Microsoft

GenAI: From One Token To An AI Revolution

Subhasish Ghosh

Senior Technical Program Manager, Azure OpenAI Service

Customer eXperience Engineering (CxE), Azure AI Platform, Microsoft

13 July 2024

AGENDA

1	GenAI Concepts, Transformers Azure OpenAI (AOAI) Offerings	*30 mins*
2	Top 3 Microsoft Build 2024 AI Announcements	*15 mins*
3	Industry Use-Cases	*15 mins*
4	Demo (Python SDK, C#)	*15 mins*
5	Quiz!!! 6 Qs Q&A 	*15 mins*
		90 mins

GenAI Concepts
Azure OpenAI Offerings

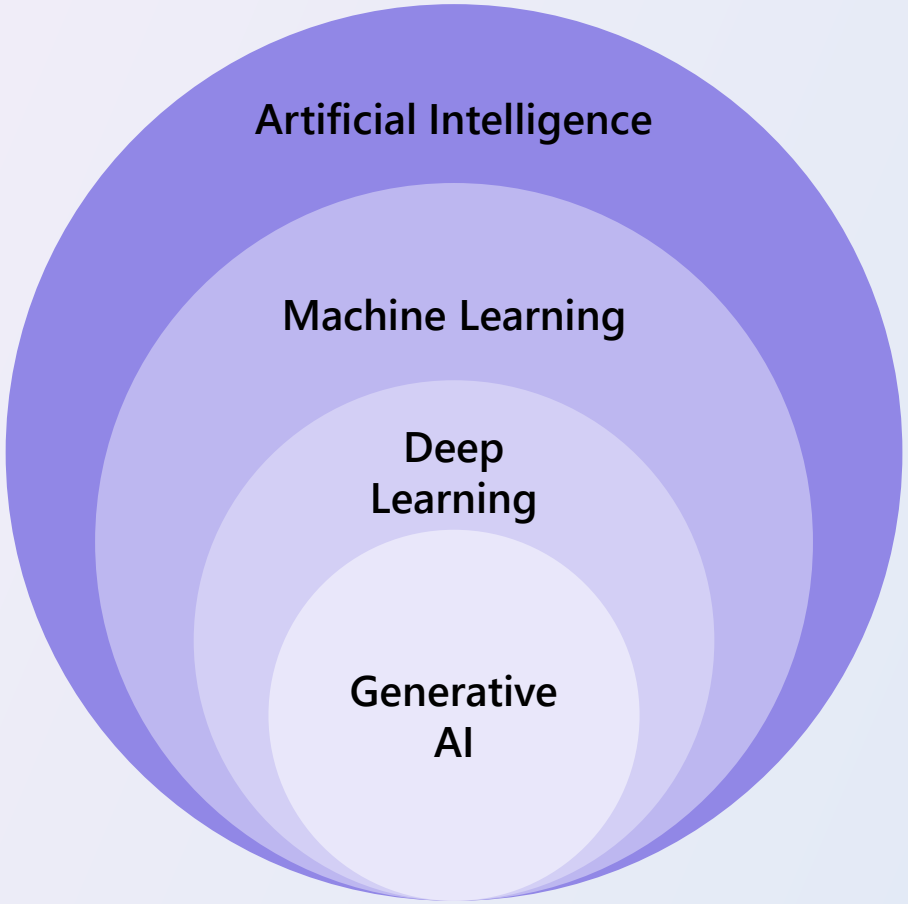
The Vision.

Can the computers some day understand us, instead of us understanding them?

Probably.”

Alan Mathison Turing (1912-1954)
Founding father of AI, Cryptography, Computer & Cognitive Sciences

The journey continues with generative AI



1956

Artificial Intelligence

The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997

Machine Learning

Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012

Deep Learning

A machine learning technique in which layers of neural networks are used to process data and make **decisions.**

