[L]LMs and Applications

Where the Magic Meets Reality

Vasudeva Varma

What Are LLMs?

AI systems trained on vast datasets to predict, generate, and understand human-like language.

Capabilities

Answer questions Summarize text Translate languages
Create content Handle complex linguistic tasks

Why LLMs are Important?

Powerful Performance: Handle tasks like reasoning, summarization, and creative generation with remarkable accuracy.

Emergent Behaviors: Perform tasks (summarizing novels or generating code) without explicit training

Adaptability: Easily fine-tuned for specific needs with minimal effort.

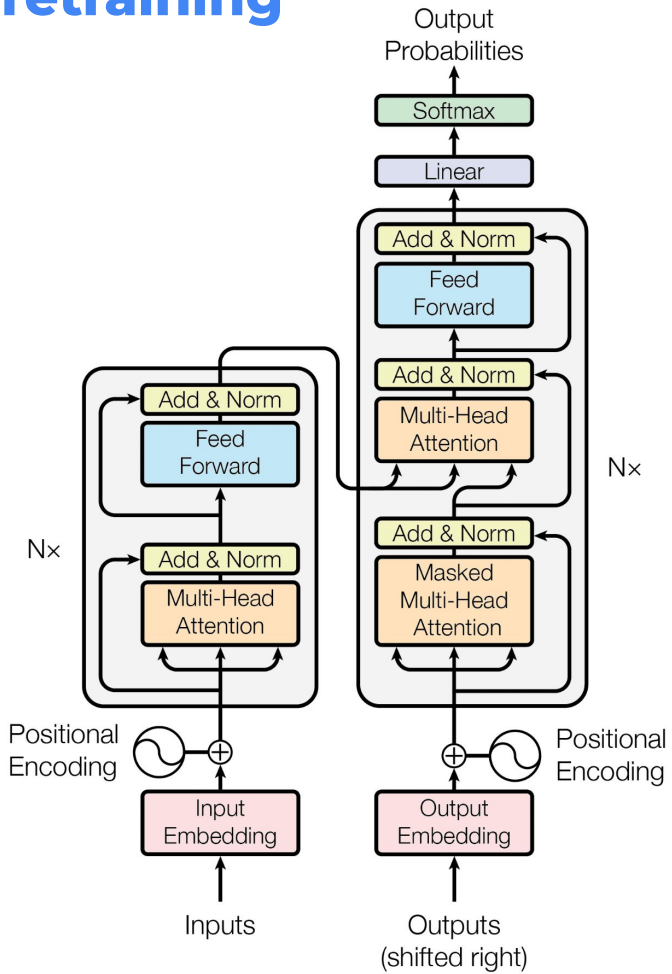
Everyday Impact: Enhances tools like chatbots, translators, and content creators for widespread use.

Fundamentals:

Pretraining
Fine Tuning
Prompt Engineering
Optimizing
Augmenting

and... agents

Pretraining



Transformer Architecture

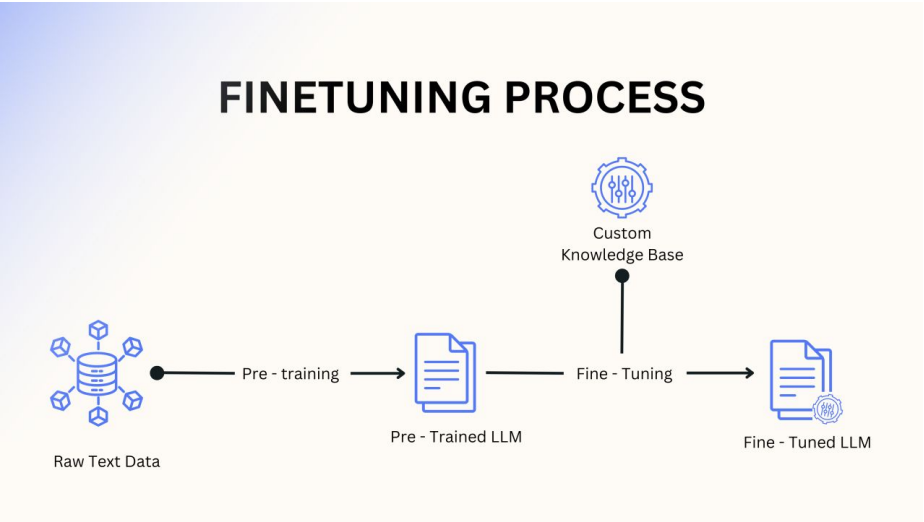
Processes sequences of text in parallel, understanding word relationships effectively.

