

# A Hands-on Introduction to Generative AI

Sathish Sundaramoorthy



ITS   
SEMINOLE  
SHOWCASE

# Agenda

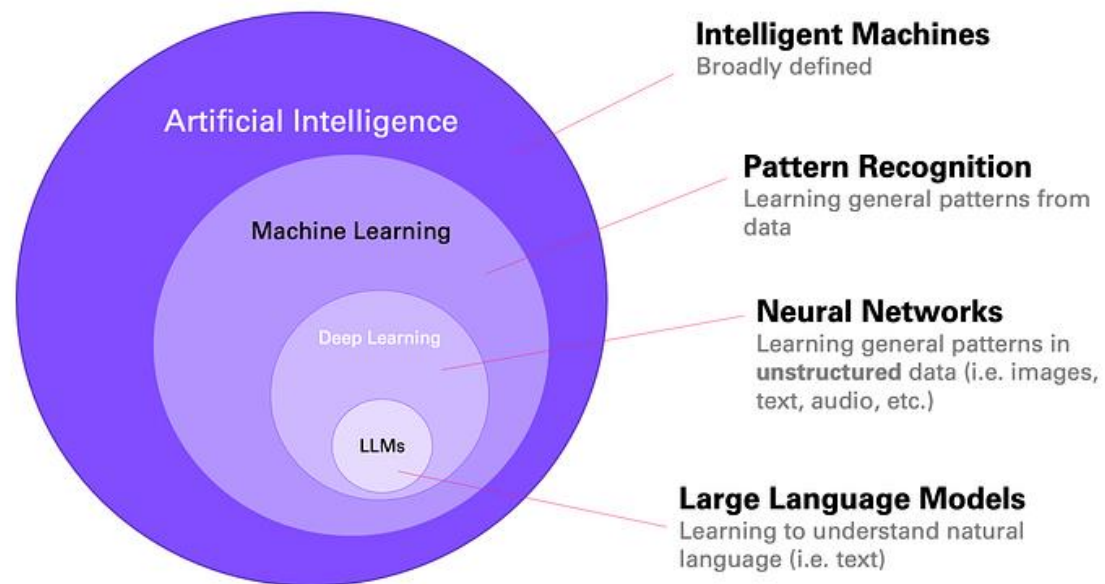
- Introduction to Generative AI
- History & Foundations of Large Language Models(LLMs)
- Understanding LLMs
- Techniques for Leveraging LLMs
- Challenges and considerations
- Use cases and Real-world Applications
- Demo
- Future of Generative AI

# About Me



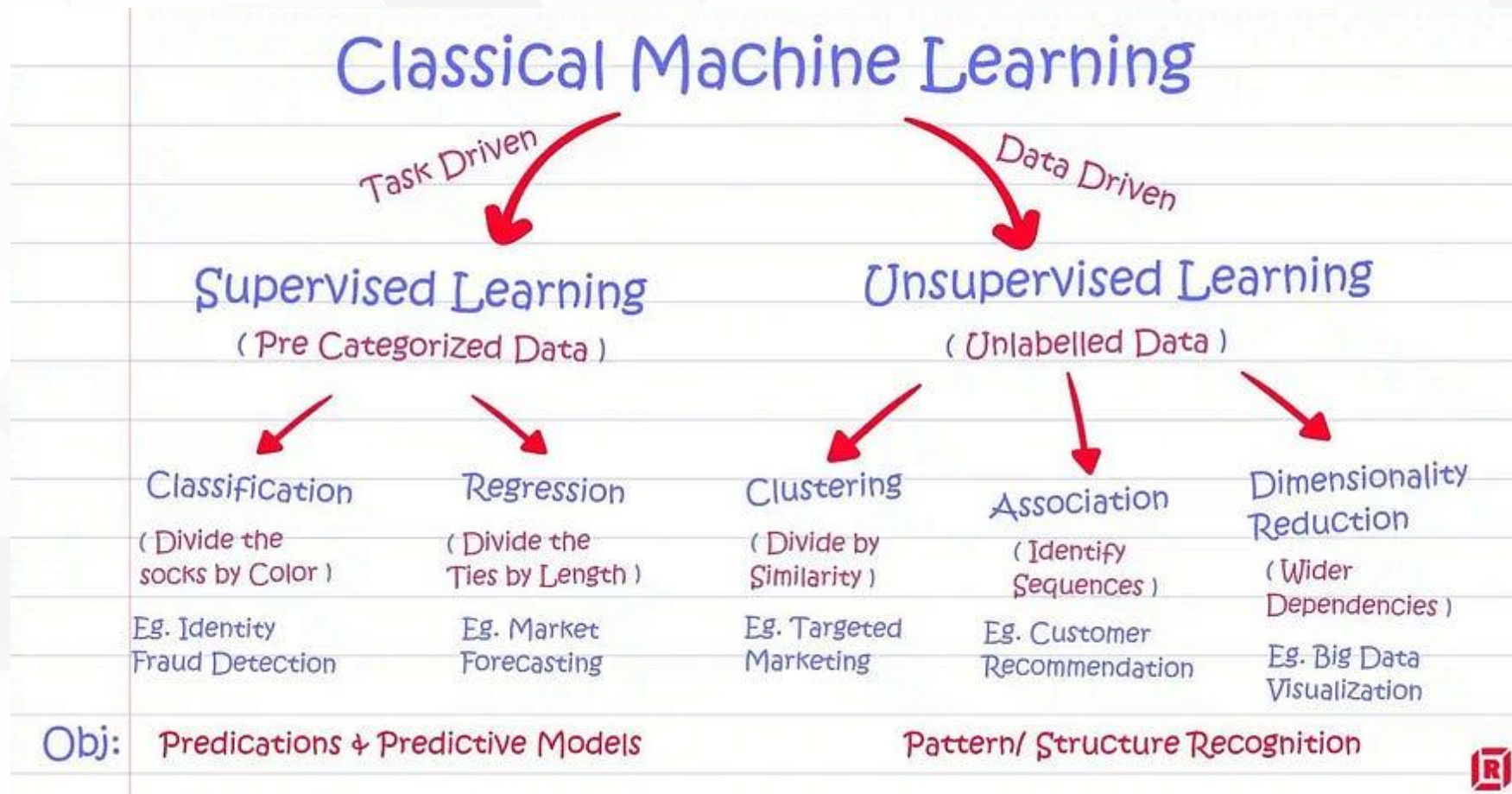
- 12+ years of experience in ERP domain.
- Masters of Science in Machine Learning and Artificial Intelligence
- Working with FSU for 2 years as ERP Analyst
- HR / Financials / Campus solutions
- Passionate about latest technology trends in AI.
- Presented in Alliance 2024 and Reconnect 2023 on PO Roll Automation, Robotic Process Automation and Kibana Analytics.

