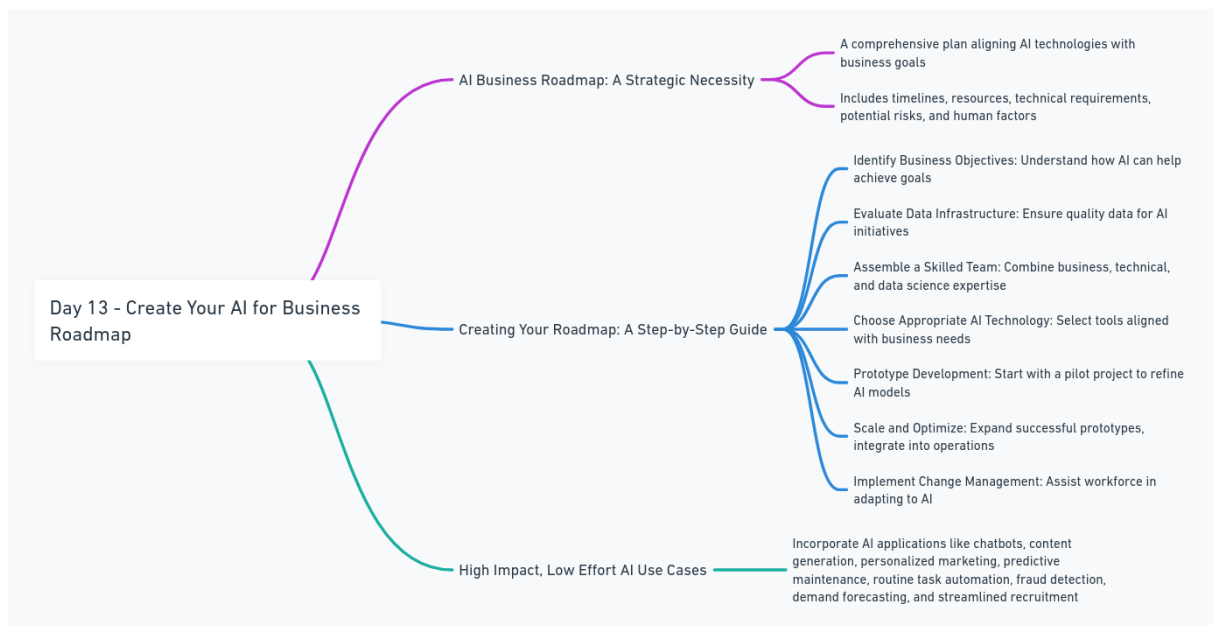
Creating Your Roadmap: A Step-by-Step Guide

- **Identify Business Objectives:** Understand how AI can help achieve your goals, whether it's through automation, predictive analytics, AI chatbots, or innovative product development.
- **Evaluate Data Infrastructure:** AI needs quality data. Assess your data collection, storage, and cleanliness to ensure your AI initiatives can thrive.
- **Assemble a Skilled Team:** Combine business insight, technical skills, and data science. Include business strategists, AI specialists, and IT professionals, or seek external expertise as necessary.
- **Choose Appropriate AI Technology:** Select AI tools like ML, NLP, RPA, or Computer Vision, aligned with your business needs.
- **Prototype Development:** Start small with a pilot project to address specific challenges, refining AI models based on performance.
- **Scale and Optimize:** Expand successful prototypes, integrating them into broader business operations and continuously optimizing.

- **Implement Change Management:** Develop strategies to assist your workforce in adapting to AI, including training and understanding AI benefits.

High Impact, Low Effort AI Use Cases

Incorporate high-impact, low-effort AI applications such as AI chatbots for customer service, content generation, personalized marketing, predictive maintenance, routine task automation, fraud detection, demand forecasting, and streamlined recruitment. These use cases provide a strong foundation for your AI journey, ensuring meaningful returns with minimal initial effort.

