Lecture 1: Generative AI - Shaping the Future of Creativity and Innovation

Modulo Cover

4	Introduction to Generative Models	The Concept of Generative Modelling, Comparison with Discriminative Models, Taxonomy of Generative Models (Probabilistic vs. Non-Probabilistic), Applications of Generative Models (Image Generation, Text Creation, etc.), Benefits and Challenges, Generative Adversarial Networks (GANs): Architecture, Training Process, Applications (Deepfakes, Style Transfer) - Variational Autoencoders (VAEs): Architecture, Training Process, Applications (Anomaly Detection, Data Augmentation) - Autoregressive Models, Transformer-Based Models, Attention Mechanisms - Diffusion Models, Multi-modal models: Fundamentals	*AICTE-prescribed syllabus:* *https://www.aicte-india.org/sites/default/files/Model_Curriculum/CS%20(AI&ML).pdf* *International Academia:* *https://www.coursera.org/learn/generative-ai-introduction-and-applications?*	6	(i) Use of Transformer for generating synthetic datasets (ii) Use of stable diffusion model in converting black and white images to color or transforming low-resolution images to high-resolution images (iii) Voice and media processing
Textbook: Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play by David Foster (Chapter 1, 3, 4, 5, 8, 9) - Handouts					
5	Large Language Models (LLMs)	- Introduction to Language Models (LMs) - Benefits and Capabilities of LLMs - Text Generation with LLMs (Creative Writing, Code Generation, Chatbots) - Machine Translation with LLMs - Text Summarization with LLMs - Question Answering with LLMs - Sentiment Analysis and NER with LLM - Hallucination in Generative Models: Understanding and Mitigating Untrue or Unrealistic Outputs - Fine-tuning LLMs with domain-specific information (introduce the concepts of matrix multiplication, LORA, etc.)	*International Academia:* *https://www.coursera.org/learn/generative-ai-with-llms*	4	(i) A project to demonstrate how to use one of the LLMs like Llama, GPT, or Gemini (ii) The objective is to provide hands-on experience in fine-tuning, API integration, and user interface building on top of an LLM
Textbook: Prompt Engineering for Generative AI by James Phoenix and Mike Taylor (Chapters 2, 3) - Handouts					

Assignment 1:

Wahedi, Hadid J., Mads Heltoft, Glenn J. Christophersen, Thomas Severinsen, Subrata Saha, and Izabela Ewa Nielsen. "Forecasting and inventory planning: an empirical investigation of classical and machine learning approaches for svanehøj's future software consolidation." Applied Sciences 13, no. 15 (2023): 8581.

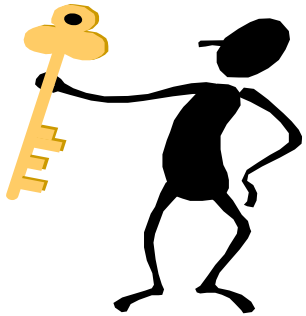
1. **Artificial Neural Network (ANN)**
2. **Long Short-Term Memory (LSTM)**
3. **Support Vector Regression (SVR)**
4. **Random Forest (RF)**
5. **Wavelet-ANN (W-ANN)**
6. **Wavelet-LSTM (W-LSTM)**
7. **Q-learning**
8. **Deep Q Network (DQN)**

Problem Solution

Real-World Problem

Formulation of Abstract Problem

Solve the Abstract Problem



Implement the Solution

Interpret the Solution

Creative problem-solving can be conceived as a search for solutions on a landscape

WHY???

AI can perform exceptionally well in tasks with **clear rules, patterns, and objectives**, it is less clear whether AI can aid in creative problem-solving tasks, which often require **abstract, nuanced, and iterative thinking, social interactions, and broad search** for distant knowledge and alternative perspectives

WHAT???

Generative AI—a type of AI technology capable of producing new content, such as text, images, audio, or video, based on patterns learned from existing data—can enhance creative problem-solving through human-guided AI partnerships

Fun Fact???

generate a creative solution for a new type of sustainable urban transportation

Here's a creative solution for a new type of sustainable urban transportation:

Concept: Urban Glide - A Modular, Solar-Powered Pod System

Overview: Urban Glide is a network of lightweight, solar-powered pods that operate on elevated tracks above urban streets. The system combines autonomous technology, modular design, and renewable energy to offer a clean, efficient, and scalable mode of trans[illegible]tation for city dwellers.

