Potential for Generative AI in CFO Organizations

Large Language Models (LLM) at CGFS

SPEAKER: MICHEAL BARTH
MODERATOR: KYLE BROOKS & PAUL MARSHALL, THE MIL CORPORATION

Diplomacy in Action

Introductions

Michael Barth

Michael Barth is the Unit Chief for Software Development Operations and Implementations within Systems Development and Maintenance. Michael has been the department for 13 years working in various roles involving infrastructure, software modifications, and process improvements. Michael is currently leading a multiyear project to modernize the current development network and implement a multi-cloud hybrid solution as well as test emerging technologies such as AI for their use within CGFS. Prior to working for the Department Michael graduated from DePaul University in Chicago with a master's in information technology and a master's in Business.

Kyle Brooks

Mr. Brooks has over 20 years of experience working with federal financial management systems. He specializes in Business Process Reengineering (BPR), Robotic Process Automation (RPA), Artificial intelligence (AI), and IT driven Project Management. Over the past 5 years, he has been working within the Bureau of the Comptroller and Global Financial Services (CGFS) at DoS to establish and develop an RPA and Intelligent Automation program.

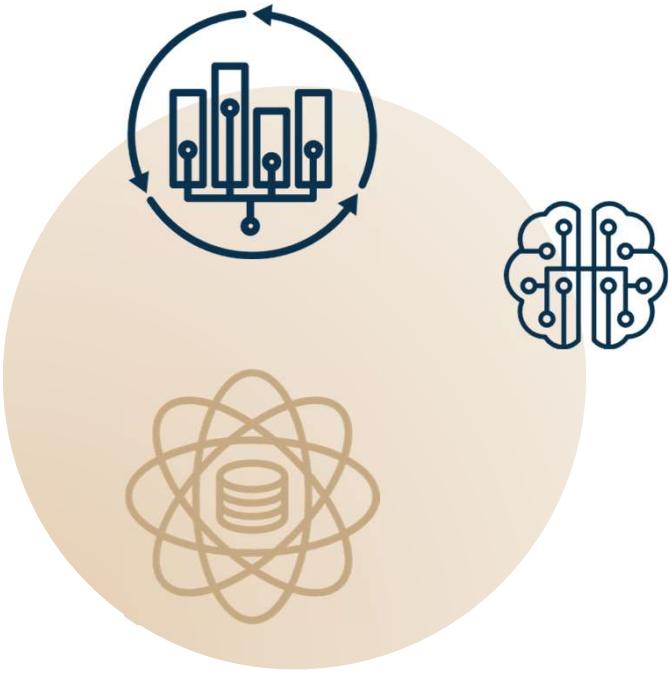
Our topic for today

Possible CFO use cases for Large Language Models

With live demonstrations from the Department of State

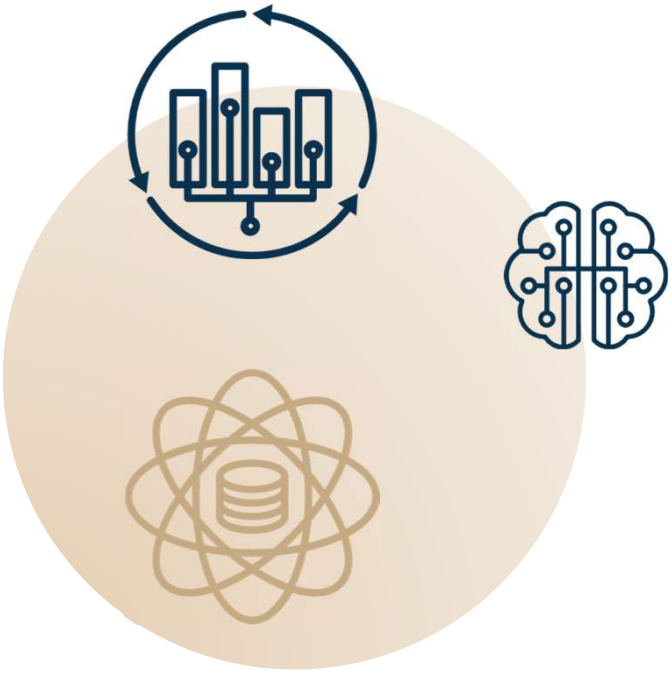
Question 1

How did you first learn about Large Language Models, and do you feel like the Department of State is heading in the right direction regarding LLM's?