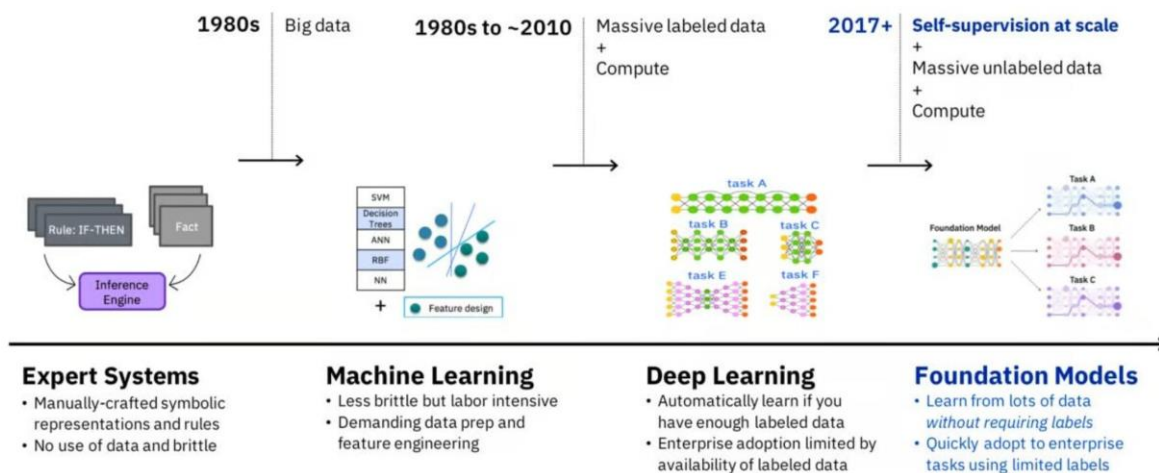


Day 14 - Future Trends in AI

Welcome to Day 14 of our 15-day journey through the world of Generative AI. Today, we're venturing into the future to explore the exciting innovations we can expect in the next decade. Let's dive in!

Where Do We Come From?

Understanding AI's journey is key to predicting its future. The past decade laid the foundation for advancements in machine learning and neural networks. Now, we're poised to build on this legacy, driving AI towards more sophisticated and nuanced applications.



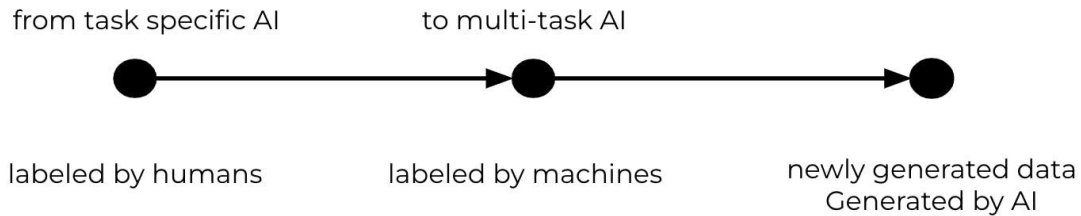
In **traditional machine learning**, individual siloeed models require task-specific training and a significant amount of human-supervised learning. The limit on performance & capabilities for supervised learning are humans.

In contrast, **foundation models** are massive multi-tasking systems, adaptable with little or no additional training, utilizing pre-trained, self-supervised learning techniques. The limit on performance & capabilities is mostly on computing and data access (not labeling).

Synthetic Data

If the limit to a better model is more data, why don't create it artificially? The rise of synthetic data is a game-changer. It's about creating artificial datasets that can train AI without compromising privacy or relying on scarce real-world data. This innovation is set to revolutionize fields from healthcare to autonomous driving, making AI training more accessible, ethical, and comprehensive.

how do we get more data?



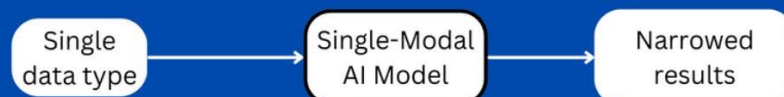
We will see increased usage of synthetic data because it overcomes data scarcity by allowing limitless generation, ensures balanced data distribution to avoid biases, and is cost-effective, bypassing the expensive and time-consuming process of real-world data collection and labeling.

