edureka!
a Veranda Enterprise | ILLINOIS TECH

POST GRADUATE PROGRAM IN GENERATIVE AI AND ML

Natural Language Processing

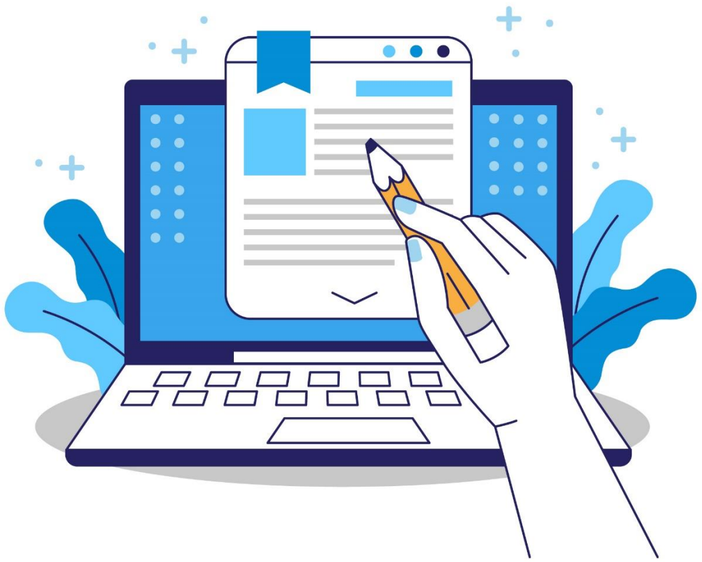
edureka!
a Veranda Enterprise

ILLINOIS TECH

Speech and Multimodal NLP

Topics

- Fundamentals of Speech Recognition
- Automatic Speech Recognition (ASR) Models (e.g., Whisper, DeepSpeech)
- Challenges in Speech Processing
- TTS Applications in NLP
- Demonstration
- Introduction to Multimodal Learning
- Vision-Language Models (CLIP, BLIP)
- Integrating Text and Speech in NLP Applications
- Generative Models for Multimodal Content
- Model Compression (Quantization, Pruning, Knowledge Distillation)
- Real-Time NLP for Edge Devices
- Explainability in Deep NLP Models
- Emerging Trends (GPT-5, HyperNetworks, Autonomous AI Agents)



Learning Objectives

By the end of this lesson, you will be able to:

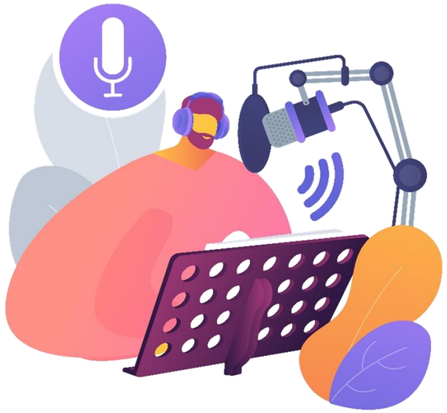
- Describe how speech and text are processed in NLP using ASR, TTS, and tokenization techniques
- Compare ASR models (e.g., Whisper, DeepSpeech) and vision-language models (e.g., CLIP, BLIP)
- Analyze challenges in speech processing and real-time NLP for edge deployment
- Apply multimodal and generative models to integrate and generate text, audio, and images
- Evaluate the use of model compression, explainability, and ethical principles in NLP systems
- Summarize emerging trends in NLP such as GPT-5, HyperNetworks, and autonomous AI agents

Fundamentals of Speech Recognition

What is Speech Recognition?

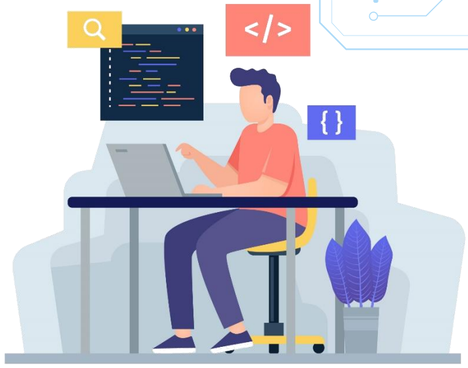
Speech Recognition is the process of converting spoken language into text, allowing computers to understand and respond to human speech.

