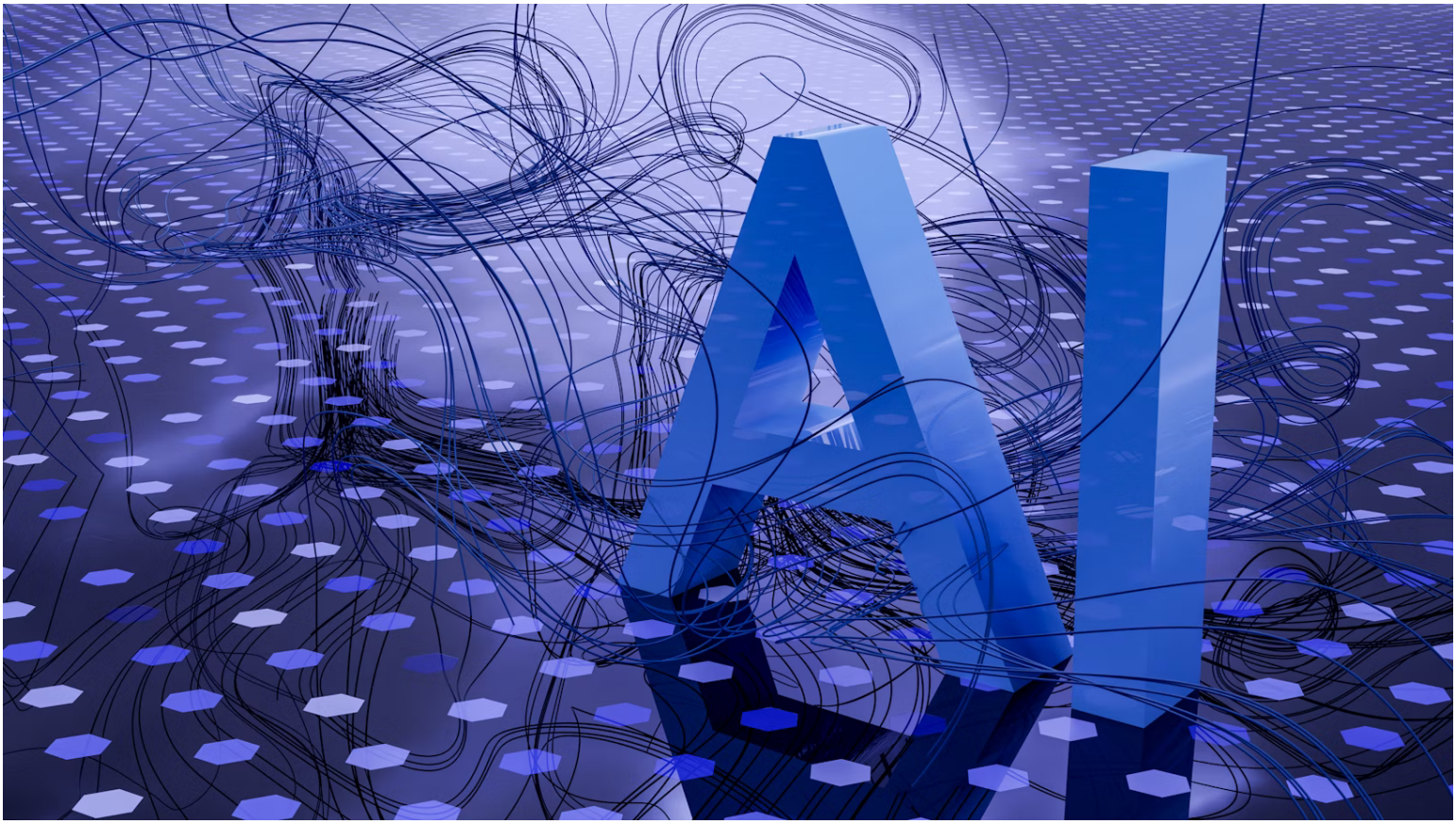
README.md



Exploring the LLM Frontier: From Hugging Face to RAG and Beyond

Generative AI is a subset of artificial intelligence that focuses on creating new content. Unlike traditional AI models that analyze data, generative AI models learn patterns from existing data and generate new, original content. This can range from text, images, music, and even code.

Why Learn About Generative AI?

Understanding generative AI is crucial for several reasons:

- Innovation: It's driving groundbreaking advancements in various industries, from art and design to healthcare and science.
- Career Opportunities: The demand for professionals skilled in generative AI is rapidly growing, presenting exciting career prospects.
- Problem-Solving: Generative AI offers innovative solutions to complex challenges, from content creation to drug discovery.
- Ethical Considerations: As generative AI becomes more powerful, it's essential to understand its potential

impacts and ethical implications.