2021

Generative AI

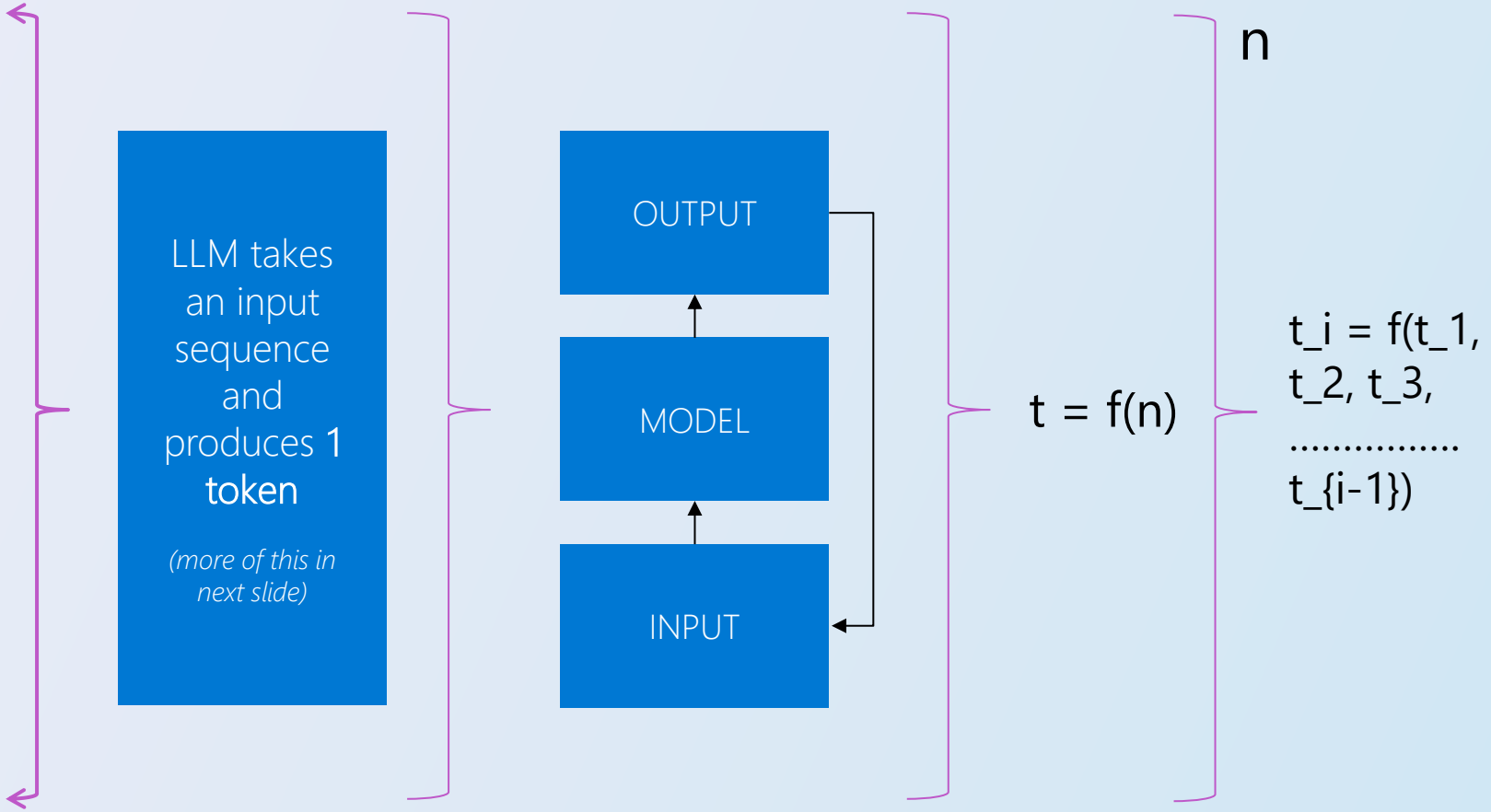
Create **new** written, visual, and auditory content given prompts or existing data.

From One Token to an AI Revolution

- A token is the smallest (further indivisible) unit of analysis. E.g., in OpenAI Large Language Models (LLMs), for English, 1 token is 4 characters (~ 0.75 word). Word *including* could be *includ* + *ing*, representing 2 tokens.

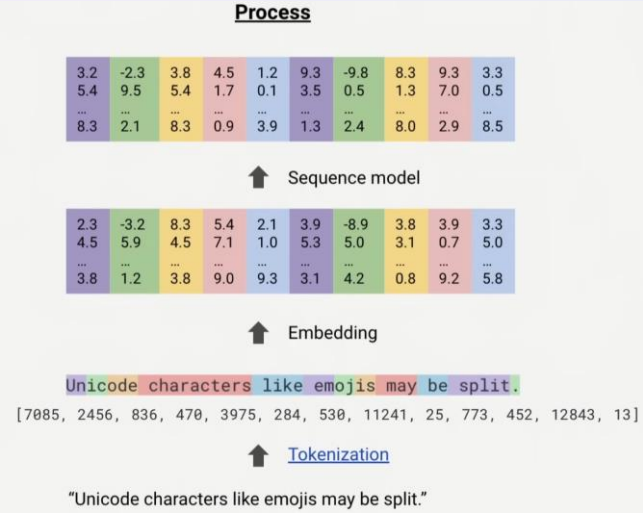
- An OpenAI Generative Pre-Trained Transformer (GPT) model to produce a token during inferencing goes through following 9 essential steps:

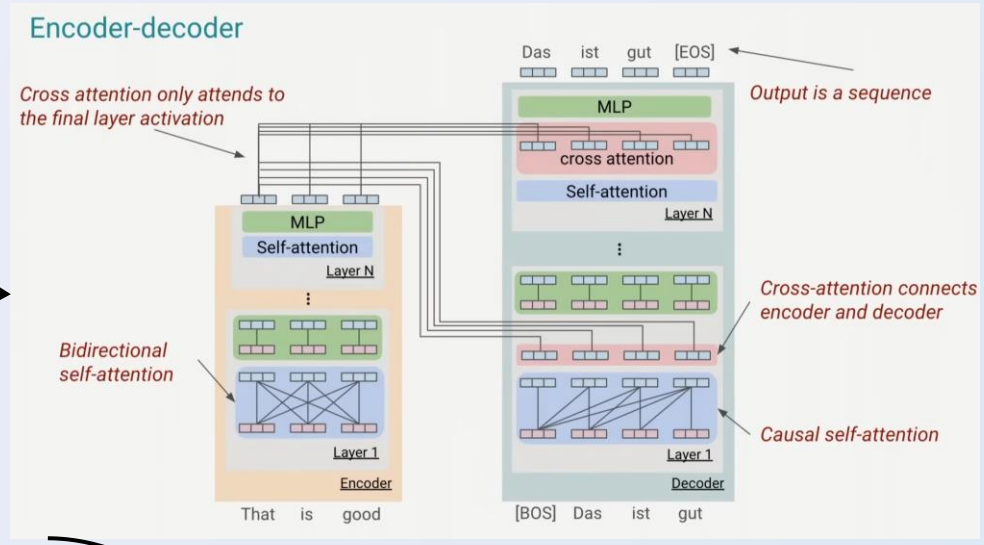
1. Tokenization
2. Model Input
3. Model
4. Output Generation
5. Sampler
6. Next Token Selection
7. Next Token Addition
8. Token Generation Completion
9. Text Reconstruction



Transformers – An Overview

Variants:
1. Encoder-Decoder
2. Encoder-only
3. Decoder-only





$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

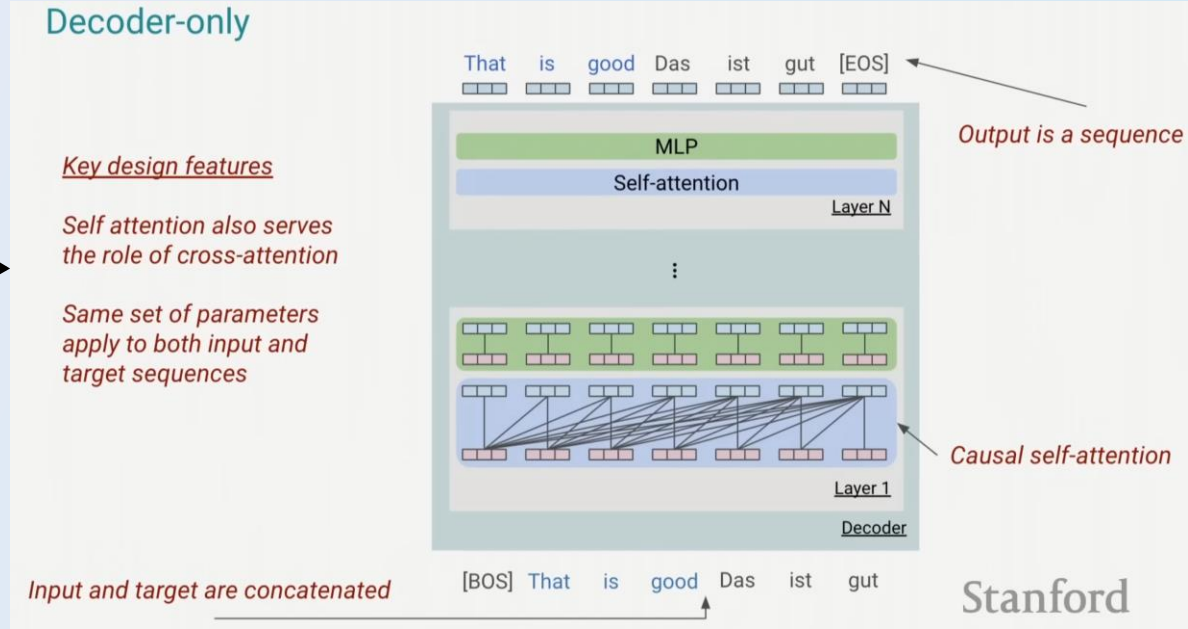


Photo Courtesy: Stanford University
'Shaping the Future of AI from the History of Transformer' by Hyung Won Chung, OpenAI at Stanford University, 11 April 2024