Key Components:

- **Self (Multi-headed) attention**: Determines the importance of each word in the input (parallelly on different parts)
- **Feedforward Networks**: Adds depth and enhances meaning.
- **Positional Encoding**: Retains the order of words for better comprehension.

Result: **Pretrained model** enabling fast, scalable processing of large language tasks.

Fine Tuning



Please see:

[TechForward Research Seminar Series - JANUARY EDITION](#) **(imp: 44:05 to 53:20 of Manohar's talk)**

Fine-Tuning Techniques

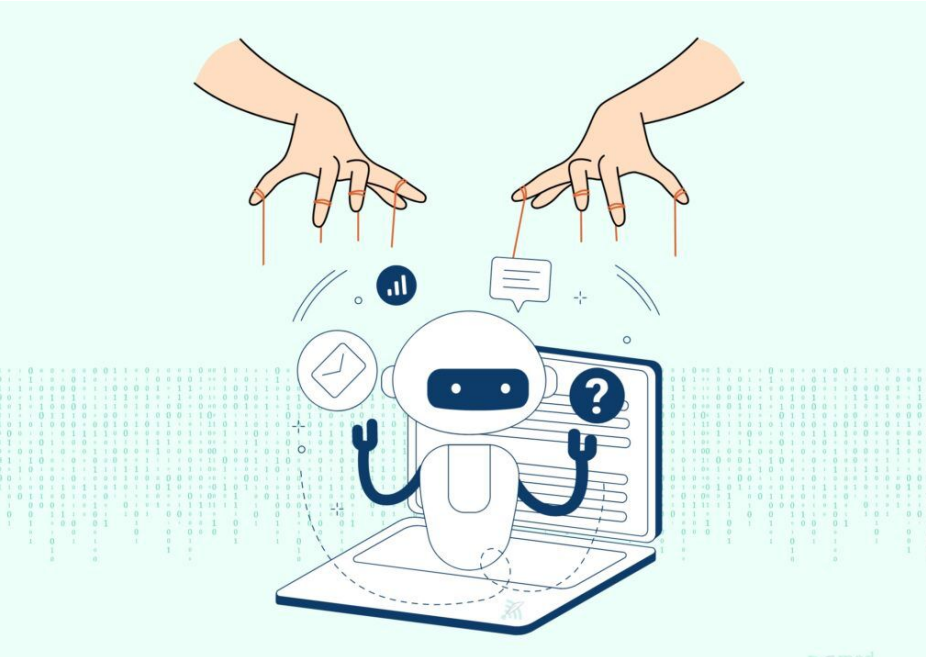
Create task-optimized, efficient LLMs to adapt for specific tasks like customer service or scientific analysis.

Key Methods:

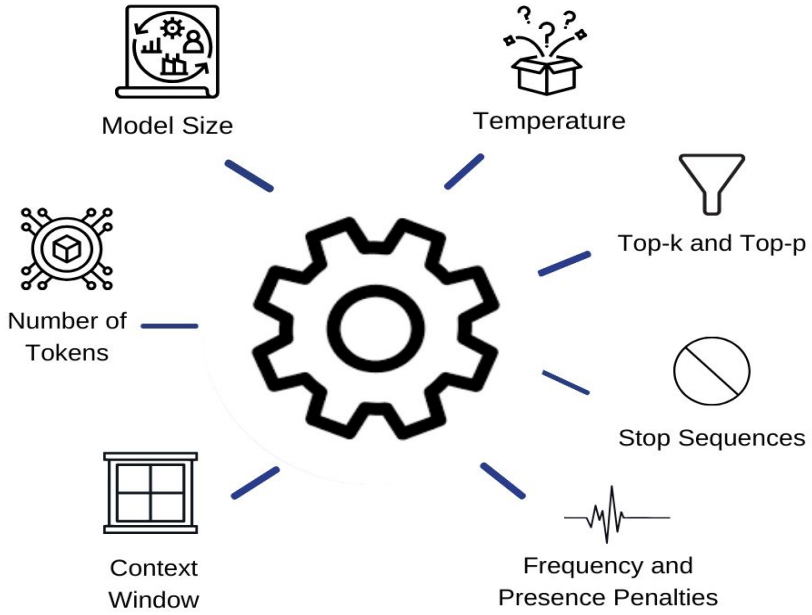
- **Supervised Fine-Tuning (SFT): Uses labeled data to train for task-specific outputs.**
- **Reinforcement Learning from Human Feedback (RLHF): Refines outputs to align with user preferences.**
- **Parameter Efficient Fine-Tuning (PEFT): Uses techniques like adapters for cost-effective customization.**