Question 2

Do you believe that CGFS will benefit from Large Language Models? What will have the largest impact?

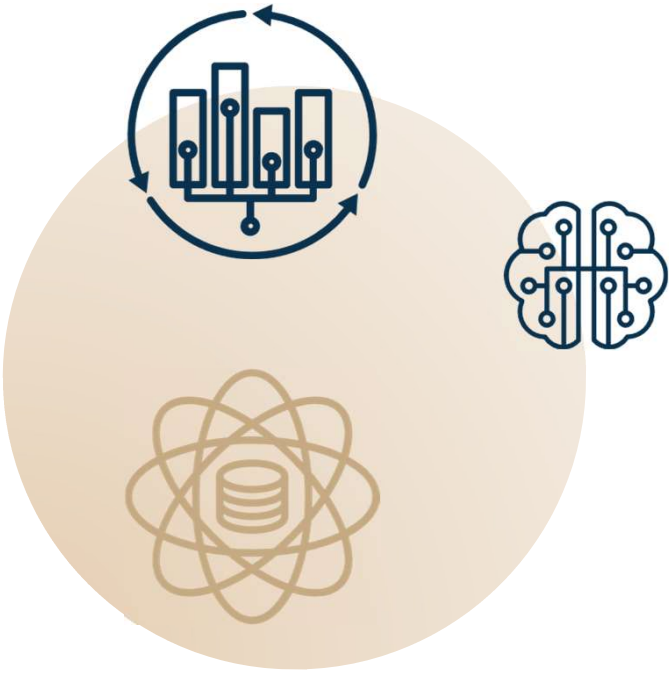


Demonstration 1

- Show a custom Retrieval-Augmented Generation (RAG) model
- Create a Vector Database in Real Time
- Talk with the Database once it is created

Question 3

Can you share some of the LLM use cases that CGFS is looking into?

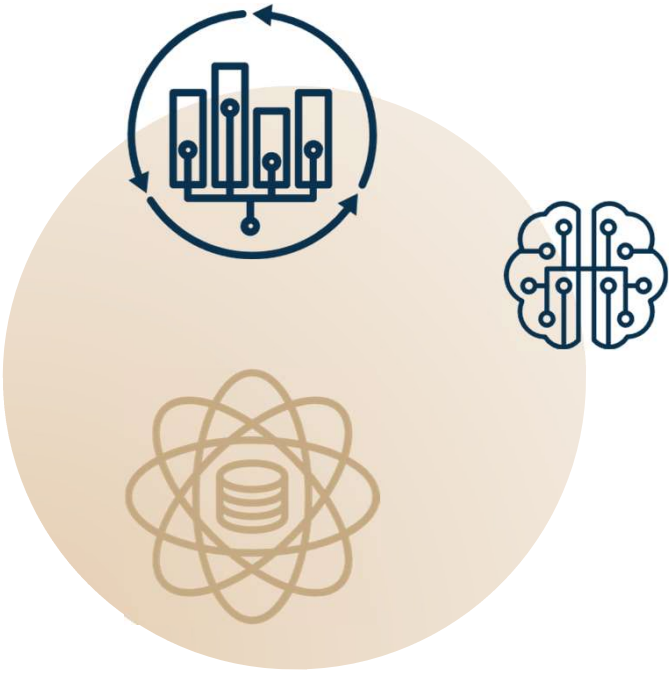


Demonstration 2

Show ChatGPT vision and how it can be utilized with financial management visualizations.

Question 4

What challenges or concerns do you have with LLMs at CGFS?