Applications of Generative AI

The applications of generative AI are vast and expanding rapidly. Some key areas include:

- Content Creation: Generating text, images, music, and videos. Drug Discovery: Accelerating the development of new drugs by designing potential molecules.
- Design: Creating innovative designs for products and architecture.
- Customer Service: Developing advanced chatbots and virtual assistants.
- Education: Personalized learning experiences and intelligent tutoring systems.

Hugging Face: A Hub for Generative AI

Hugging Face has become a central platform for the generative AI community. It provides:

- Pre-trained models: A vast library of pre-trained models for various tasks, making it easier to get started.
- Datasets: Access to high-quality datasets for training and evaluation. Community: A thriving community of AI enthusiasts and researchers.
- Tools: A platform for building and sharing generative AI applications.

By capitalizing on Hugging Face, developers and researchers can accelerate their work in generative AI and build upon the collective knowledge of the community.

The Power Trio: Langchain, Gradio, and RAG for LLM Applications

Langchain, Gradio, and RAG form a potent combination for developing and deploying LLM applications. Each tool plays a crucial role in the process:

Langchain: The Architect

- Orchestrates LLM interactions: Connects LLMs to external data sources, databases, and APIs. Creates complex workflows: Builds intricate pipelines for tasks like summarization, question answering, and generation.
- Manages prompts and responses: Handles prompt engineering and response formatting.

Gradio: The Interface Builder

- Creates user-friendly interfaces: Builds interactive demos and web applications for LLMs.
- Visualizes outputs: Displays LLM generated text, images, or other media.
- Facilitates user input: Allows users to interact with the LLM through various input methods.

RAG: The Knowledge Booster

- Enhances LLM knowledge: Provides access to external information sources.
- Improves accuracy and relevance: Ensures LLM responses are grounded in factual data.
- Expands LLM capabilities: Enables the LLM to handle more complex and informative tasks.

Combined Power

When used together, these tools enable you to:

- Build sophisticated LLM applications: Create complex systems that leverage the strengths of LLMs and external data.
- Deploy models effectively: Create user-friendly interfaces for your LLM applications.
- Improve model performance: Enhance LLM accuracy and relevance through RAG. *Accelerate development: Streamline the development process with pre-built components.

In essence, Langchain, Gradio, and RAG are essential tools for anyone looking to develop and deploy cutting-edge LLM applications. They provide the foundation for building powerful, user-friendly, and informative AI systems.

Instructors: Carlos Lizárraga / Enrique Noriega.

Location: Albert B. Weaver Science-Engineering Library. Room 212.

When: Thursdays at 2PM.

[Program not definitive!]

Calendar

Date	Title	Topic Description	Wiki
09/05/2024 2PM	Hugging Face Models (NLP)	Hugging Face offers a vast array of pre-trained models for Natural Language Processing (NLP) tasks. These models cover a wide spectrum of applications, from text generation and translation to sentiment analysis and question answering.	
09/12/2024 2PM	Hugging Face Models (Computer Vision)	Hugging Face has significantly expanded its offerings beyond NLP to encompass a robust collection of computer vision models. You can find pre-trained models for a wide range of tasks, from basic image classification to complex image generation.	
09/19/2024 2PM	Hugging Face Models (Multimodal)	Hugging Face offers a diverse range of multimodal models, capable of processing and understanding multiple data modalities such as text, images, and audio. These models are at the forefront of AI research and development, enabling innovative applications.	
09/26/2024 2PM	Running LLM locally: Ollama	Ollama is an open-source platform designed to make running large language models (LLMs) on your local machine accessible and efficient. It acts as a bridge between the complex world of LLMs and users who want to experiment and interact with these models without	

		relying on cloud-based services.
10/03/2024 2PM	Introduction to LangChain	Langchain is an open-source Python library that provides a framework for developing applications powered by large language models (LLMs). It simplifies the process of building complex LLM-based applications by offering tools and abstractions to connect LLMs with other data sources and systems.
10/10/2024 2PM	Getting Started with Phi-3	Phi-3 is a series of small language models (SLMs) developed by Microsoft. Unlike larger language models (LLMs) that require substantial computational resources, Phi-3 models offer impressive performance while being significantly smaller and more efficient.
10/17/2024 2PM	Getting started with Gemini	Gemini is a large language model (LLM) developed by Google AI. It's designed to be exceptionally versatile, capable of handling a wide range of tasks and modalities, including text, code, audio, and images. This makes it a significant advancement in the field of artificial intelligence.
10/24/2024 2PM	Introduction to Gradio	Gradio is an open-source Python library that allows you to quickly create user interfaces for your machine learning models, APIs, or any Python function. It simplifies the process of building interactive demos and web applications without requiring extensive knowledge of JavaScript, CSS, or web development.
10/31/2024 2PM	Introduction to RAG	Retrieval-Augmented Generation. It's a technique that enhances the capabilities of Large Language Models (LLMs) by combining them with external knowledge sources.

Created: 07/18/2024 (C. Lizárraga)

Updated: 07/19/2024 (C. Lizárraga)

DataLab, Data Science Institute, University of Arizona.