# Introduction



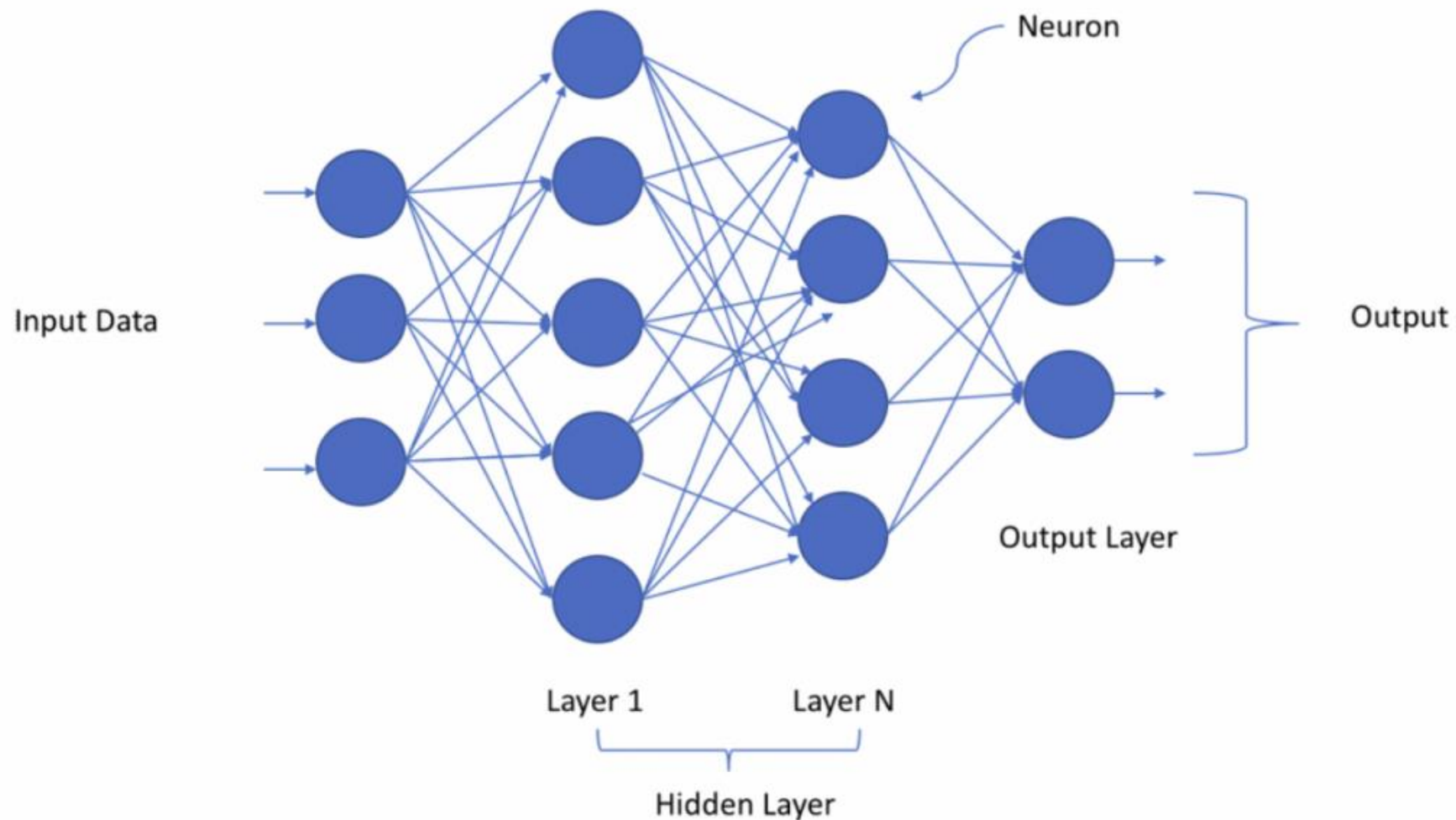
+  
•

# Machine Learning



# Deep Learning

- CNN
- RNN



# Generative AI

- Generative AI models learn to generate new content (text, images, audio, etc.) based on training data.
- Examples: GPT-3 for text generation, DALL-E for image generation, Jukebox for music generation.
- Given a **prompt**, the model predicts the expected response, creating new original data like images, text, audio, video.
- Creativity is powered by large datasets.