Multimodality

Multimodality is the future of AI's interaction with the world. By integrating text, image, sound, and more, AI can understand and respond to complex queries with unprecedented accuracy. This holistic approach will deepen AI's integration into daily life, from smarter virtual assistants to more intuitive educational tools.

train ai with data generated by ai

Single-modal AI Model

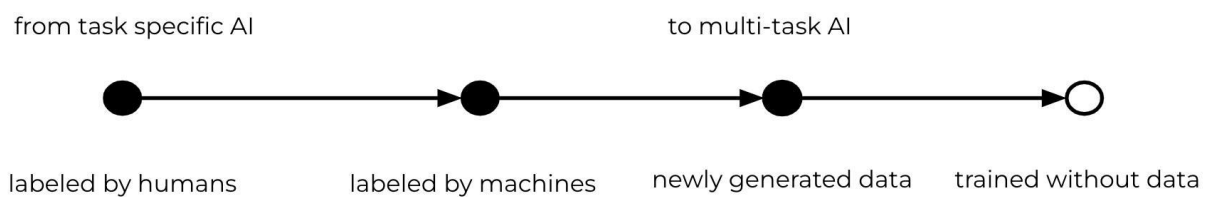


Multimodal AI Model?



Reinforcement Learning

Want to take it to the next level? What if you could train the AI without data? Meet Reinforcement Learning, a technique poised to make significant strides. By learning through trial and error, AI systems will become more autonomous and capable of solving complex, real-world problems. This means smarter algorithms in everything from financial forecasting to climate change modeling.



Check out [this video](#) to see a demonstration of Reinforcement Learning in action.