The Potential of Large Language Models

Large language models (LLMs) such as ChatGPT , Gemini, and Claude have demonstrated amazing capabilities in carrying out tasks previously considered unattainable by artificial intelligence.

ChatGPT, one of the most prominent LLMs, has seen unprecedented adoption since its launch in November 2022.

- Current generative AI and other technologies can potentially automate work activities that absorb 60 to 70 percent of employees' time today (Segev Wasserkrug, IBM Research)!!!!

- As LLMs continue to advance, they are poised to revolutionize how software is developed, enabling faster, more efficient, and more productive coding practices.

LLMs Overview

An LLM is a language model designed to calculate a probability distribution over word sequences in a language (Jurafsky and Martin). An LLM is a conditional probability language model: given a sequence of words (or parts of words known as tokens), $x_1, \ldots, x_t$, $i \in \{1, \ldots, t\}$ where each x_i belongs to the set of all possible tokens X, the model calculates the conditional probability $Pr(x_{t+1} | x_t, \ldots, x_1)$ of the next token x_{t+1} for each possible $x_{t+1} \in X$.

Jurafsky D, Martin JH (2024) N-gram Language Models (Draft available online), 3 edition, URL https://web.stanford.edu/~jurafsky/slp3.

LLMs Overview

What differentiates LLMs from other types of conditional probability language models are their size and the way they are trained. In terms of their size, LLMs are deep neural network (DNN) machine learning models– mathematical models based on the principles of interconnected neurons – **with billions or trillions of parameters**

There are now many open-source and proprietary LLMs. Open-source LLM examples include Flan-T5 (80 million to 11 billion parameters), Gemma (2 and 7 billion-parameter versions), Llama (7 to 70 billion parameters) and Mixtral (8 sets of 7 billion parameters)

The Technology Underlying LLMs

Attention mechanisms: an attention function can be thought of similarly to how a person would look for relevant information in a library. Specifically, it can be thought of as mapping a query and a set of key-value pairs to an output. In the context of LLMs, the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. The attention mechanism used in the original transformer architecture is called scaled dot-product attention. The input consists of queries and keys of dimension d_k and values of dimension dv. We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values.

Problem Solving

When the best method for solving a problem is uncertain, one strategy to enhance innovative search is to utilize a variety of different approaches, or **“parallel paths,”** as this breadth can improve overall solution quality (Nelson 1961). The parallel path effect suggests that developing multiple solutions to the same problem increases the likelihood of achieving a high-quality outcome. Arguably, utilizing various approaches is particularly critical when the objective is to maximize the quality of a few top ideas instead of many average ones

Crowdsourcing

Crowdsourcing contests leverage a diverse pool of solvers with differing backgrounds and experiences to increase the number of parallel paths to solve a problem and improve solution quality. However, crowdsourcing can be resource intensive and statistically inefficient because of the volume of low-quality submissions

Using LLMs To Advance Human-AI Creative Problem Solving

AI is a broad field within computer science that seeks to create systems capable of performing tasks that **typically require human intelligence**. This includes activities such as learning, reasoning, problem-solving, perception, and understanding language. Machine learning (ML), a subset of AI, focuses on algorithms that allow machines to analyze data, learn from it, and make predictions. Unlike traditional programming, ML models evolve their performance as they process more data, eliminating the need for explicit programming in every scenario.

Generative AI leverages ML models, trained on large data sets to produce outputs that mirror the input data distribution. LLMs are a type of generative AI specifically designed to process and generate human language

Components of LLM

1. Tokenization: The input text is divided into smaller units called tokens, which can be words, subwords, or characters. This process allows the model to process the text more efficiently.

2. Embedding: Each token is mapped to a high dimensional vector representation, capturing the semantic and syntactic relationships between tokens. This embedding layer allows the model to understand the meaning and context of words

Components of LLM

3. Transformer architecture: LLMs commonly use transformer neural networks. Transformers utilize self-attention mechanisms, which allow the model to weigh the importance of different tokens within a sequence, enabling it to capture long-range dependencies and context more effectively.

4. Autoregressive language modeling: This approach trains the model to predict the next token in a sequence based on all preceding tokens. The model learns to generate text by iteratively predicting each subsequent token, conditioned on the previously generated ones.