Foundation Models

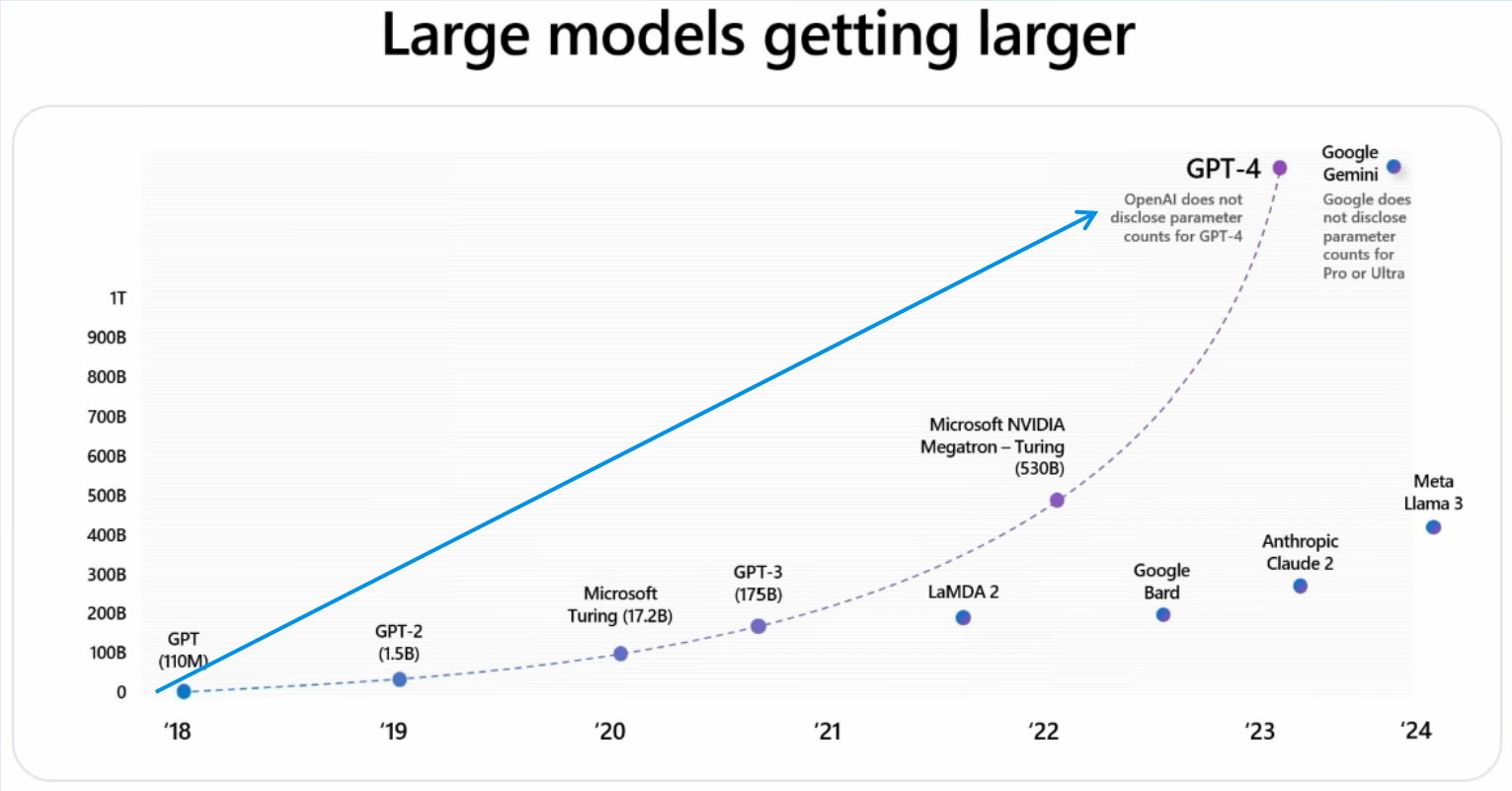
- Google invented the Transformer model ('Attention is All You Need', 2017, Vaswani et al.). Microsoft + OpenAI partnered to produce ChatGPT with GPT-4, and so on. Publicly released on 30 Nov 2022, ChatGPT reached 1 Million users in 5 days.
- Foundation Models have 2 essential components:

  1) Homogenization

  2) Emergence properties

  Industrialized Transformer LLMs, f(x) = homogenization abilities that "emerge" from training Billion-parameter LLMs.

  A parameter is a number inside a LLM that we can adjust to make it more / less accurate. Here's a list of 1 Billion+ Parameter LLMs. E.g., GPT-3 was a 175 Billion parameters LLM.



An excellent session to watch, 'Inside Microsoft AI Innovation with Mark Russinovich, MSBuild BRK256'

Quick Recap

- GPT-4 is built on GPT-3. GPT-3 is built on the GPT-2 architecture.
- But a fully trained, GPT-3 Transformer is a *Foundation Model*. Can effectively do 2 things:
  - Perform many tasks which it was *not* explicitly trained to do via Emergence.
  - GPT-3/GPT-4-generative abilities apply to various NLP tasks, including programming via Homogenization.