Key Components

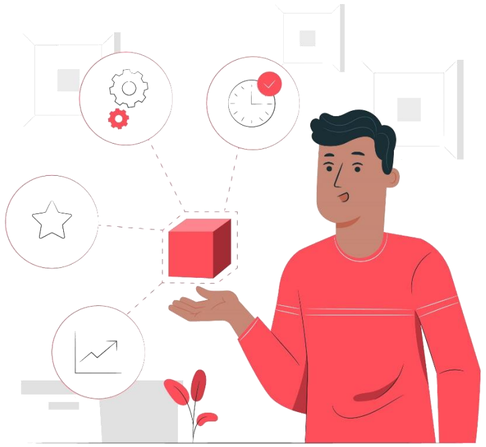


Audio Input

Acoustic Model



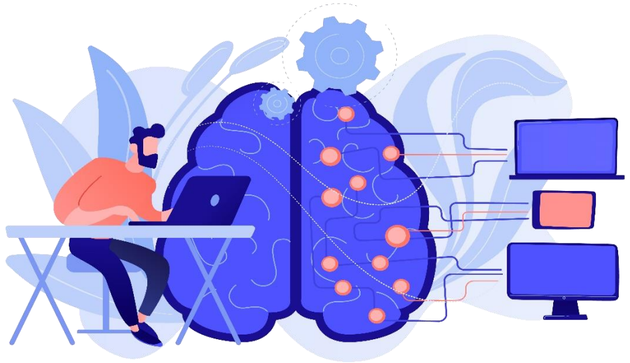
Decoder



Feature Extraction

Language Model

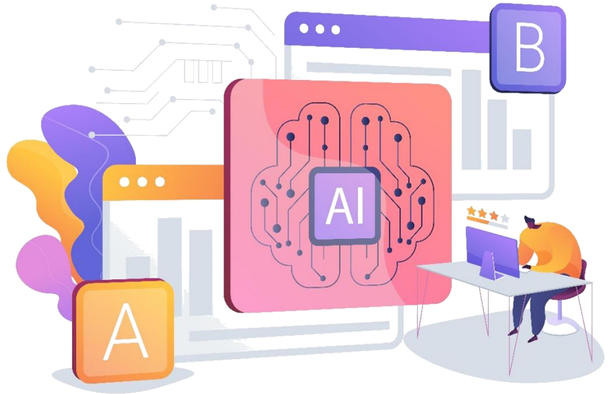
Deep Learning in Modern Speech Recognition



