INTERNET2 TECH ex 24 BOSTON DEC 9-13

# AI Chatbots for Network Troubleshooting

Karl Newell
Network Software Architect, Internet2

# Agenda

- Large Language Models
- Retrieval Augmented Generation
- Fine tuning
- Agentic Chatbots
- Retrieval Augmented Fine Tuning

# Large Language Model

- **LLM**
  - type of artificial intelligence system trained on vast amounts of text data
  - can understand, generate, and manipulate human language by recognizing patterns in that data
  - perform tasks like writing, answering questions, and analyzing text while attempting to mimic human-like language understanding and generation.

- **ChatGPT, Claude, Llama, Mistral, many more**

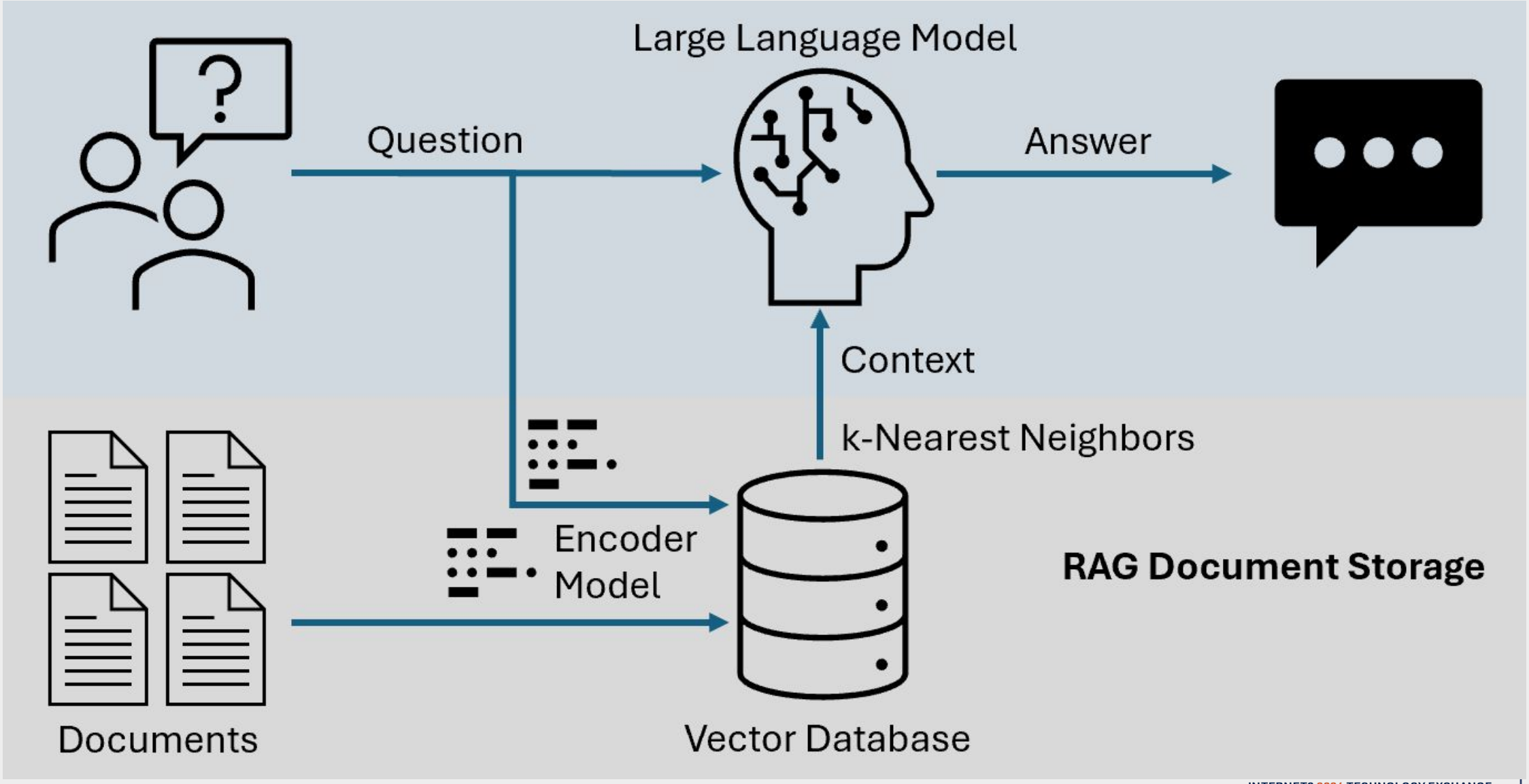
# LLMs

- **Insufficient for network troubleshooting**
  - Do not contain domain specific knowledge
  - Not capable of gathering data from the network

# Retrieval Augmented Generation

- **RAG**
- **Combines language models with external knowledge retrieval**
- **Benefits:**
  - Up-to-date information
  - Reduced hallucinations
  - Domain-specific knowledge
  - Cost-effective compared to fine-tuning

# RAG


Large Language Model
Question
Answer
Context
k-Nearest Neighbors
Encoder
Model
RAG Document Storage
Documents
Vector Database

# RAG

- **Insufficient for network troubleshooting**