Prompt Engineering

Prompt Engineering:



Sampling Techniques & Parameters:

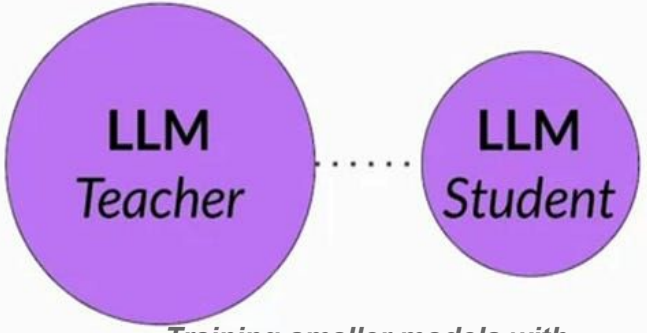


Optimizing

LLMs require significant computation and memory

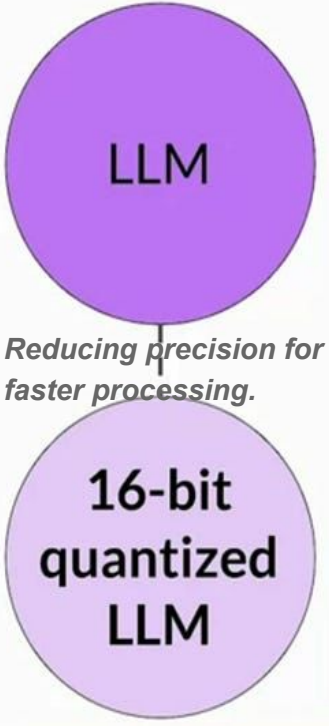
How to Accelerate Inference *to achieve Optimize speed, cost, and efficiency*

Distillation



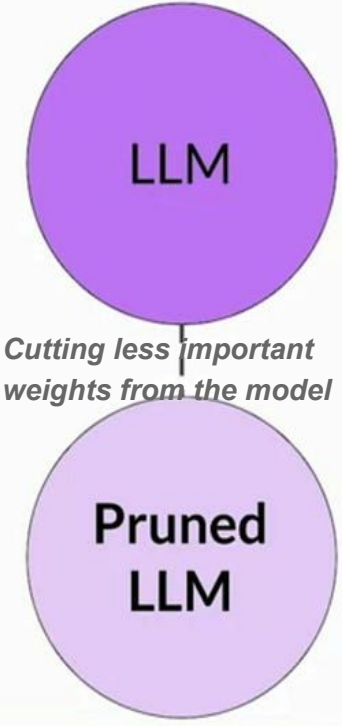
Training smaller models with similar capabilities.

Quantization



Reducing precision for faster processing.