Components of LLM

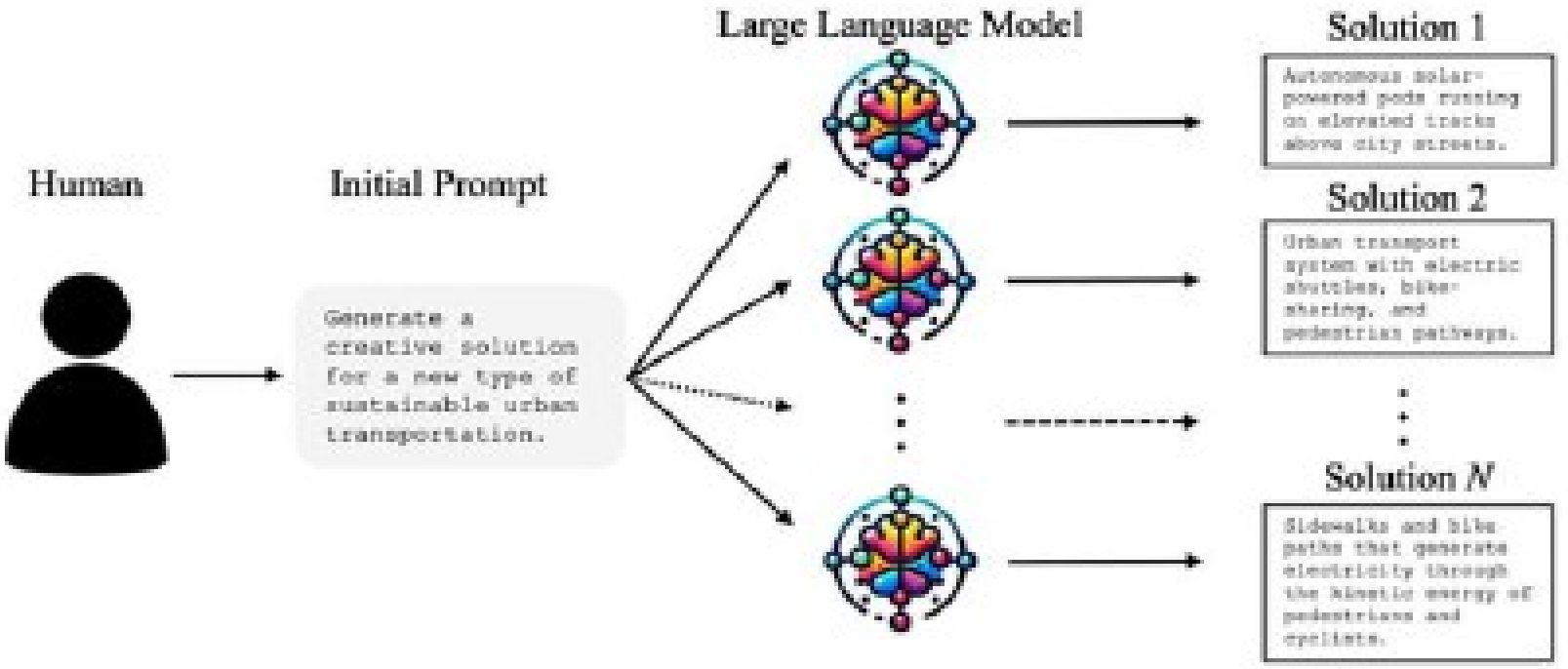
5. Optimization: The model's parameters are updated through an iterative process called gradient descent, which minimizes the difference between the model's predictions and the actual next tokens in the training data. This process allows the model to learn the patterns and relationships within the language. Because of a large number of parameters and complexity, optimizing LLMs requires substantial computational resources

6. Fine-tuning and alignment: State-of-the-art LLMs are typically further refined through supervised learning, where the models are provided with human annotated data sets that exemplify desired behaviors or task-specific outputs.