# Traditional vs Generative AI

discriminative  
technique



Classify

**Discriminative model**  
(classify as dog or cat)



generative  
technique



Generate

**Generative model**  
(generate a cat)

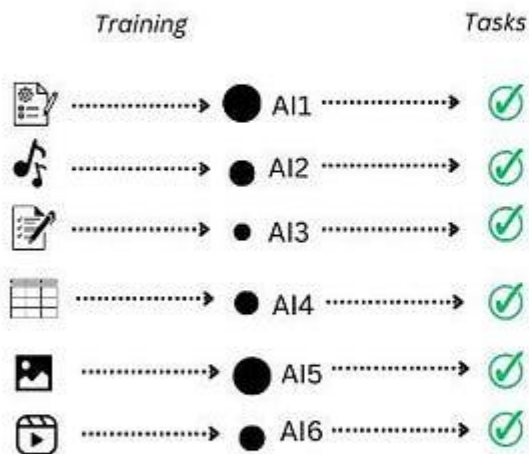




# Traditional vs Generative AI

## traditional ML vs generative AI

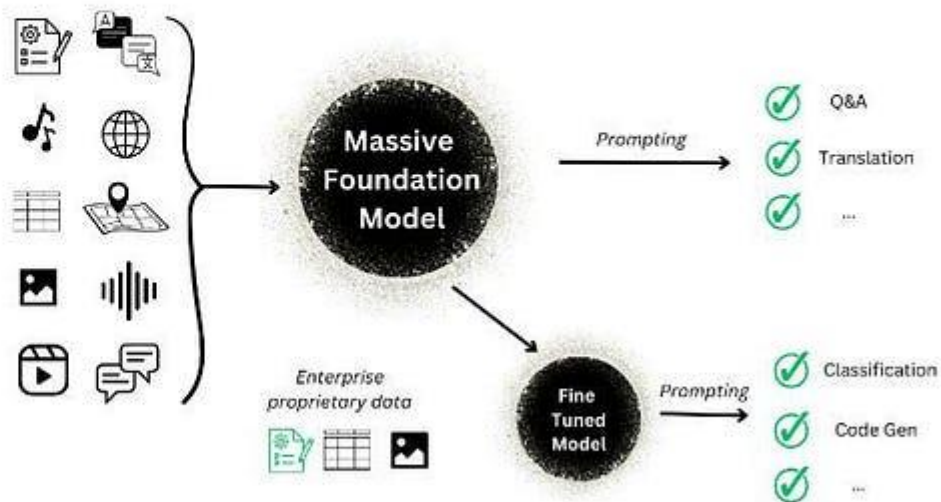
### predictive ML



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

### genAI

Massive external data



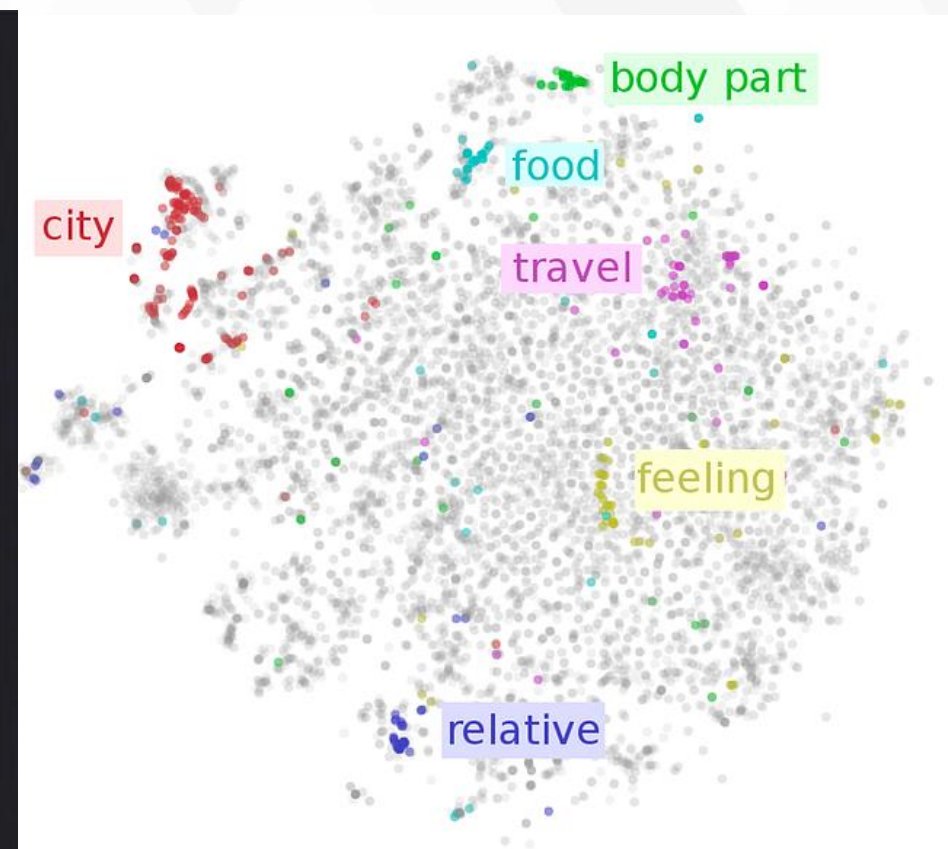
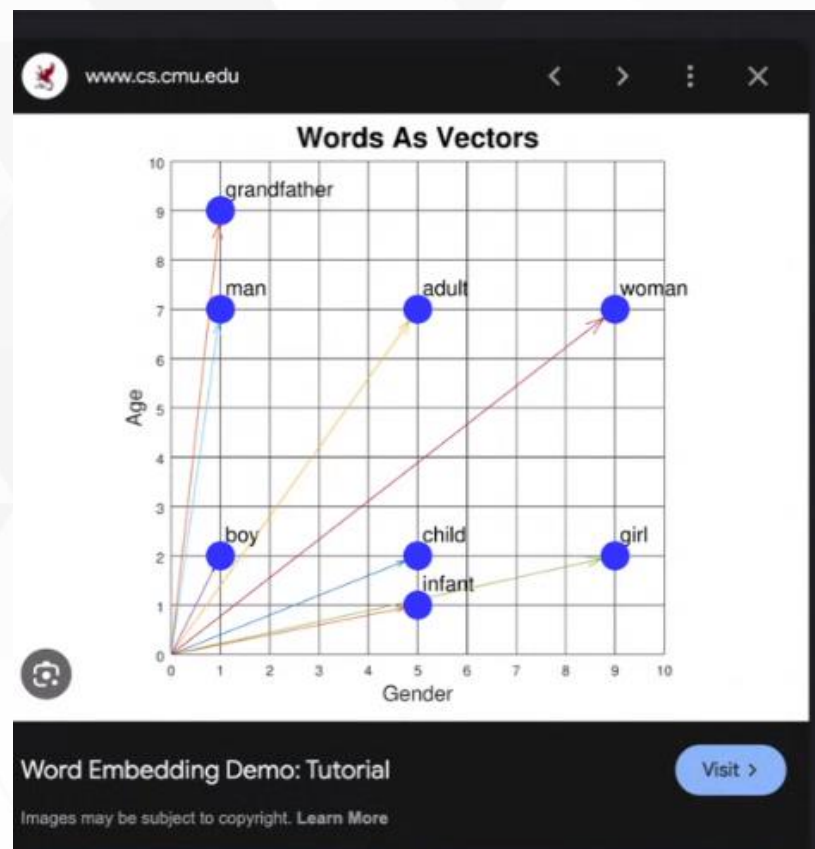
- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