  - Likely doesn’t include current network operational data
  - Would be data/time intensive to keep a vector DB updated

# Fine tuning

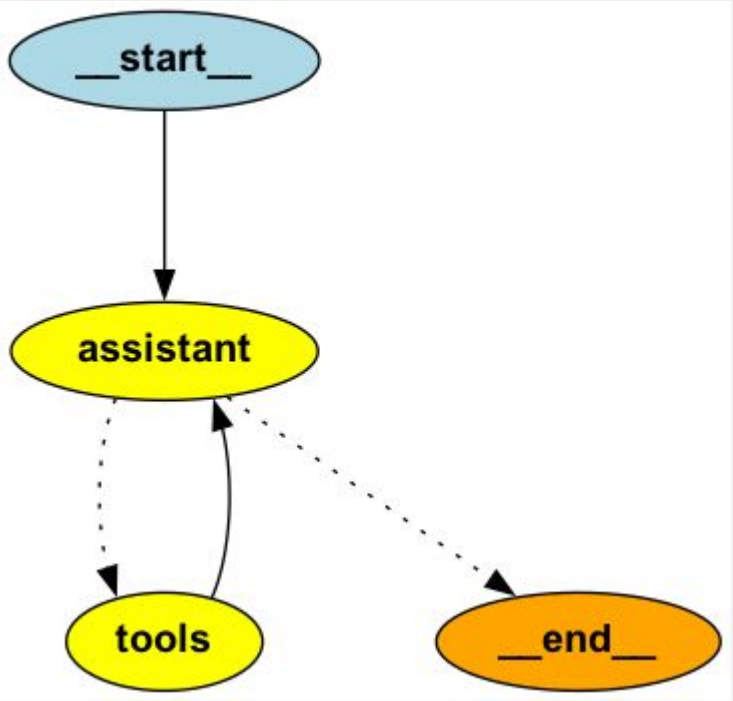
- **LLM fine tuning**
  - further training an LLM on a specific dataset to optimize its performance for particular tasks or domains
  - requires less computational resources than training a model from scratch.

# Fine tuning

- **Insufficient for network troubleshooting**
  - Expensive to constantly tune a model
  - The fine tuned data is rapidly out of date

# Agentic Chatbot

- **Leverages the ReAct framework**
  - **Re**asoning and **Act**ing
  - Thought, Action, Observation (TAO)
  - You provide a set of tools (e.g., Python functions) and the LLM determines which tool to run and next steps based on the results of the tool.



# Agentic Chatbot

- **Useful for network troubleshooting**
  - Leverages the reasoning capabilities of an LLM
  - Provides current network operational data to the LLM

# Retrieval Augmented Fine Tuning

- **RAFT**
  - enhances traditional LLM fine-tuning by retrieving relevant information from a knowledge base during the training process
  - allows the model to learn not just from its training examples but also from dynamically retrieved context

# RAFT

- Utility unknown

INTERNET2 TECH ex 24 BOSTON DEC 9-13

# Questions?