Demonstration 3

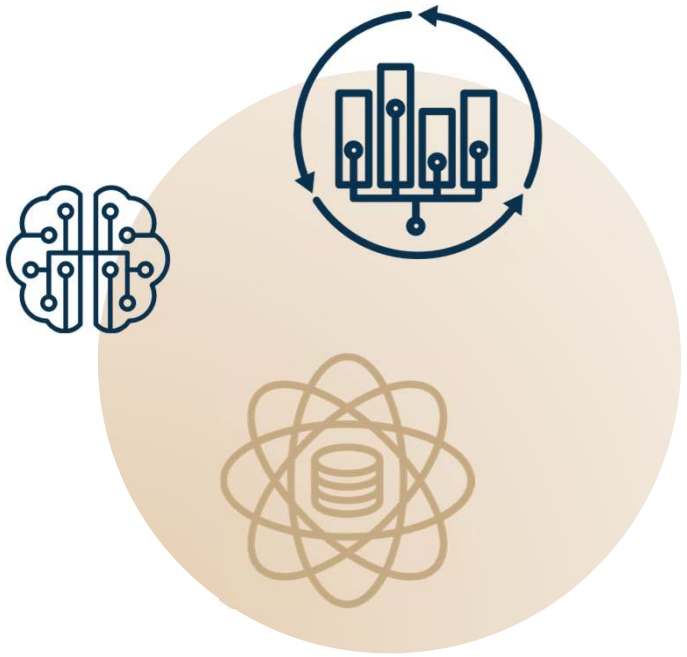
Show how custom GPTs can be used to work together to solve tasks and streamline processes.

Questions?

Large Language Models

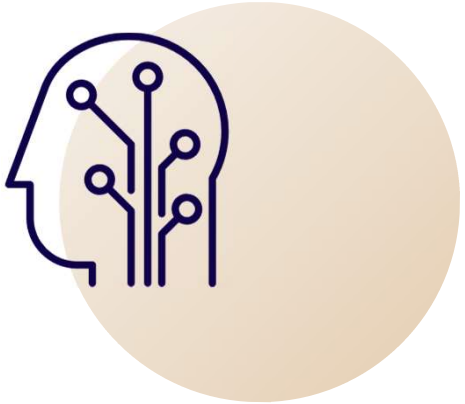
What is a Large Language Model?

- Type of artificial intelligence system that is trained on vast amounts of text data to generate, understand, and interact with human-like language.
- LLM's have shown emergent capabilities such as performing arithmetic, answering questions, summarizing passages.
- Different from a standard chat-bot, LLM's do not follow a script, and can provide recommendations and generate new ideas based off the user input.

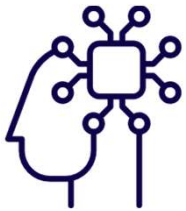


Neural Network/Deep Learning

- They are inspired by the neural networks in the human brain, consisting of layers of interconnected “nodes" that process data in a hierarchical manner.
- Uses a combination of input, hidden, and output node layers to mimic how neurons pass information between layers.
- Models are trained on large data sets.
- Uses a weight and threshold system to manage/pass/transform the information from node to node between layers.



Human Brain



Artificial Neural Network (ANN)