# Word embeddings



- Word2Vec
- CBOW model
- Skip Gram model

A visualization showing the word embeddings for 'banana' and 'monkey'. The word 'banana' is associated with the vector [0.85 0.12 ... 0.23] and the word 'monkey' is associated with the vector [0.96 0.55 ... 0.32]. A blue arrow points from the 'banana' vector to the 'monkey' vector.



# Language modeling

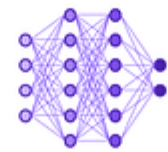
Imagine the following task: **Predict the next word in a sequence**

[ The cat likes to sleep in the \_\_\_ ] → What **word** comes next?

**Can we frame this as a ML problem?** Yes, it's a **classification** task.

[ The cat likes to sleep in the ]

Input



Neural Network  
(LLM)



Word	Probability
ability	0.002
bag	0.071
<b>box</b>	<b>0.085</b>
...	...
zebra	0.001

Output

Now we have (say)  
~50,000 classes (i.e.  
words)

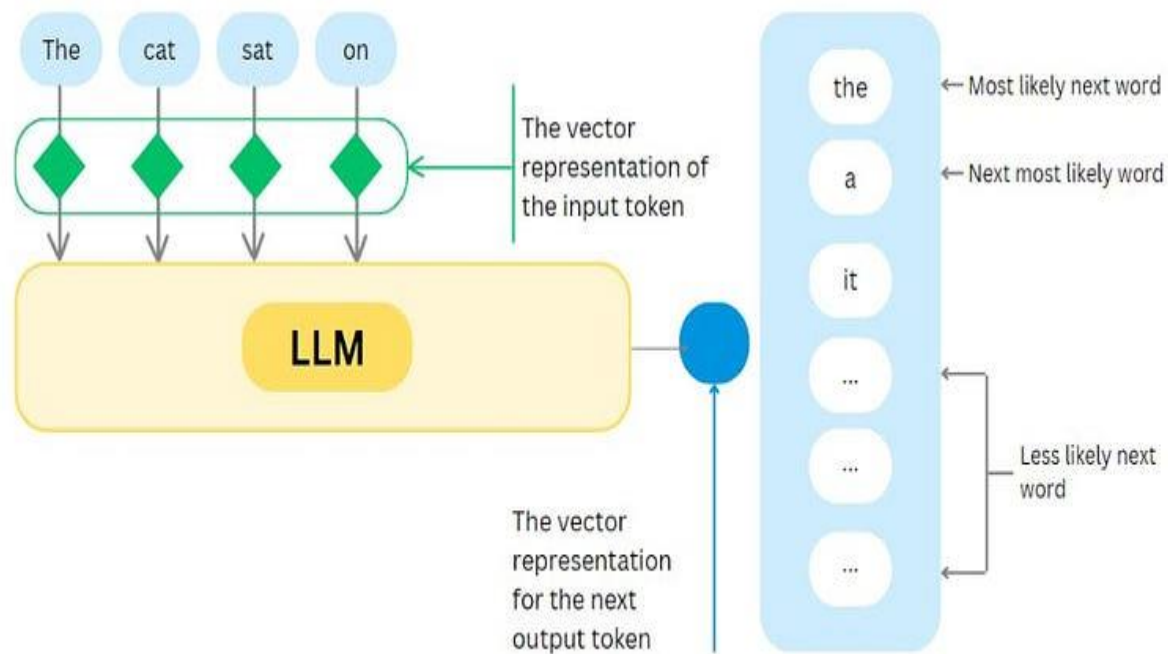


# Large Language Models

- Weights
- Parameters
- Tokenization

## 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.



# LLMs

QUESTION ANSWERING

ARITHMETIC



LANGUAGE UNDERSTANDING

**8 billion parameters**

# Transformers

## 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 performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

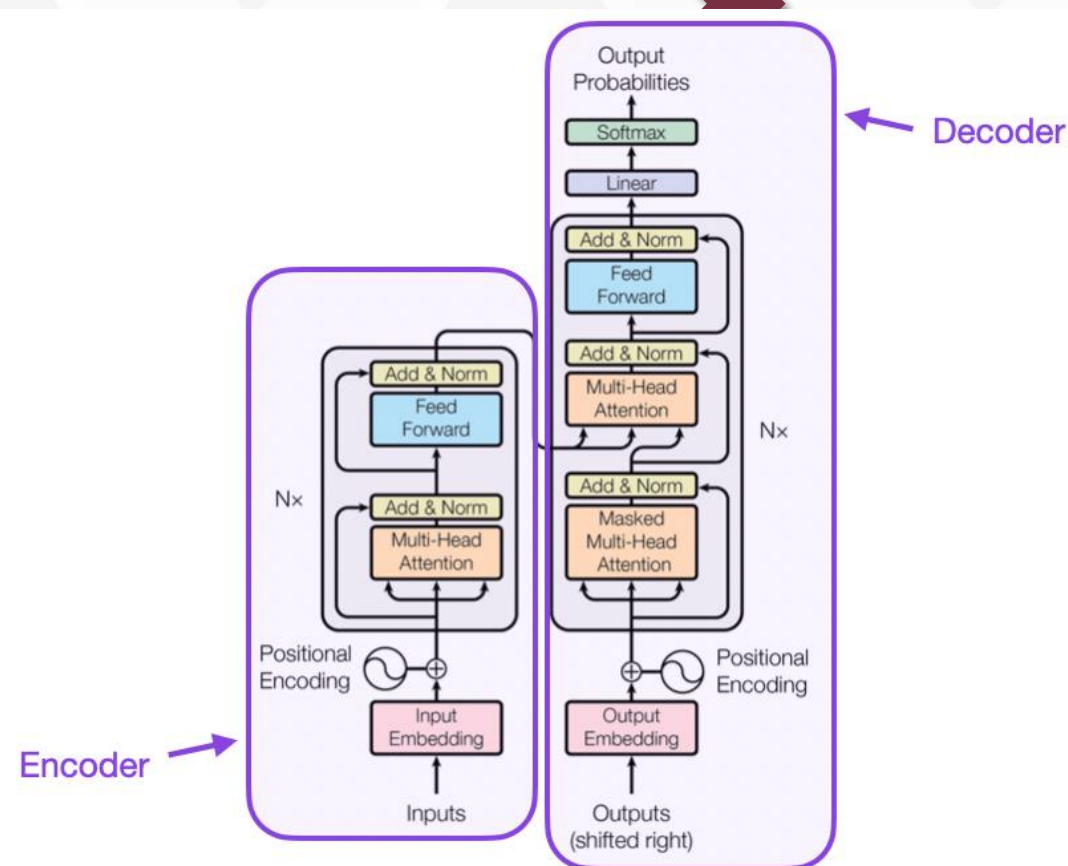


