Emergence of Multimodal AIs

For Prelims: Emergence of Multimodal AIs, **AI (Artificial Intelligence)**, Human-like Cognition, OpenAIs ChatGPT, Google's Gemini model.

For Mains: Emergence of Multimodal AIs and their implications, Developments and their applications and effects in everyday life

Source: TH

Why in News?

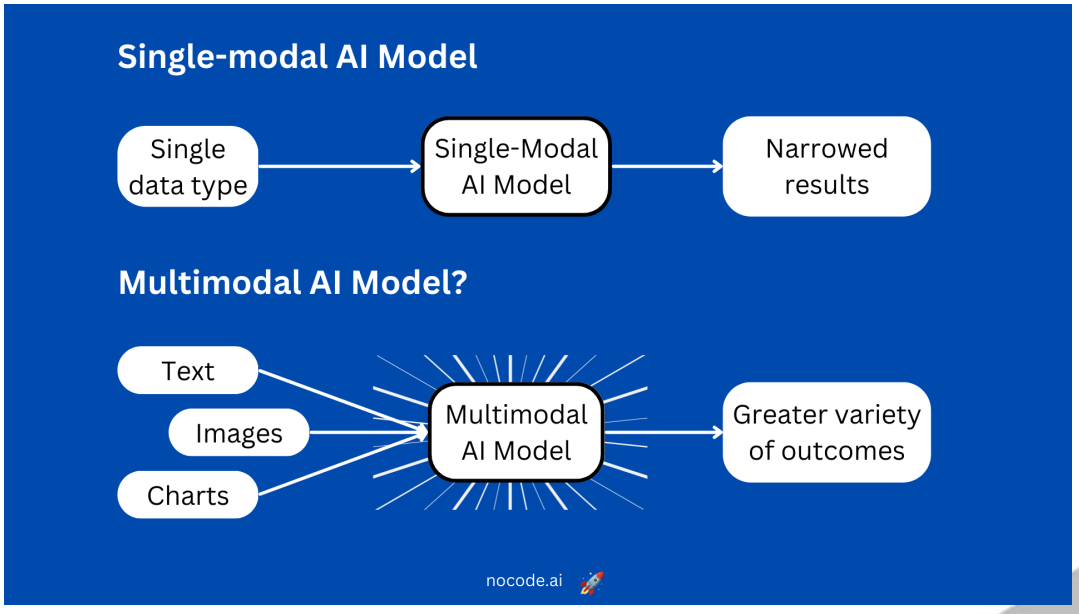
There has been a paradigm shift within **AI (Artificial Intelligence)** towards Multimodal Systems, allowing users to engage with AI through a combination of text, images, sounds, and videos.

- These systems aim to replicate human-like cognition by encompassing multiple sensory inputs.

What are Multimodal AI Systems?

- **About:**
  - Multimodal AI is artificial intelligence that combines **multiple types, or modes, of data to create more accurate determinations**, draw insightful conclusions or make more precise predictions about real-world problems.
  - Multimodal AI systems train with and use video, audio, speech, images, text and a range of traditional numerical data sets.
  - **For Example:** Multimodal audio systems follow **similar principles,** with Whisper, OpenAI's open-source speech-to-text translation model, **serving as the foundation for GPT's voice processing capabilities.**

//



- **Recent Developments in Multimodal AI:**
  - **OpenAIs ChatGPT:** OpenAI announced enhancements to its GPT-3.5 and GPT-4 models, allowing them to analyze images and engage in speech synthesis, enabling more immersive interactions with users.
    - It is working on a project named "Gobi," which aims to create a multimodal AI system from scratch, distinct from the GPT models.
  - **Google's Gemini Model:**
    - Another major player in the field is **Google' new yet-to-be-released multimodal large language model Gemini.**
      - Due to its huge collection of images and videos from its search engine and YouTube, **Google had a clear edge over its rivals in the multimodal domain.**
      - It puts immense pressure on other AI systems to rapidly advance their multimodal capabilities.

What are the Advantages of Multimodal AI over Unimodal AI?

- Multimodal AI, unlike unimodal AI, leverages **diverse data types such as text, images**, and audio, offering a richer representation of information.
- This approach enhances contextual understanding, resulting in more accurate predictions and informed decisions.
- By fusing data from **multiple modalities, multimodal AI achieves better performance**, increased robustness, and the ability to handle ambiguity effectively.
- It broadens **applicability across various domains and enables cross-modal learning.**
- Multimodal AI provides a more **holistic and human-like understanding of data,** paving the way for innovative applications and a deeper comprehension of complex real-world scenarios.

What are the Applications of Multimodal AI?

- It finds applications in **diverse fields, including autonomous driving,** robotics, and medicine.
  - For example, In medical field, the analysis of complex datasets from **CT Scans** and identifying genetic variations, simplifying the communication of results to medical professionals is very crucial.
- Speech translation models, such as Google Translate and Meta's SeamlessM4T, also **benefit from multimodality,** offering translation services across various languages and modalities.
- Recent developments include Meta's ImageBind, a multimodal system capable of processing text, visual data, audio, temperature, and movement readings.
  - The potential for integrating additional sensory data like touch, smell, speech, and brain MRI signals is explored, enabling **future AI systems to simulate complex**

environments.

What are the Challenges of Multimodal AI?

- **Data Volume and Storage:**
  - The diverse and voluminous data required for **Multimodal AI poses challenges** in terms of data quality, storage costs, and redundancy management, making it expensive and resource-intensive.
- **Learning Nuance and Context:**
  - Teaching AI to **understand nuanced meanings** from identical input, especially in languages or expressions with context-dependent meanings, proves challenging without additional contextual cues like tone, facial expressions, or gestures.
- **Limited and Incomplete Data:**
  - Availability of complete and **easily accessible data sets is a challenge.** Public data sets may be limited, costly, or suffer from aggregation issues, affecting data integrity and bias in AI model training.
- **Missing Data Handling:**
  - Dependency on **data from multiple sources can result in AI malfunctions** or misinterpretations if any of the data sources are missing or malfunctioning, causing uncertainty in AI response.
- **Decision-Making Complexity:**
  - Neural networks in Multimodal AI may be complex and challenging to interpret, making it **difficult to understand how AI evaluates** data and makes decisions. This lack of transparency can hinder debugging and bias elimination efforts.

Conclusion

- The advent of multimodal AI systems represents a significant advancement in the field of artificial intelligence.
- These systems have the potential to revolutionize various industries, enhance human-computer interactions, and address complex real-world problems.
- As AI continues to evolve, multimodality is poised to play a pivotal role in achieving artificial general intelligence and expanding the boundaries of AI applications.

UPSC Civil Services Examination, Previous Year Question (PYQ)

Q. Introduce the concept of Artificial Intelligence (AI). How does AI help clinical diagnosis? Do you perceive any threat to privacy of the individual in the use of AI in the healthcare? **(2023)**