CNN + RNN + CTC

Hugging Face

Pretrained Models in Hugging Face

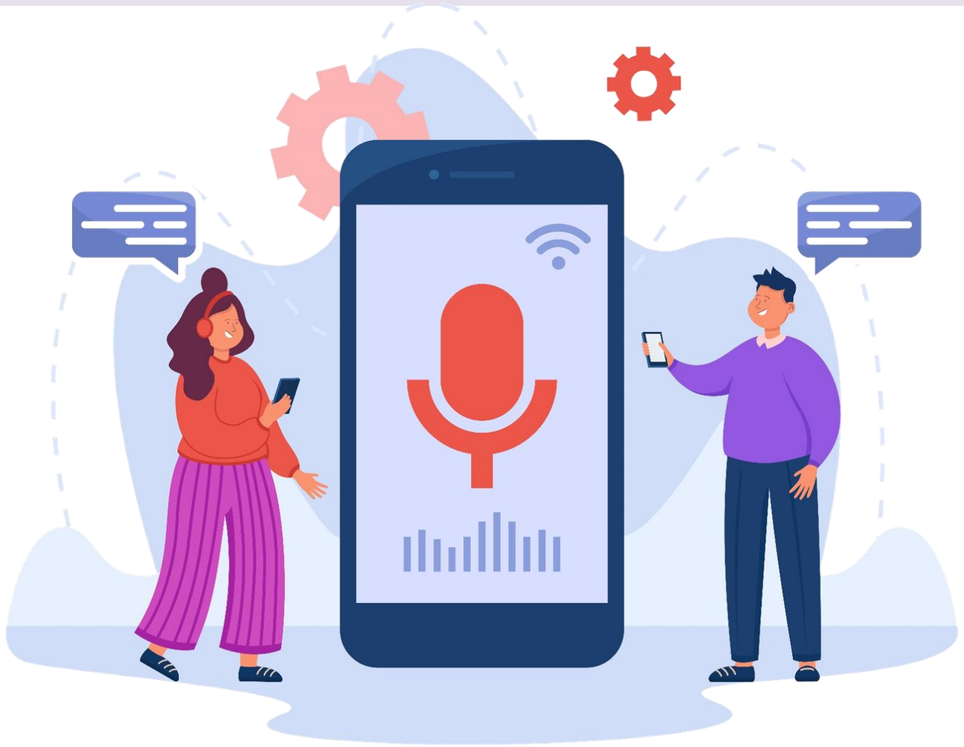


Transformer-Based Models

Automatic Speech Recognition (ASR) Models (e.g., Whisper, DeepSpeech)

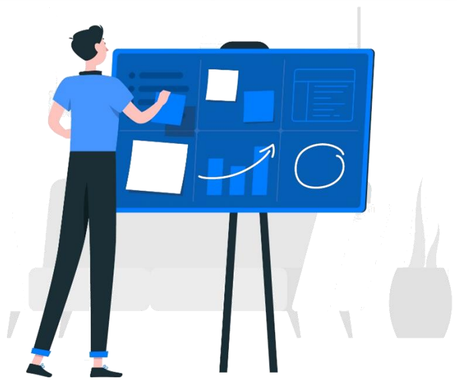
Automatic Speech Recognition (ASR) Models

ASR models are deep learning systems that convert spoken audio into written text by learning patterns in speech, sound, and language.



DeepSpeech – A Open-Source ASR Model

DeepSpeech is a speech recognition model by Mozilla based on RNNs with the CTC loss for aligning speech frames with words.