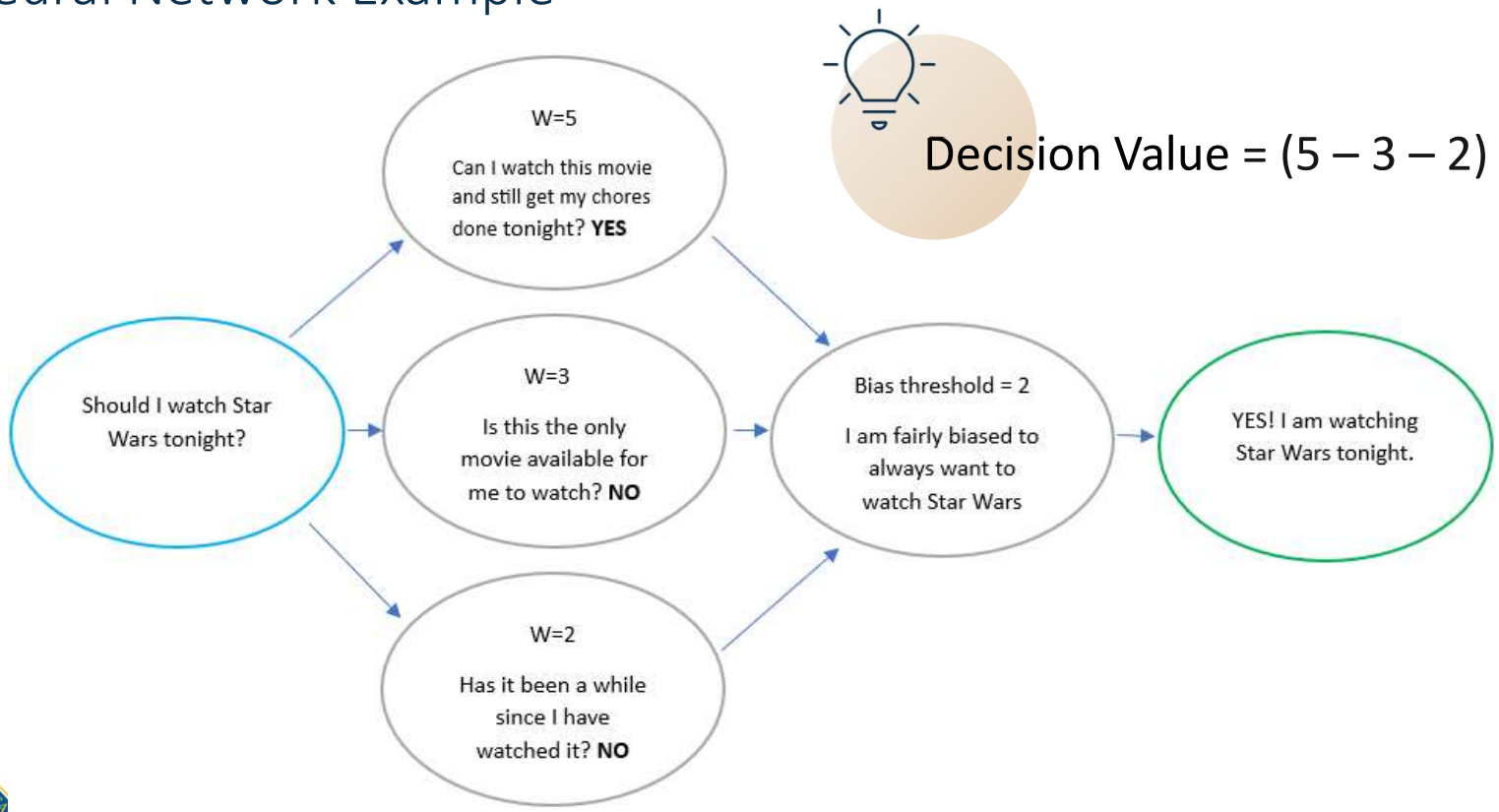
- Basic Unit: Neurons
- Connections: Synapses
- Operation: Electrical And Chemical signals are sent.
- Learning: Synapses (connections) are strengthened and weekend by training.
- Complexity: 80-100 billion Neurons with an estimated 100 trillion synapses (connections).

- Basic Unit: Nodes
- Connections: Parameters or Weights
- Operation: Passing of numerical values
- Learning: Training is performed with mass amounts of data that creates the connections and adjusts the Parameters (weights).
- Complexity: ChatGPT 4 Has 1.76 trillion parameters (weights).

Transformer Model

- Transforms input-text into a more meaningful representation of data.
- Developed by Google in 2017 that is based on a Graph Neural Network (GNN).
- Analyzes and understands the relationships between words in a sentence or a paragraph.
- Attention, focus, and understanding of text changes dynamically as it reads it. It works to understand the context.

Neural Network Example



Decision Value = (5 – 3 – 2) + (2) = 2

How are they trained?