Token, Parameter, Weight… Correlation in LLM

Token:

Smallest unit of Analysis. Basic unit of meaning in a language.

Parameter:

Numerical value determining behavior of LLM.

Weight:

Type of param which when pre-trained, determines some form of Machine IQ.

Sentence Tokenization:
['This is a sentence.', 'This is another one.']

Word Tokenization:
['This', 'sentence', 'contains', 'several', 'words', '.']

Regular Expression Tokenization:
['Let', 's', 'see', 'how', 'to', 'tokenize', 'a', 'sentence']

Treebank Tokenization:
['There', 'are', "n't", 'that', 'many', 'tokenizers', '.']

… and others, White Space tokenization, BPE, Punkt, Multi-word Expression, Sub-word tokenizers etc.

Weights are of multiple types:

- Embedding weights: semantic meaning of token
- Self-attention weights: determine influence of tokens on each other within a sequence
- Feedforward weights: compute layer's output
- Bias weights: increase accuracy by virtue of adding bias weights to outputs of various layers in LLM

Models Listed as of, 13 July 2024

OpenAI

Ensure that artificial general intelligence (AGI) benefits humanity

Microsoft

Empower every person and organization on the planet to achieve more

Azure OpenAI Service

GPT-4o (Omni)	GPT-4 Turbo with Vision	GPT-4	Embeddings
gpt-4o-2024-05-13 (NEW)	gpt-turbo-2024-04-09 (NEW)	gpt-4 (0125-preview)* gpt-4 (vision-preview)* gpt-4 (1106-preview)*	text-embedding-3-large text-embedding-3-small text-embedding-ada-02

GPT-3.5 Turbo	DALL-E 3	Whisper	Text to Speech
gpt-35-turbo (0125) (NEW)	dall-e-3, 3.0	whisper, 001	tts, 001 tts-hd, 001

Please refer to https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models for a complete list of Models with information on model ID, both GA & Preview information, description, Max Request (Tokens), Training Data date, Availability details etc.

**Recommended not to use preview models in Production scenarios.*

Removing Confusions

Removing Confusions…. Dejargonization

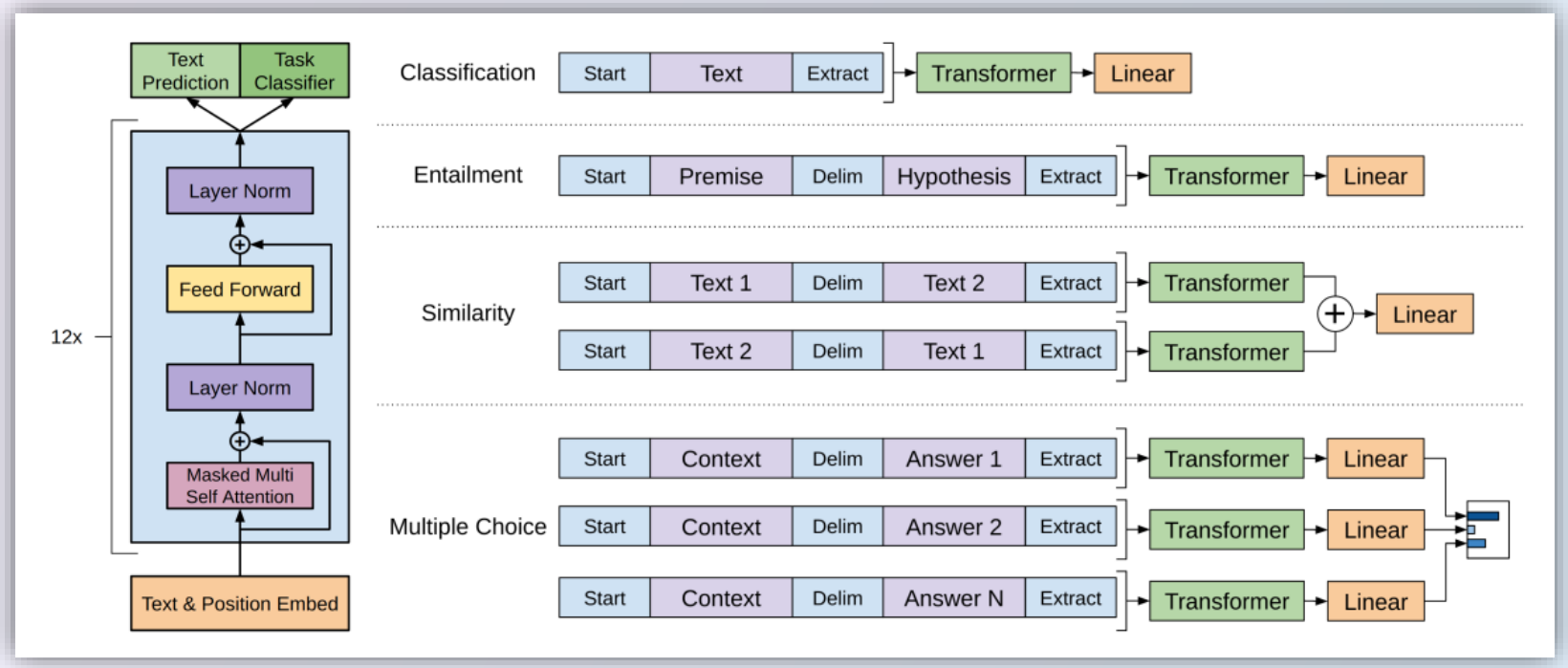
Model ID	Context Length (aka Context Window Size)	Max Request (tokens)
gpt-35-turbo (0613)	**4096**	**4096**