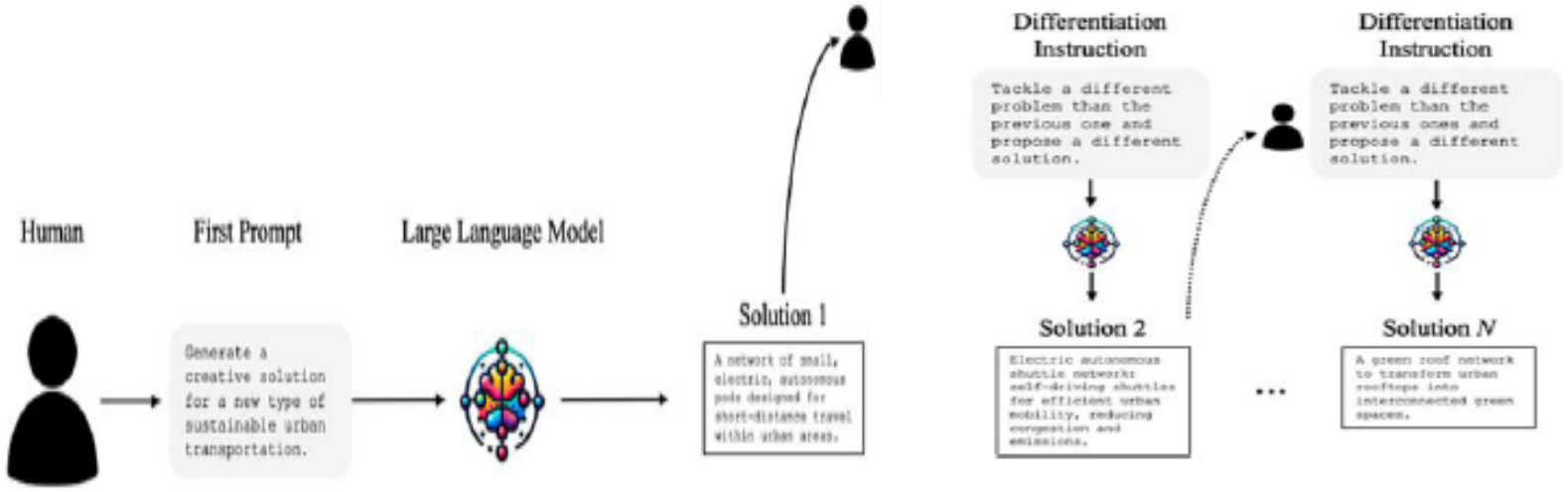
Strategic Prompt Engineering

Prompt engineering, the process of designing input prompts to guide the model's output, plays a crucial role in shaping the generated text. As LLMs currently lack independent agency, the quality and relevance of their outputs heavily depend on humans' ability to skillfully craft prompts, emphasizing the necessary collaboration between humans and AI

Independent Search



Differential Search



Prompting Techniques for LLMs

- **Providing Detailed and Precise Instructions:** It is important to make the prompt as precise and detailed as possible, as this helps direct the LLM to the desired response.
- **Role-Playing:** In this technique, the LLM is instructed to take on a specific persona, point of view, or role within the context of the task. This can help guide the model's responses and outputs to align with the desired perspective or domain expertise. Role-playing can be combined with other prompting techniques to tailor the LLM's behavior further.
- **Zero-Shot Learning:** In this paradigm, an LLM is directly tasked with a new task without additional examples or training. Success relies on clear instructions and the LLM's ability to extrapolate from its vast knowledge base.

Prompting Techniques for LLMs

Few-Shot Learning: Here, several examples are provided that demonstrate both the input and output formats of the desired task. This in-context learning helps the model understand the task structure and biases it towards generating similar outputs.

Chain-of-Thought (CoT) Prompting is intended for complex reasoning tasks. The CoT technique is a few-shot prompting technique in which the provided examples include the detailed steps required to reach the result. A CoT prompt therefore directs the model to break down complex problems into smaller steps and provide intermediate reasoning. CoT prompts encourage LLMs to explicitly verbalize their reasoning process, leading to improved accuracy and transparency. Recent research suggests that for more complex reasoning tasks, simply adding the phrase "Let's think step by step" before the answer can significantly boost LLM performance, even in a zero-shot setting

Prompt Chaining/Iterative Prompting: This approach involves breaking down a complex task into a series of smaller, interconnected prompts. The output from one prompt becomes the input for the next, allowing the LLM to build up to the final desired output gradually. This iterative process can help improve the coherence and quality of the generated content, especially for tasks that require multiple steps or evolving context .