Figure 1: The Transformer - model architecture.

# GPT



---

## What does **Generative Pre-trained Transformer (GPT)** mean

+ .

### **Generative**

Means "next word prediction."

As just described.

### **Pre-trained**

The LLM is pretrained on massive amounts of text from the internet and other sources.

### **Transformer**

The neural network architecture used (introduced in 2017).

# Phases of training LLMs (GPT-3 & 4)

## 1. Pretraining

Massive amounts of data from the internet + books + etc.

**Question:** What is the problem with that?

**Answer:** We get a model that can babble on about anything, but it's probably not **aligned** with what we want it to do.

## 2. Instruction Fine-tuning

Teaching the model to respond to instructions.

Model learns to respond to instructions.

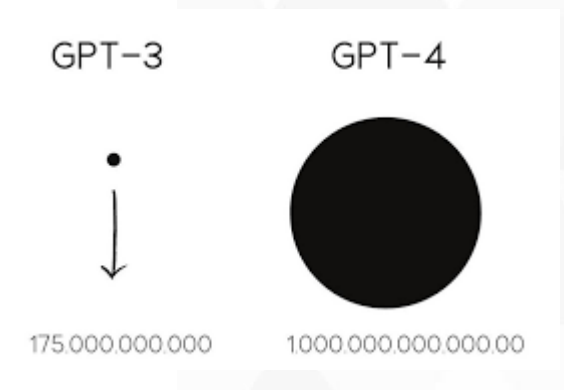
→ Helps **alignment**

*"Alignment" is a hugely important research topic*

## 3. Reinforcement Learning from Human Feedback

Similar purpose to instruction tuning.

Helps produce output that is closer to what humans want or like.

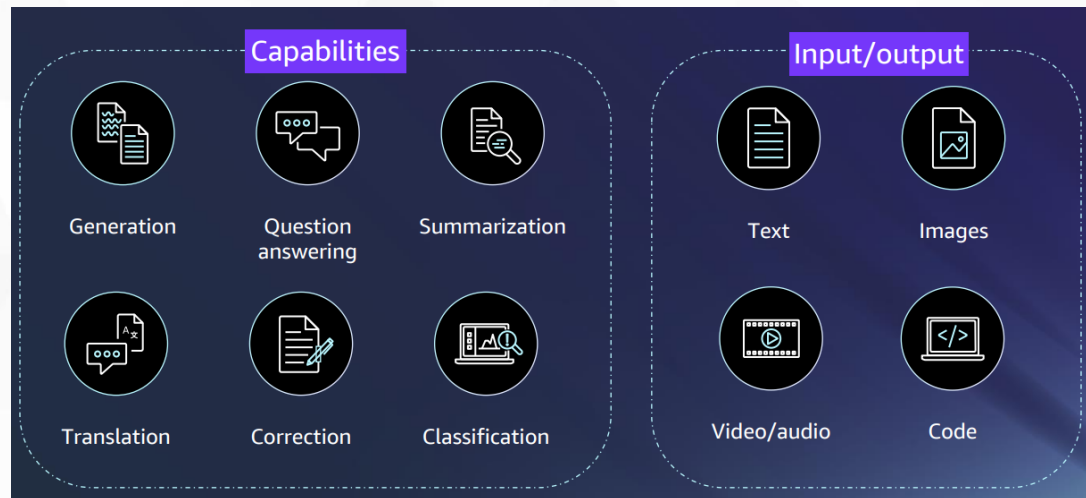






Powerful Programming Language?