(0613) signifies snapshot of main LLM on a specific date.	*X = size (LLM's available memory with Q preloaded) + answer to be generated.*	*Area in Context Length reserved for "length of generated response".*

E.g., I'm using Azure OpenAI Model, **gpt-35-turbo** in my App. This LLM has limit of 4096 tokens. And, I set MAX_TOKENS= 1,000. Thus:

1. Generated response would be within 1000 tokens.
2. 4,096-1,000 = 3,096 tokens are left for your prompt that you can now send to the GPT LLM.

OpenAI GPT Models – Supervised or Unsupervised?



GPT training process in 2 stages:

Stage 1: Unsupervised Pre-Training

$$L_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta)$$

Stage 2: Supervised Fine-Tuning

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m).$$

Please refer to "Improving Language Understanding by Generative Pre-Training" by Radford et. al (2018), OpenAI for more details.

OpenAI GPT introduced a task-agnostic LLM through generative unsupervised Pre-Training & discriminative Fine-Tuning. In summary, GPT is self-supervised than totally unsupervised.

What Drives Innovation… I mean… Improvement?

Decoder Only:

Min regularization, maximum freedom resulting in innovative solutions.

Scale:

You require more parameters to capture & define contextual subtleties.

- *Too many params = Costly + Useless*
- *Too less params = Poor accuracy*
- *Balance to be found with trial & error.*

Task Generalization:

Essence of Generative AI models (e.g., OpenAI GPT) is to provide the most effective soln to potentially optimal number of tasks.

Diffusion

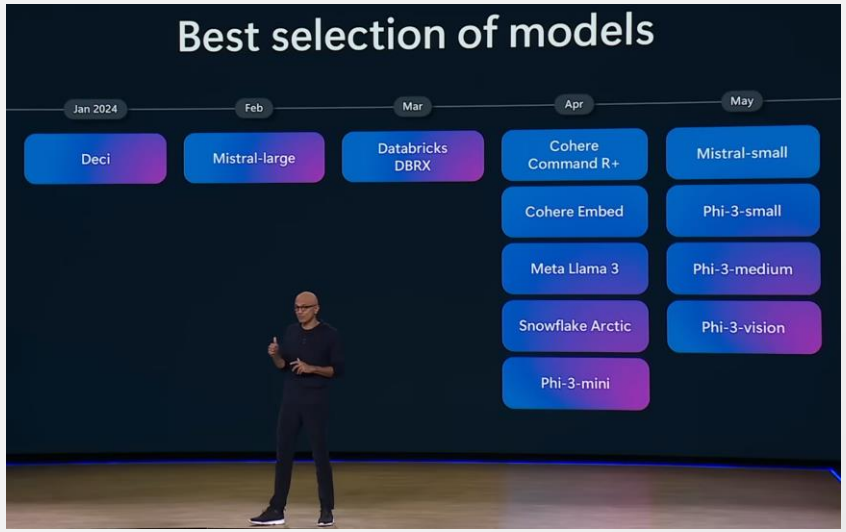
Application of Generative AI in different application sectors & its generality.