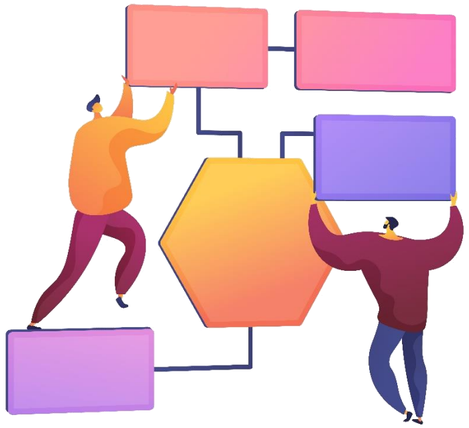
Model Architecture

Training Approach

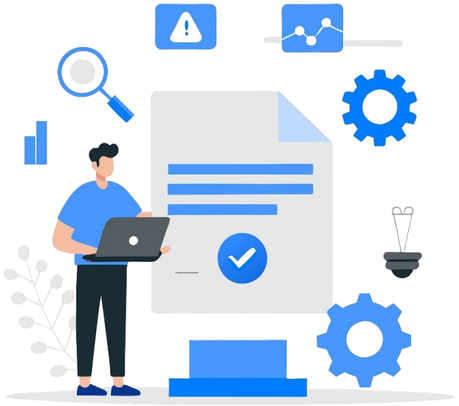
Limitations

Whisper – A Transformer-Based ASR Model

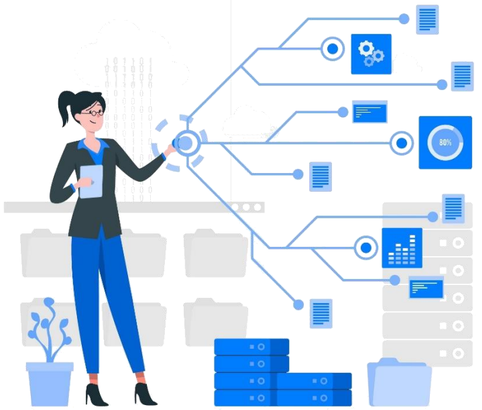
Whisper is a powerful, multilingual ASR model developed by OpenAI, trained on 680,000 hours of web audio using a sequence-to-sequence transformer.



Model Design



Key Features



Model Sizes

Comparison & Applications of ASR Models

Different ASR models serve different needs — from lightweight offline processing to accurate multilingual transcription in the cloud.

Feature	DeepSpeech	Whisper (OpenAI)
Architecture	RNN + CTC	Transformer (Seq2Seq)
Multilingual	English only	50+ languages
Noise Handling	Limited	Strong
Open Source	Yes	Yes
Extra Features	Basic transcription	Transcription + Translation

Challenges in Speech Processing

What Makes Speech Processing Difficult?

Speech processing involves decoding spoken language into meaningful text or actions. This process faces numerous challenges due to the unstructured, variable, and noisy nature of human speech.

Variability in Speech

Environmental Challenges

Real-Time Constraints

Language-Related Challenges

Code-Switching

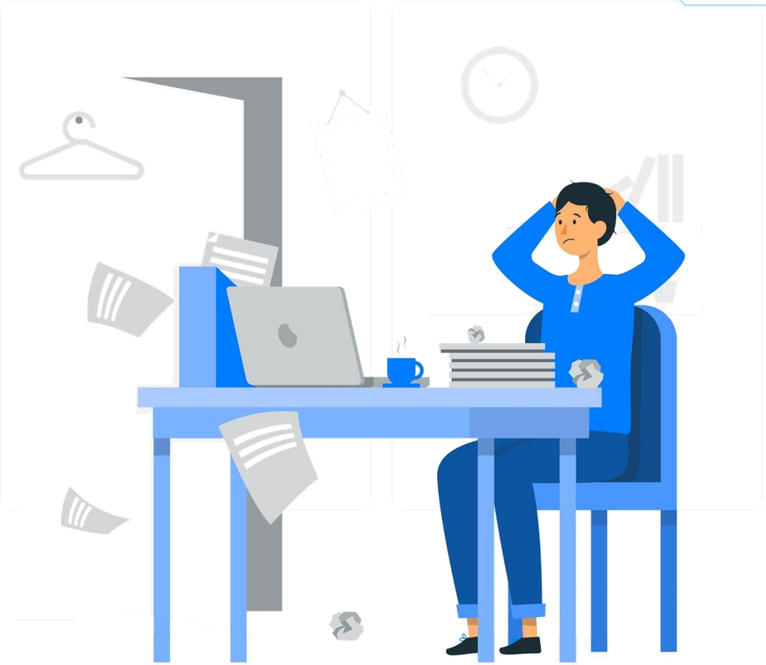
- Mixing languages within a sentence
- Common in multilingual regions (e.g., English + Hindi = Hinglish)

Low-Resource Languages

- Lack of labeled audio datasets
- No robust pretrained ASR models

Dialects, Accents & Regional Slang

- Same language spoken differently (e.g., Indian English vs. US English)
- Words like “innit” (UK) or “yaar” (India) often misrecognized



Technical & Model-Centric Challenges

Speech recognition models have limitations based on their architecture, training data, and decoding techniques — leading to errors in transcription and understanding.

Out-of-Vocabulary (OOV)

- Domain-specific terms (e.g., "photovoltaic", "BTS", "cringe")
- Newly coined or trending slang

Misalignment & Punctuation

- Speech is continuous; models struggle to infer pauses, sentence boundaries
- Lack of proper punctuation makes outputs unreadable

Resource Constraints

- Real-time speech apps must run on mobile/edge devices
- Large transformer-based ASR (like Whisper) needs high compute

TTS Applications in NLP

What is TTS in NLP?