Pruning



Cutting less important weights from the model

Augmenting

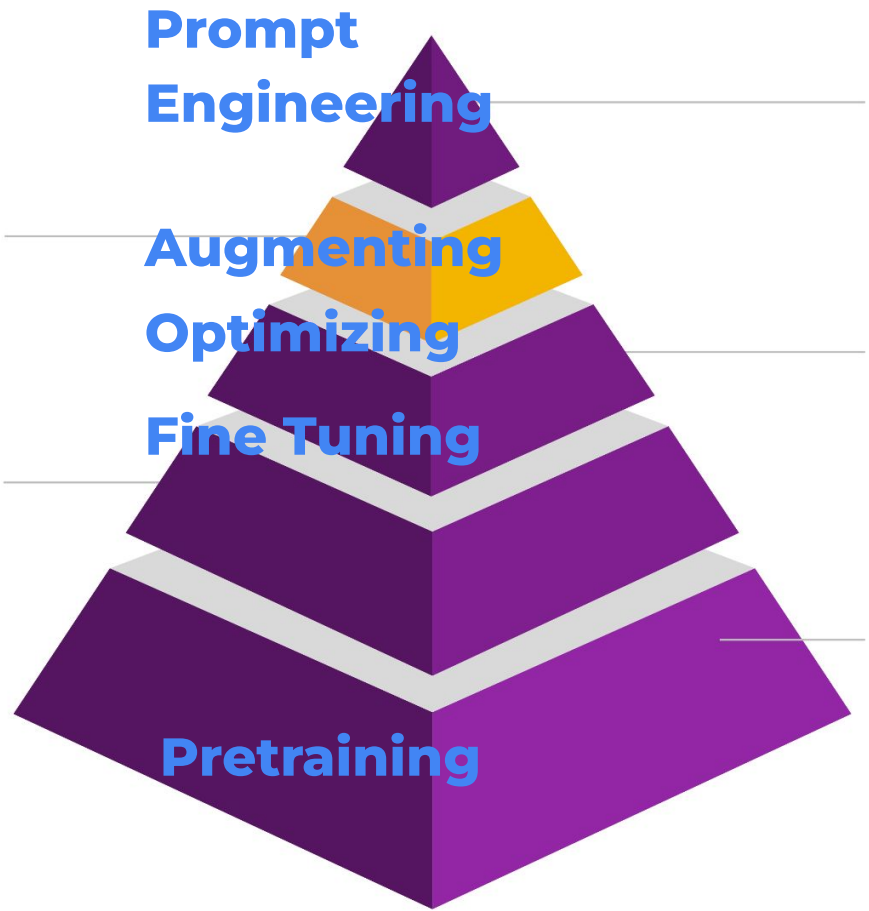
Retrieval-Augmented Generation (RAG) *combines retrieval systems with LLMs*