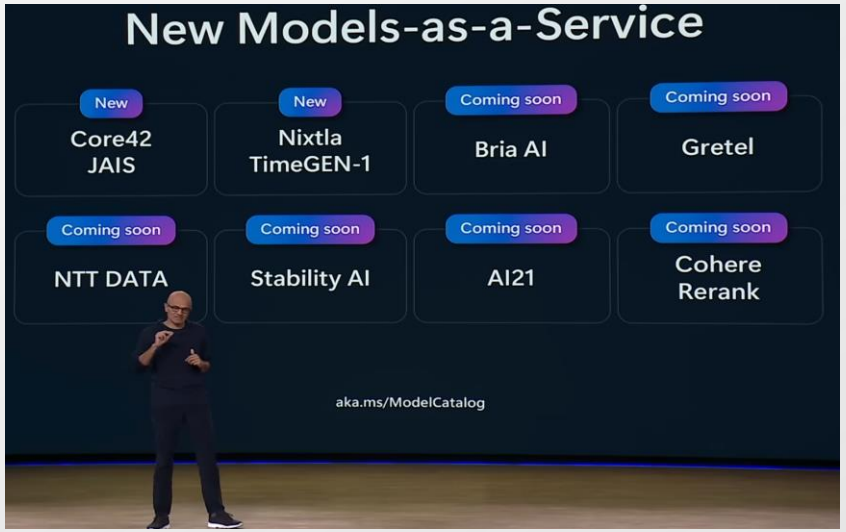
a) *Self-Service assistants*
b) *Development assistants*

Top3 Microsoft Build 2024 AI Announcements

Microsoft Build 2024 AI Announcements

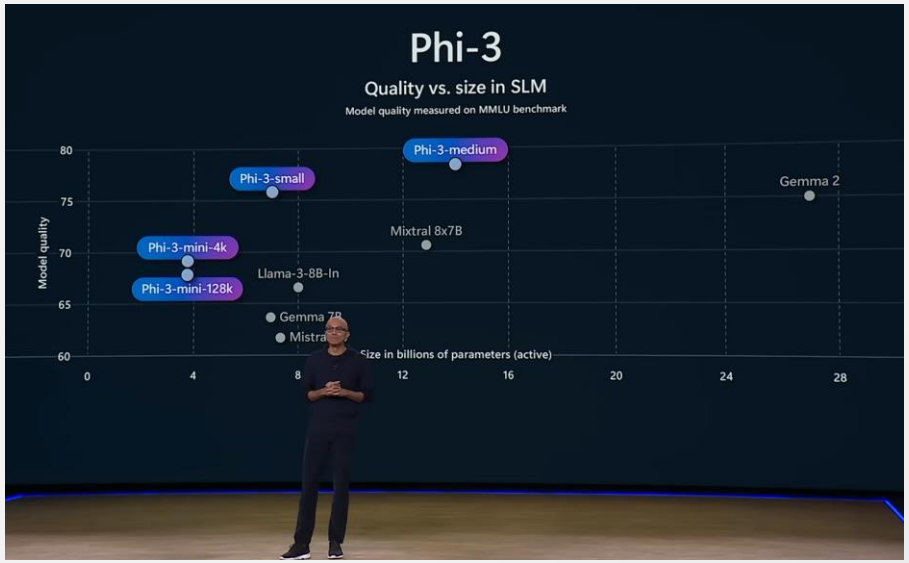
1





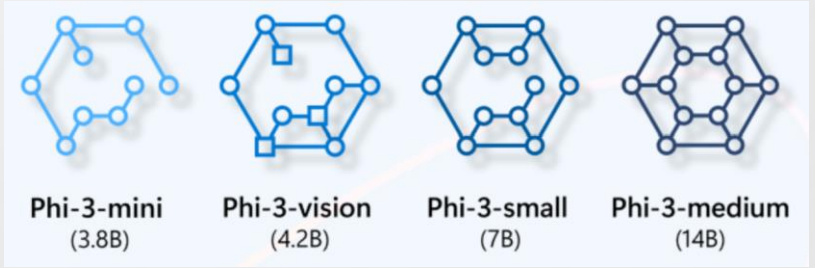
Microsoft Build 2024 AI Announcements

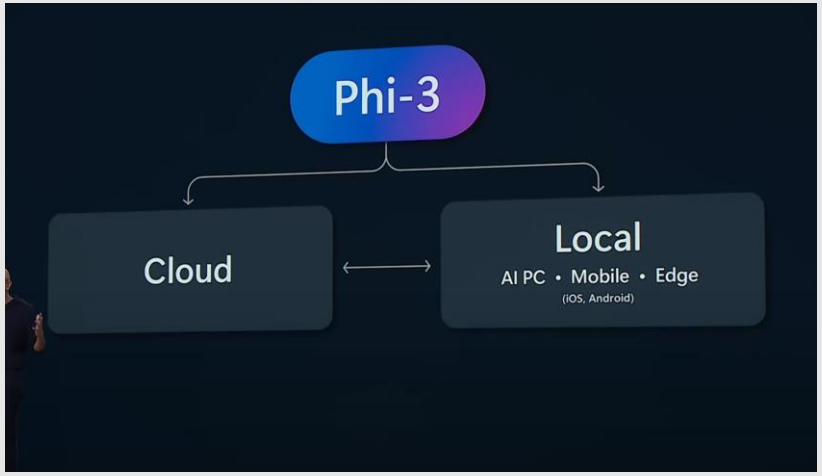
2



Microsoft's Small Language Models (SLMs)