Text-to-Speech (TTS) is the NLP task of converting written text into natural-sounding spoken audio using linguistic and acoustic models.

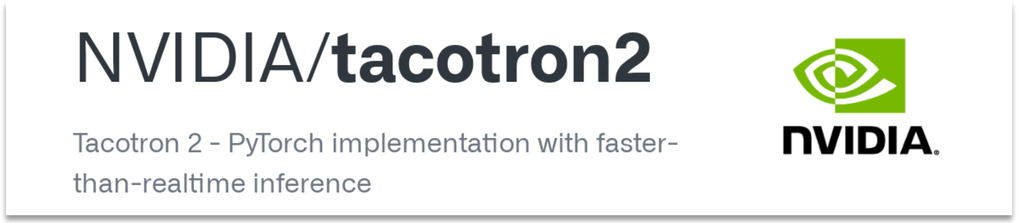
Purpose of TTS

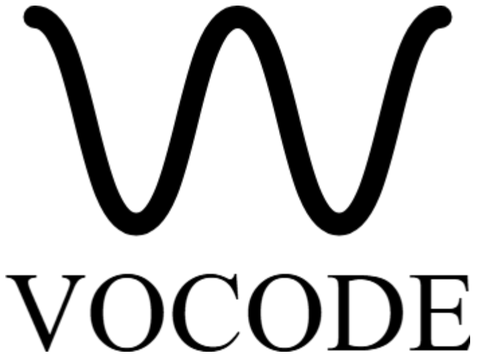
Components of a TTS System

Evolution of TTS

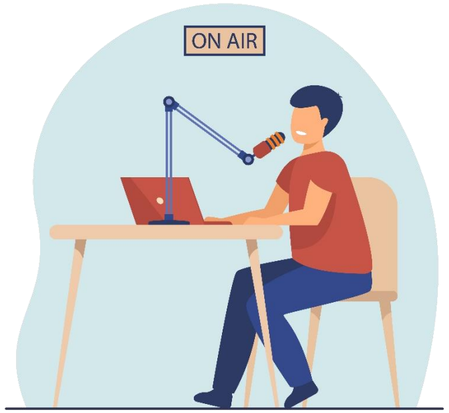
Neural TTS Models and Advancements

- Modern TTS systems use deep learning models to generate realistic, expressive speech by learning from large amounts of audio-text data.





Challenges and Future of TTS in NLP



Expressiveness & Emotion

Personalization

Future Directions

Multilingual & Code-Switching Support

Introduction to Multimodal Learning

What is Multimodal Learning?

Multimodal learning is an AI technique where a model learns from multiple data modalities — like text, images, and audio — simultaneously to improve understanding and performance on complex tasks.

What is a Modality?

- A specific type of data:
- Text (language, words)
- Image (pixels, objects)
- Audio (waveforms, speech, music)

Why Combine Modalities?

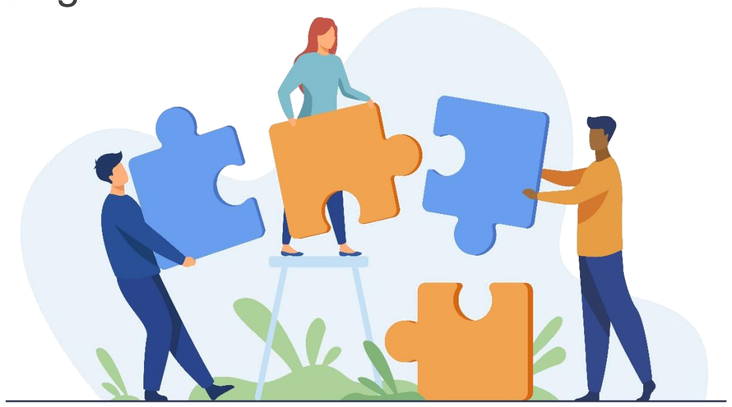
- Humans use multiple senses for understanding
- Fuses complementary signals
- Enables better context, reasoning, and prediction

Benefits of Multimodal Learning

Robustness to Missing Data

Contextual Understanding

Better Performance



Real-World Adaptability

Key Architectures & Techniques

- Early Fusion vs. Late Fusion
- Cross-Attention Mechanisms
- Multimodal Transformers
- Popular Frameworks

Vision-Language Models (CLIP, BLIP)

What Are Vision-Language Models (VLMs)?