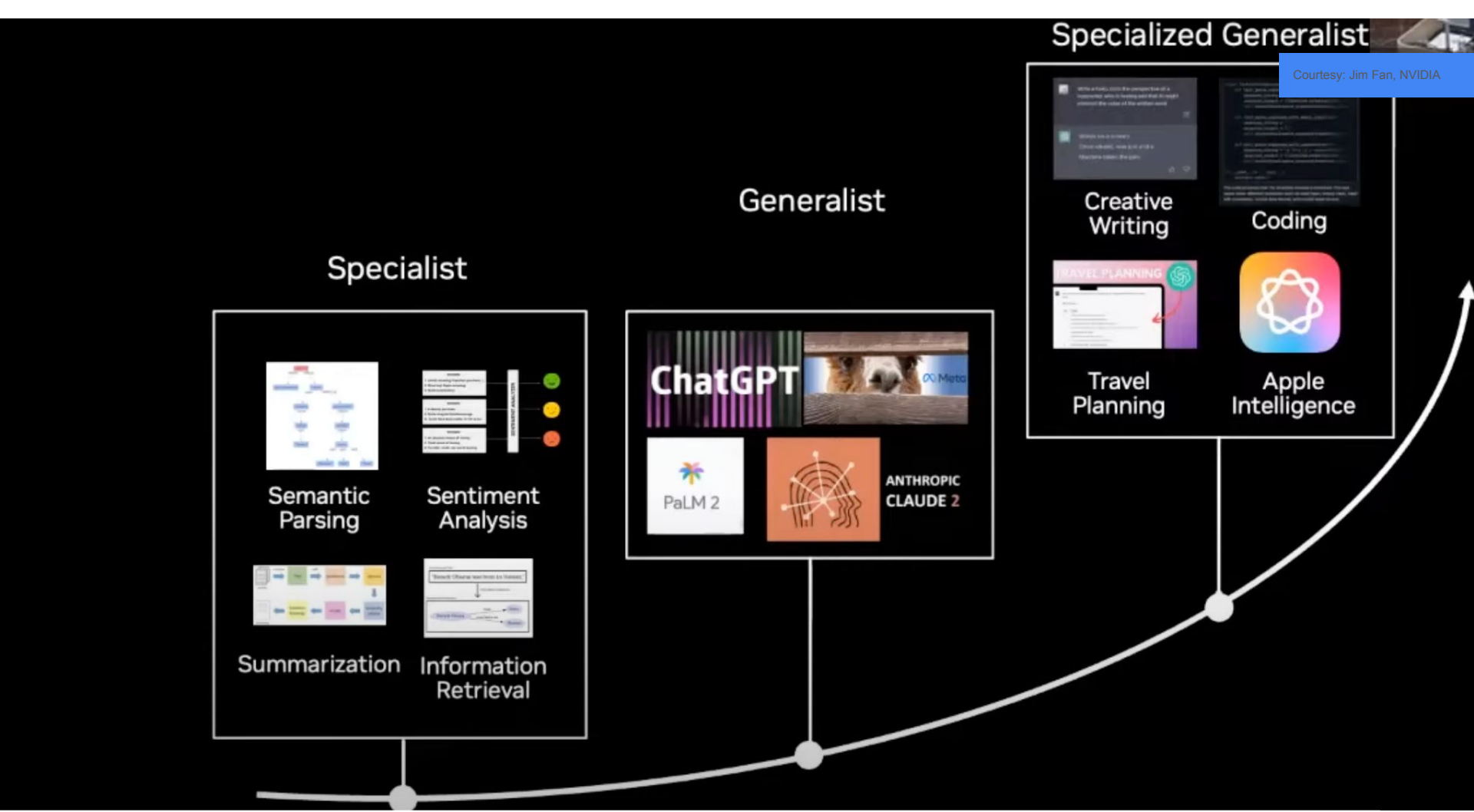
How It Works

1. **Query Understanding**: The LLM ***interprets*** the user's question.
2. **Information Retrieval**: The query is matched against a ***Document Repository***
3. **Contextual Generation**: The retrieved data is fed into the LLM, which ***generates a grounded, factual response.***

Key Advantages

- **Enhanced Accuracy**: Incorporates real-world, dynamic data, generating **up-to-date, accurate information**
- **Domain-Specific Adaptability**: Tailored to specialized datasets (e.g., healthcare, legal).
- **Reduced Hallucinations**: Limits reliance on outdated or inferred knowledge.

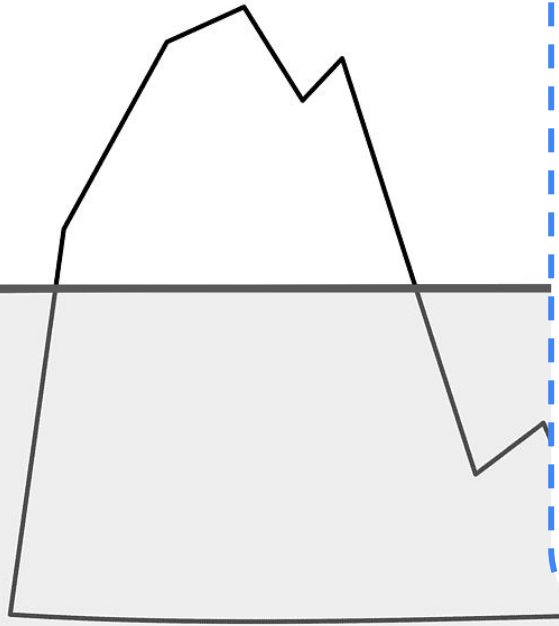

Prompt Engineering
Augmenting
Optimizing
Fine Tuning
Pretraining
Prompt Engineering
“In-context learning”
Retrieval Augmented Generation (RAG)
Parameter Efficient Fine Tuning (PEFT)
Instruction Tuning / Supervised Fine-Tuning
Pretraining


Specialized Generalist
Courtesy: Jim Fan, NVIDIA
Generalist
Specialist
Creative Writing
Coding
Travel Planning
Apple Intelligence
ChatGPT
PaLM 2
ANTHROPIC CLAUDE 2
Semantic Parsing
Sentiment Analysis
Summarization
Information Retrieval

I Don't Work on LLMs...