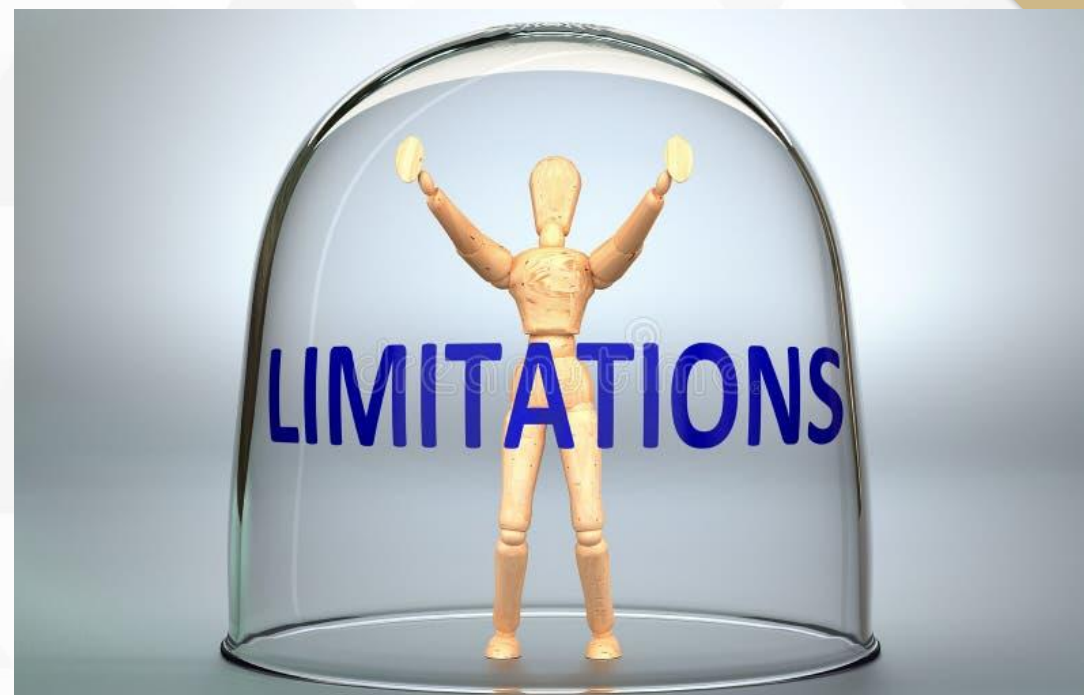
# What can LLMs do?



- Content creation (text, images, music, etc.)
- Creative assistance (ideation, brainstorming, etc.)
- Task automation (summarization, code generation, etc.)
- Synthetic data generation for machine learning
- Keyword/skill extraction
- Information Extraction from Image
- Object Recognition
- Sentiment Analysis
- Named Entity Recognition
- Translation
- Natural Language to SQL
- .....

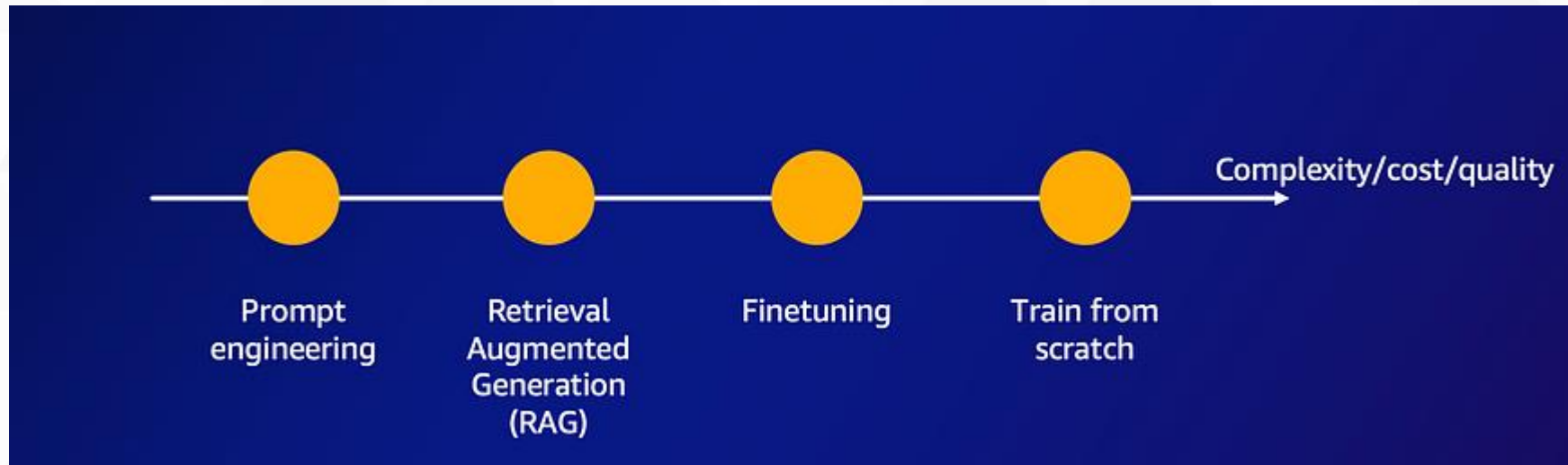
# Limitations and Challenges

- Difficulty in ensuring the generated content is truthful, **accurate**, and up-to-date.
- Tendency for models to "**hallucinate**" or contradict themselves.
- Generative models can perpetuate societal **biases** present in their training data.
- Safety and Misuse concerns
- Computational and Environmental costs