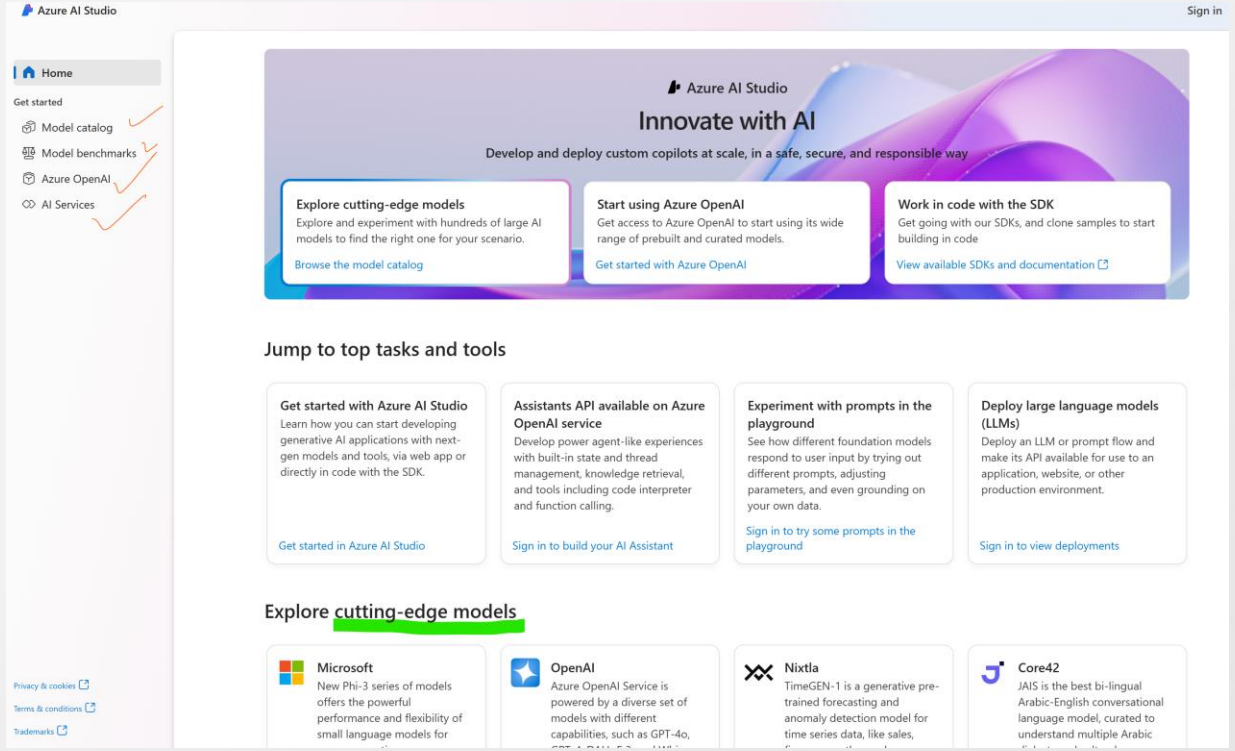
- 'Small' refers to # of parameters of model
- Lightweight and used wherein computational resources are limited (e.g., mobile/IoT edges) or where real-time inferencing is necessary.
- Microsoft offers Phi-3 Family of SLMs.





Microsoft Build 2024 AI Announcements

3



API & Model Choice | Complete AI toolchain | Responsible AI tools & practices | Production at Scale

1,600 Models from Microsoft, OpenAI, Databricks, Core42, Cohere, Mistral, Meta & others.

GenAI Industry Use-Cases

Azure OpenAI Service: 10 ways generative AI is transforming businesses > https://azure.microsoft.com/en-us/blog/azure-openai-service-10-ways-generative-ai-is-transforming-businesses/

Microsoft AI Customer Stories > https://www.microsoft.com/en-in/ai/ai-customer-stories

Industries are accelerating adoption of AI

AI is being embedded in standard business processes

31%

Use virtual agents

23%

Use computer vision

32%

Use natural language text understanding

24%

Use robotics

Source: The AI Index 2022 Annual Report, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022, Figure 4.3.3.

Demo

sugh@microsoft.com

- GPT-3.5-Turbo & GPT-4 optimized for conversational interfaces.

  // Meaning, conversation-in & message-out

- **Demo #1** on how to use GPT-3.5-Turbo from OpenAI Python SDK (1.35.10) with Chat Completion API

- **Demo #2** on Semantic Kernel with C# in Visual Studio 2022

  // Semantic Kernel is an OSS SDK from Microsoft (Python, C#, Java)

Quiz

6 Questions
Chance to win Swags!

Q&A
Thank you!