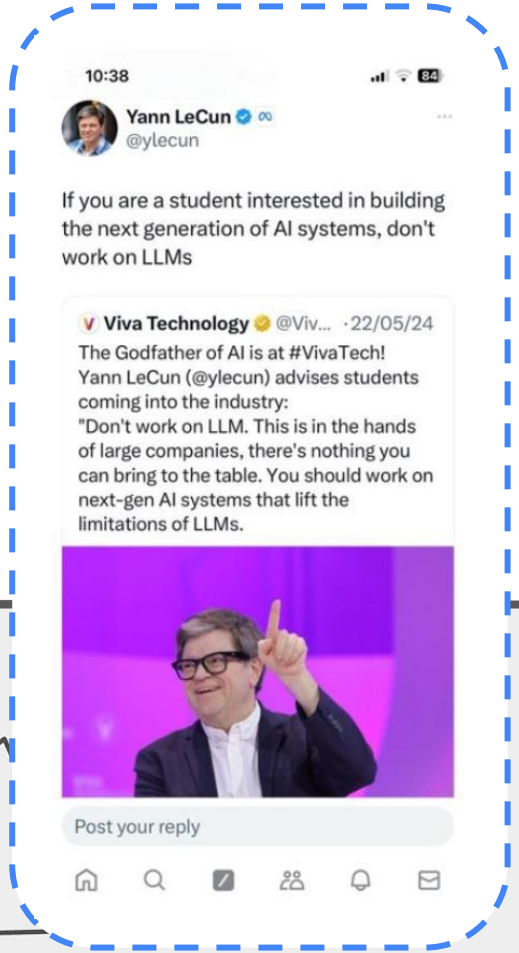
My focus is on addressing the limitations of LLMs

"Most of the progress in AI has been on the part of the iceberg that is above the water.



There's a lot of progress yet to be made on the part of the iceberg that is below the water."

-Yann LeCun



Who will work on local issues?

The biggest opportunity for India is to look at the sectors where the Indian economy is already strong and figure out the applications of AI to those sectors that enable it to maintain its advantages

Andrew Ng(Co-founder of Coursera and founder of DeepLearning.Ai)

3 Use cases/Applications

Healthcare

Science Research

Education/Learning

Role of Gen AI and LLM in HealthCare

Experiences with healthcare Apps

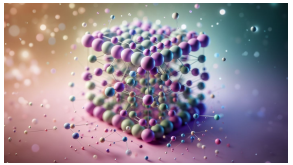
- ➔ Personalised Diagnosis
- ➔ Doctor Patient Communication
- ➔ Drug Discovery
- ➔ Insurance

The other side of the coin

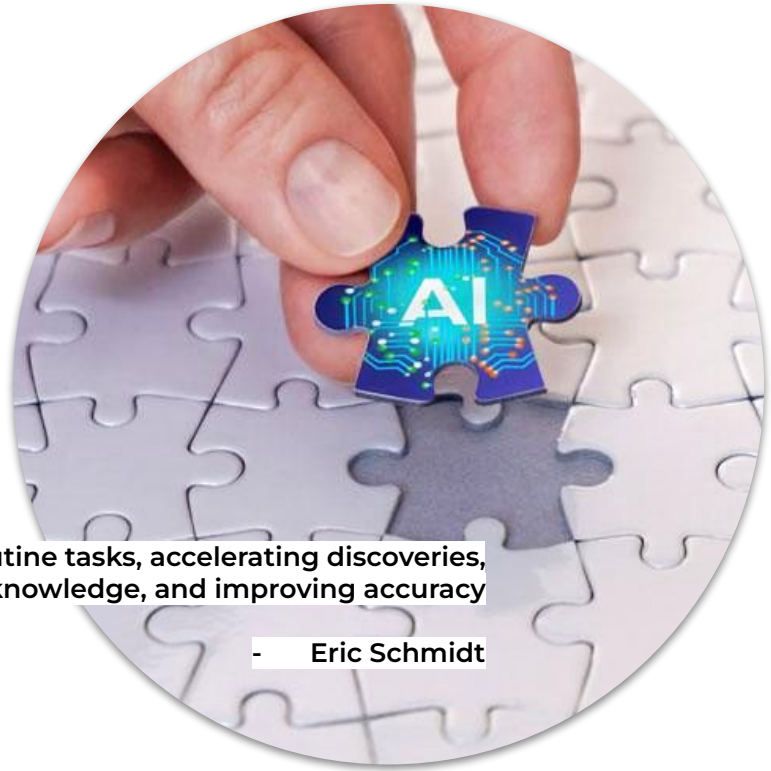
- Risk / accuracy
- Fairness/ Bias
- Accountability
- Transparency
- Ethics
- Safety

Role of Gen AI and LLM in Scientific Research

AlphaFold



Clara



How will AI revolutionize science and science research?