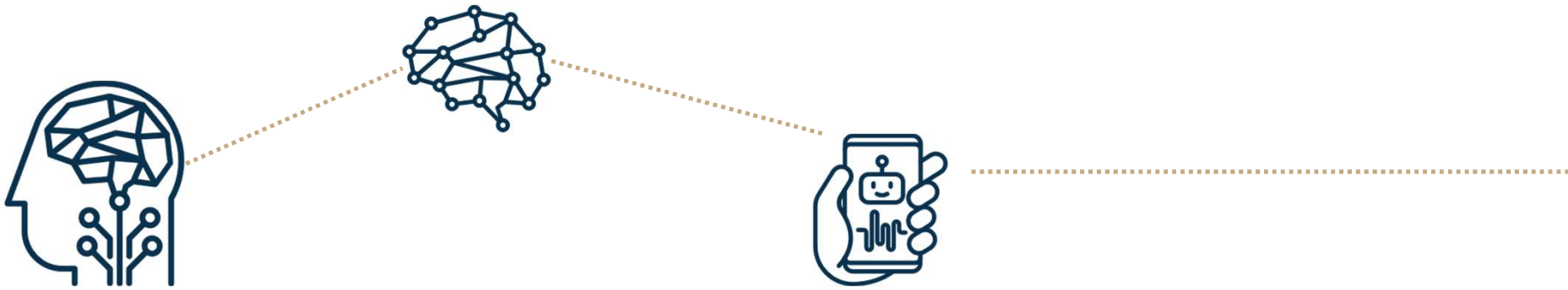
Single-Turn

Prompt: “What is five times seven?"

Response: “Five times seven is thirty-five.”

Multi-Turn

Prompt: Hey, did you catch the game last night?
Response: No, I missed it. Was it good?
Prompt: Yeah, it was incredible! Our team won in the last minute.
Response: That sounds exciting! I'll have to watch the highlights.



What is ChatGPT

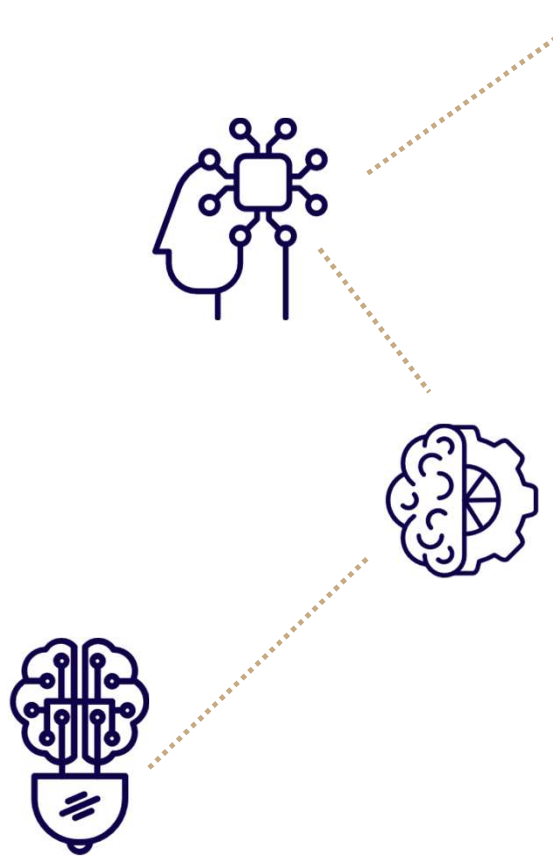
- Stands for Generative Pre-Trained Transformer
- Uses the Transformer neural-network structure
- Financially backed by Microsoft and recently FedRAMP approved for commercial cloud use on Azure platform
- One of the top-performing LLM's available
  - Competing with Google Gemini, Anthropic Claude Opus, and upcoming Meta Llama 3 (400B)
  - Latest model is the Chat GPT 4-Turbo release
- Trained on data from December 2024. **Does not train in real-time.**

How does ChatGPT work?

Needs a prompt from the user that tells it how to act and respond.

- System prompt example – You are a helpful and kind history expert that will only answer questions on medieval Europe. You will provide accurate responses and citations for all questions.
- User Prompt – What happened in the year 1231?

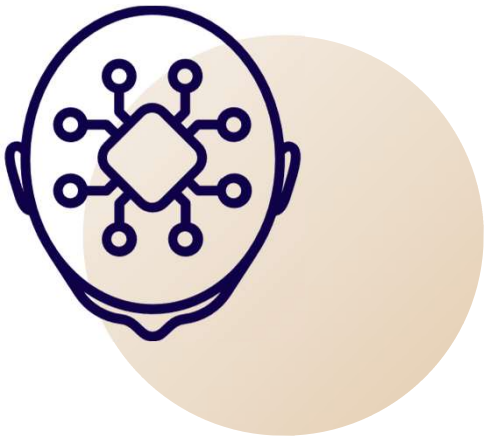
An existing model can be fine-tuned by the user by training it using proprietary information. (E.g., Bloomberg trained a model using its own financial papers and information.)