# Approach to use LLMs

- Build from scratch
  - Fine tuning
  - RAG
  - API based (Open AI API)
- Training vs inference
  - Open source vs closed source



	Prompt Engineering	RAG	Fine-tuning
Key benefit	Rapid adaptability and prototyping	Incorporation of real-time or external data for factual answers	High specialization and tailored responses.
Training requirement	No	No	Yes
External Data	No	Needs a corpus	Task-specific dataset
Computation	No overhead	Overhead for retrieval	Intensive for training, no overhead for inference
Quality Improvement	Iterative refinement	Update/expand corpus	Periodic retraining
Potential Costs	Human labor for crafting prompts	Training, storing corpus, computational overhead	Dataset, training compute, evaluation
Technical Complexity	Low technical	Moderate to high – management of corpus can be complex	Moderate – requires expertise in neural networks & dataset biases
Extra Inference Latency	No	Yes – needed for retrieval	No

*Comparison of prompt engineering, RAG, and fine-tuning*

# RAG



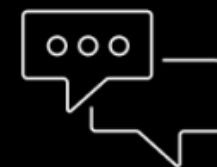
## Retrieval

Fetches the relevant content from the external knowledge base or data sources based on a user query



## Augmentation

Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model

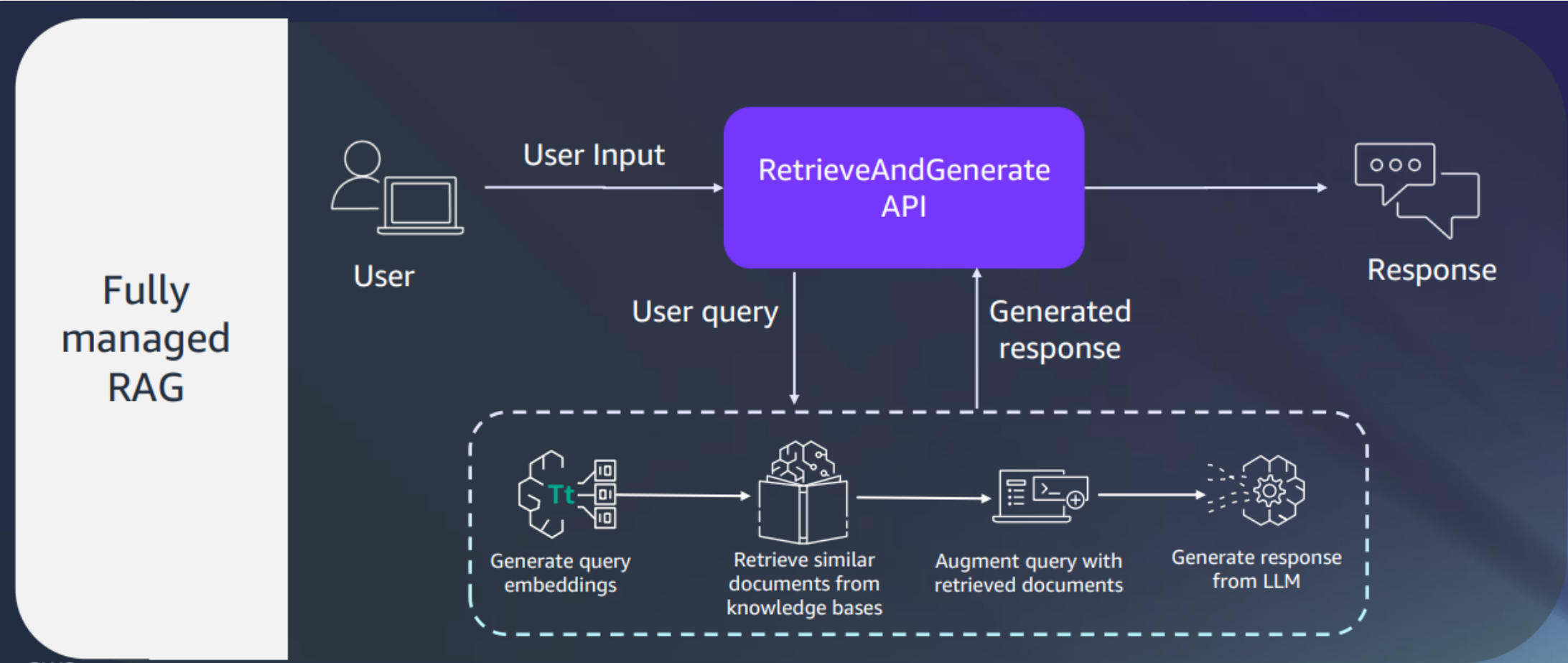


## Generation

Response from the foundation model based on the augmented prompt



# RAG



Demo





# Use cases in HR/FI



Challenge	Description	How AI Can Help
<b>Resume Analysis</b>	Efficiently reviewing and evaluating a large volume of candidate resumes.	- AI-powered resume screening tools to identify the best-fit candidates.
<b>Skill Repository</b>	Maintaining an up-to-date database of employee skills and competencies.	- AI-based skills taxonomies and knowledge extraction to update the repository.
<b>Candidate Evaluation</b>	Conducting objective and consistent assessments of candidates' qualifications and fit.	- AI-powered candidate assessment tools to provide data-driven insights.
<b>Generating Job Descriptions</b>	Creating accurate and compelling job postings to attract the right talent.	- AI-powered job description generation based on industry trends and organizational needs.
<b>Invoice Processing(Finance)</b>	Efficiently processing high volumes of invoices and ensuring timely payment.	- AI-based optical character recognition and workflow automation to streamline invoice processing.
<b>Fraudulent Transactions(Finance)</b>	Detecting and preventing fraudulent financial activities.	- AI-powered fraud detection models to identify anomalies and potential fraud.