AI will transform the scientific process by automating routine tasks, accelerating discoveries, enhancing experimentation, integrating cross-domain knowledge, and improving accuracy

- Eric Schmidt

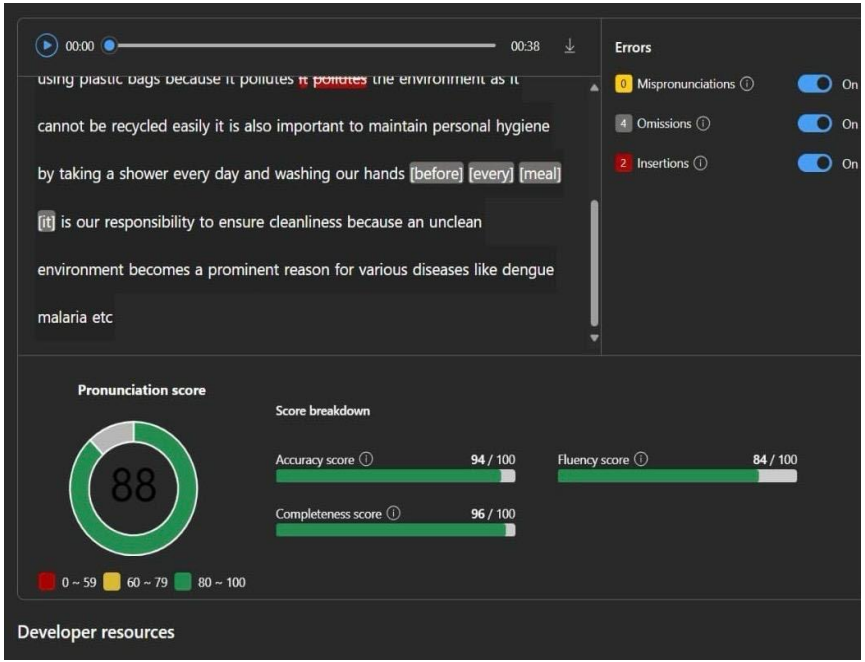
Use case

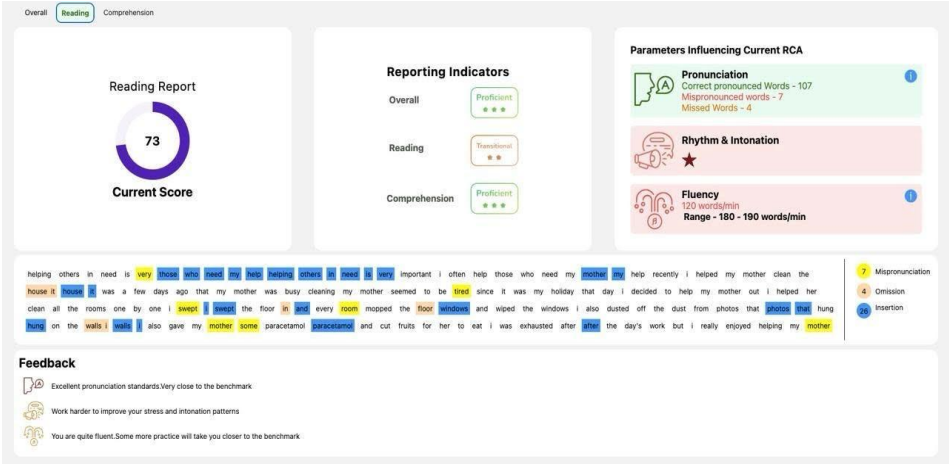
Education/Learning

Experiments in using GenAI in Learning

Behavioural and Social Learning Course
RGUKT & ISB

Building a Negotiation Course with GenAI
ISB

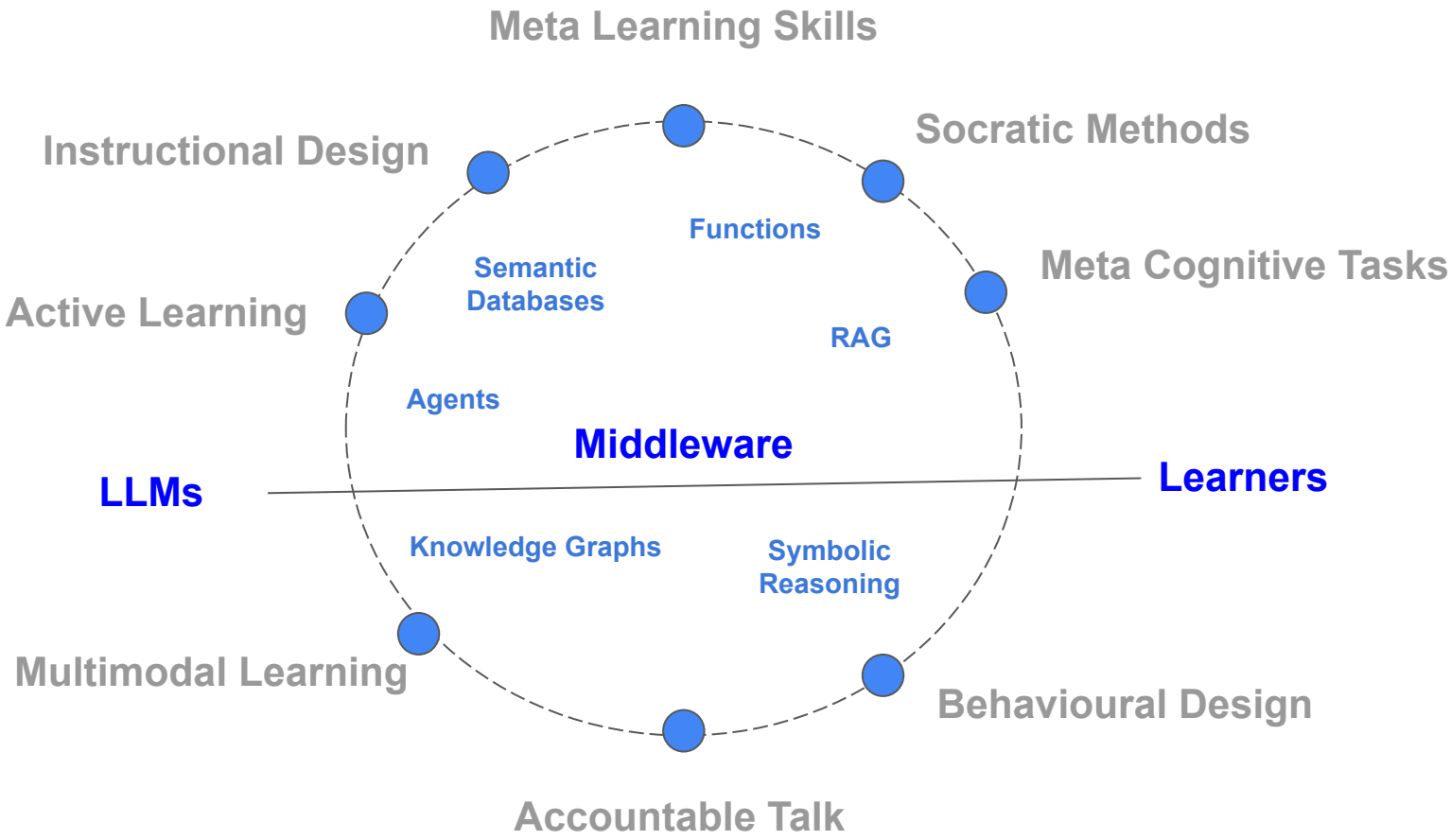




Issues with scalability and adaptability

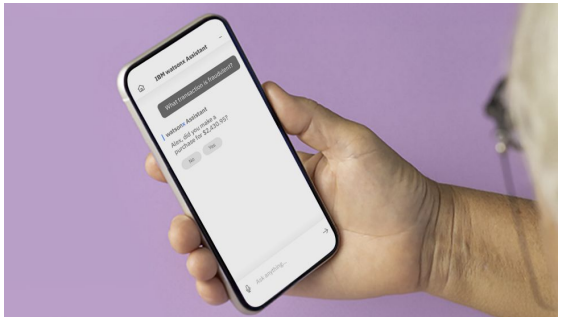
Some are natural fits and some tough ones.

- **STEM subjects?**


Meta Learning Skills
Instructional Design
Socratic Methods
Functions
Semantic Databases
Meta Cognitive Tasks
Active Learning
RAG
Agents
Middleware
LLMs
Learners
Knowledge Graphs
Symbolic Reasoning
Multimodal Learning
Behavioural Design
Accountable Talk

3 key advancements that is rapidly evolving Gen AI's capabilities (courtesy: Eric Schmidt)

Really Large Context Windows



Agents

Text-to-Action Capabilities

Key messages:

The last mile is the hardest; **if one doesn't do well, magic may fall flat on its face.**

While technology tremendously enhances productivity, ***domain knowledge is the cornerstone to success.***

Need to build middleware, responsibility layer, and domain specific foundation models

Thank you