How does ChatGPT work?

Each model has a maximum context size that it uses to hold conversational memory, or to review external data for an answer.

- ChatGPT 3.0 = 4,096 tokens (3,000 words)
- ChatGPT 3.5 = 16,000 tokens (12,000 words)
- ChatGPT 4.0 = 32,768 tokens (24,500 words)
- ChatGPT 4.0 Turbo = 128,000 tokens (96,000 words)



For example, you could paste the entire book of Frankenstein in the prompt for 4.0 Turbo and ask a question on it without the need for training.

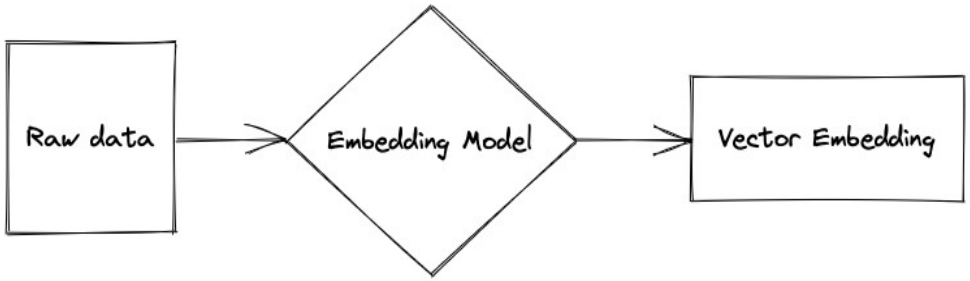
** Google recently released Gemini 1.5 Pro which has a context length of up to 10 million tokens. That is around 7 million words of text that can be consumed and understood at once. **

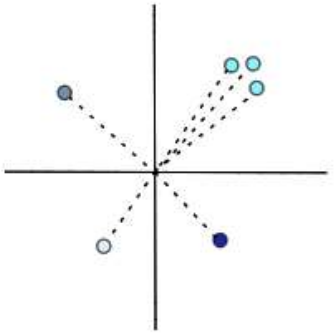
Retrieval-Augmented Generation (RAG) using Vector databases

Powerful tool for organizing and retrieving data based on similarity.

- *Example: a paragraph about Processing Vouchers will receive multiple vector values that are associated with Vouchering. These values can be compared to other paragraphs and documents to find similarities on the same topics.*

- Identifies and Organizes data in a vector graph based on relationships.
- Vector Values are determined by the trained neural network/model.





Where can I find this Technology?

Microsoft Azure – Received FedRAMP approval for the commercial cloud use of ChatGPT (Federal Information Assistant). Uses ChatGPT 4-Turbo.

Microsoft Co-Pilot – Multimodal instance of ChatGPT-4 Turbo that integrates with plugins and custom GPT's.

Bing – ChatGPT 4 can be accessed via Microsoft Bing

Google – New Gemini Ultra multimodal LLM being integrated by Google.

OpenAI.com – ChatGPT 4-Turbo can be found used via ChatGPT plus subscription

Kaggle.com – Website where you can download open-source models such as the new Meta Llama 3 (8B and 70B)

Claude.ai – Access to Anthropic's 3 models (Haiku, Sonnet, and Opus)

LLM Arena - (https://chat.lmsys.org/) – Try out all available closed and open-source models for yourself!

What is next?

- ChatGPT-5 anticipated for 2024 QTR 4 release.
  - Improved reasoning; Text to Video; Faster and close to real time communication and responses; Increased context size.
- Nano LLM's – Can run directly on User Hardware for focused tasks and do not require Internet connections.
- AI Agents – Uses Natural Language to interpret users needs and acts on applications. (Think of unbound Robotic Process Automation).
  - Example: Go check my email and see if my tickets were delivered, if so, transfer 2 of them to…