# Use cases in HR/FI



Use Case	Challenges	Description
<b>Employee Engagement and Retention</b>	Identifying turnover risks and implementing proactive retention strategies	AI-driven sentiment analysis and predictive models to address engagement and retention
<b>Learning and Development</b>	Providing personalized training content and delivery	AI-powered platforms that adapt learning based on individual needs
<b>Performance Management</b>	Conducting effective performance reviews and goal-setting	AI-enabled real-time feedback, coaching, and career development insights
<b>HR Automation</b>	Streamlining repetitive HR tasks	AI-based automation of onboarding, payroll, and employee data management
<b>Intelligent Automation (Finance)</b>	Automating financial processes	AI-driven RPA for accounts payable, receivables, and reconciliations
<b>Compliance and Risk Management (Finance)</b>	Monitoring transactions and ensuring regulatory compliance	AI algorithms to identify risks and maintain compliance





# Use cases in CS



Challenge in Higher Education	Description	How AI Can Help
<b>Scalability and Accessibility</b>	Providing quality education to a growing and diverse student population while maintaining personalized attention and support.	- Generative AI-powered chatbots, virtual tutors, and adaptive learning platforms
<b>Student Engagement and Retention</b>	Keeping students engaged, motivated, and actively participating in their learning.	- AI-driven personalized feedback, gamification, and adaptive content
<b>Instructor Workload and Burnout</b>	Overwhelming workloads for instructors, leading to burnout and negatively impacting the quality of education.	- AI can automate grading, provide teaching assistance, and streamline administrative tasks
<b>Curriculum Development and Optimization</b>	Developing and continuously improving curriculum to stay relevant and meet the evolving needs of students and employers.	- AI can analyze data to help optimize curriculum and learning outcomes
<b>Data-Driven Decision Making</b>	Leveraging the vast amount of data generated in higher education to inform strategic decisions and drive institutional improvements.	- AI-powered analytics and predictive modeling to support data-driven decision-making

# Use cases in CS



Challenge in Higher Education	Description	How AI Can Help
<b>Personalized Career Guidance</b>	Providing tailored career counseling and job placement support to a diverse student population.	AI-powered career guidance systems for personalized recommendations
<b>Admissions Process</b>	Streamlining the application and selection process for a growing pool of candidates.	Generative AI to streamline admissions, from essays to candidate evaluation
<b>Thesis Writing Assistance</b>	Providing students with support and guidance throughout the thesis writing process.	Generative AI to assist with ideation, structure, feedback, and drafting
<b>Idea Generation and Brainstorming</b>	Helping students generate and explore new ideas for research projects, essays, or creative works.	Generative AI to provide prompts and enable innovative thinking
<b>Mental Health and Well-being</b>	Addressing the growing mental health and wellness needs of students, faculty, and staff	AI-powered chatbots, virtual counseling services, and predictive analytics can improve access to mental health resources and early intervention.

# Modern AI Stack: The Emerging Building Blocks for GenAI

<p>Layer 4: <b>Observability</b></p>	<p>OBSERVABILITY, EVALUATION, SECURITY</p> <p>  Helicone            AgentOps            Humanloop            Credal.ai            CALYPSOAI            truera            eppo            BRAINTRUST            Patronus AI            LOGIO         </p>	
<p>Layer 3: <b>Deployment</b></p>	<p>PROMPT MANAGEMENT</p> <p>  vellum            LangSmith         </p>	<p>ORCHESTRATION</p> <p>  Marian            orkes            Radiant         </p>
	<p>AGENT TOOL FRAMEWORKS</p> <p>  LangChain            AutoGPT            FIXIE            LlamaIndex         </p>	
<p>Layer 2: <b>Data</b></p>	<p>DATA PRE-PROCESSING</p> <p>  gable            datologyai            Cleanlab         </p>	<p>ETL + DATA PIPELINES</p> <p>  UNSTRUCTURED            NOMIC            Lexy            Indexify         </p>
	<p>DATABASES (VECTOR, DB, METADATA STORE, CONTEXT CACHE)</p> <p>  databricks            upstash            Pinecone            NEON            WarpStream            momento         </p>	
<p>Layer 1: <b>Compute + Foundation</b></p>	<p>MODEL DEPLOYMENT + INFERENCE</p> <p>  baseten            Modal            Replicate            clarifai            Substrate            fireworks.ai         </p>	
	<p>FINETUNING + RLHF</p> <p>  LAMINI            Predibase            arcee.ai         </p>	
	<p>FOUNDATION MODELS</p> <p>  OpenAI            ANTHROPIC            MISTRAL AI            contextual.ai            Hugging Face            Llama 2         </p>	
<p>TRAINING</p> <p>  Modular            Lightning AI            OctoML         </p>		
<p>GPU PROVIDERS</p> <p>  aws            Azure            Google Cloud            CoreWeave            Lambda            FOUNDRY            together.ai         </p>		

# Future

- Ongoing research and advancements
- Potential for multimodal integration (text, image, audio)
- Ethical considerations and the need for responsible development
- Artificial General Intelligence(AGI)



# Final Thoughts

- Getting Started with Generative AI
- The **Pilot** Phase: A Strategic Approach
  - Identify **Use Cases**
  - Assess Feasibility
  - Choose the right approach
  - Evaluate Performance
  - Ensure responsible development
- Building the foundation
  - Assemble a **Cross-Functional Team**
  - Secure Necessary Resources.
  - Develop a Governance Framework

Questions?





# Please Provide Feedback!