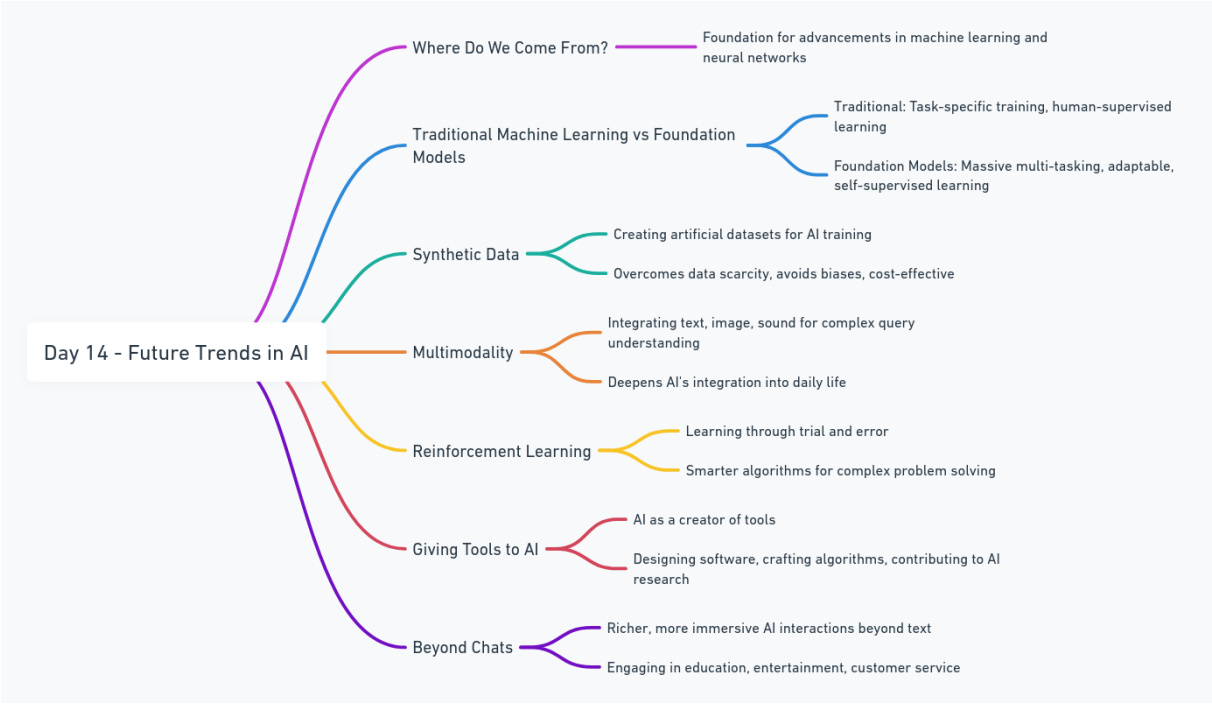
Giving Tools to AI

In the next decade, AI will evolve beyond being just a tool to becoming a creator of tools. We will see AI designing software, crafting algorithms, and contributing to AI research. AI agents will autonomously manage projects in a continuous loop, executing tasks, enhancing results, and generating new tasks based on objectives and past outcomes. Their workflow will include task execution, result enrichment, task creation, and prioritization. Equipped with integration capabilities, these agents will be able to search for information in CRMs, access databases, send emails, and more. Frameworks like BabyAGI and Auto-GPT are already emerging to test these concepts.

Beyond Chats

While chatbots and conversational AI have made leaps, the future extends far beyond text. Expect AI that can seamlessly interact across various formats, offering richer, more immersive experiences. Whether it's in education, entertainment, or customer service, AI will engage us in more meaningful, dynamic ways.

As we approach the final day of our course, remember that the future of AI is not just about technology; it's about the creativity and ingenuity of those who wield it.



Day 15 - Continuing Your AI Journey

We've reached the final day of our 15-day exploration into Generative AI. It's not an end, but a beginning to a lifelong journey in AI learning and exploration. Here's how you can continue growing in this exciting field:

Online Courses & Tutorials

Armand Ruiz created the AI Bootcamp that can provide you next level of practice. It includes videos and exercises with multiple tools, no coding knowledge is required. Get your free access [here](#)

If you would like the next level of depth, Armand Ruiz recommends the Short Courses from DeepLearning.ai, they include topics such as:

[Functions, Tools and Agents with LangChain](#)

[Vector Databases: from Embeddings to Applications](#)

[Quality and Safety for LLM Applications](#)



[Building and Evaluating Advanced RAG Applications](#)

[Reinforcement Learning from Human Feedback](#)

These require basic Python knowledge.

Follow leaders in the AI space

These are some of the AI leaders Armand Ruiz follows on LinkedIn:

- [Aishwarya Srinivasan](#): Data Scientist | LinkedIn Top Voice Data & AI | EB1A Recipient | 460k+ Followers | Ex- Google, Ex-IBM
- [Allie K. Miller](#): AI Entrepreneur, Advisor, and Investor | 1MM+ followers | Former Amazon, IBM | LinkedIn top Voice for AI 2019-2023
- [Luis Serrano](#): AI scientist | YouTuber - 120K followers | Author of Grokking Machine Learning
- [Andriy Burkov](#): ML at TalentNeuron, author of  The Hundred-Page Machine Learning Book and  the Machine Learning Engineering book

[Armand Ruiz](#), author of this 15 day learning generative AI, also publish content every day with tips and best practices to apply AI for Business. Feel free to follow him, like I do, [here](#).

Join Newsletters

Don't fall behind on AI. These newsletters summarize the daily news so you get the latest AI trends and tools you need to know. I read them all every day.

- [The Rundown](#)
- [Ben's Bites](#)
- [The Neuron](#)

Remember, the field of AI is ever-evolving, and continuous learning is key. Embrace curiosity, keep experimenting, and stay connected to the AI community. You're now equipped to take the next steps in your AI journey.

In compiling this comprehensive guide, I have integrated the 15-day email course on generative AI created by Armand Ruiz, encapsulating the essence of each day's learning complemented by AI-generated diagrams for enhanced synthesis. This document is intended as a foundational resource for those keen on exploring potential great power of generative AI. While my specialization lies in healthcare, the insights garnered from Armand Ruiz's teachings hold broad applicability across various sectors, including life sciences. I highly recommend following Armand Ruiz for a deeper understanding of generative AI and its potential.

This work is a tribute to the knowledge shared by Armand Ruiz, tailored to foster learning and innovation within and beyond the healthcare domain.

Emmanuel Lacharme

PharmD - Health economist - Global HTA & patient access expert