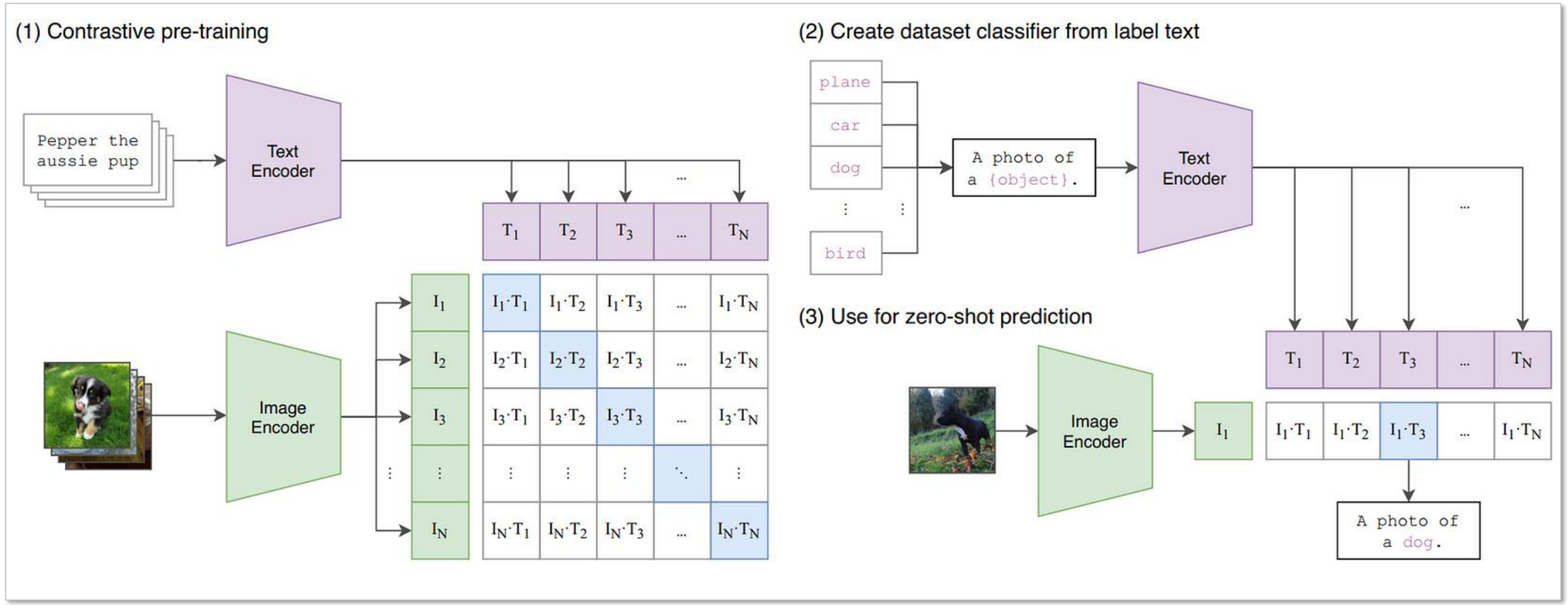
VLMs are deep learning systems designed to understand both images and text, and learn how they relate — enabling tasks like image captioning, visual question answering (VQA), and zero-shot image classification.

Input: An image of a cat wearing glasses

Model finds the best caption: “A stylish cat with sunglasses”

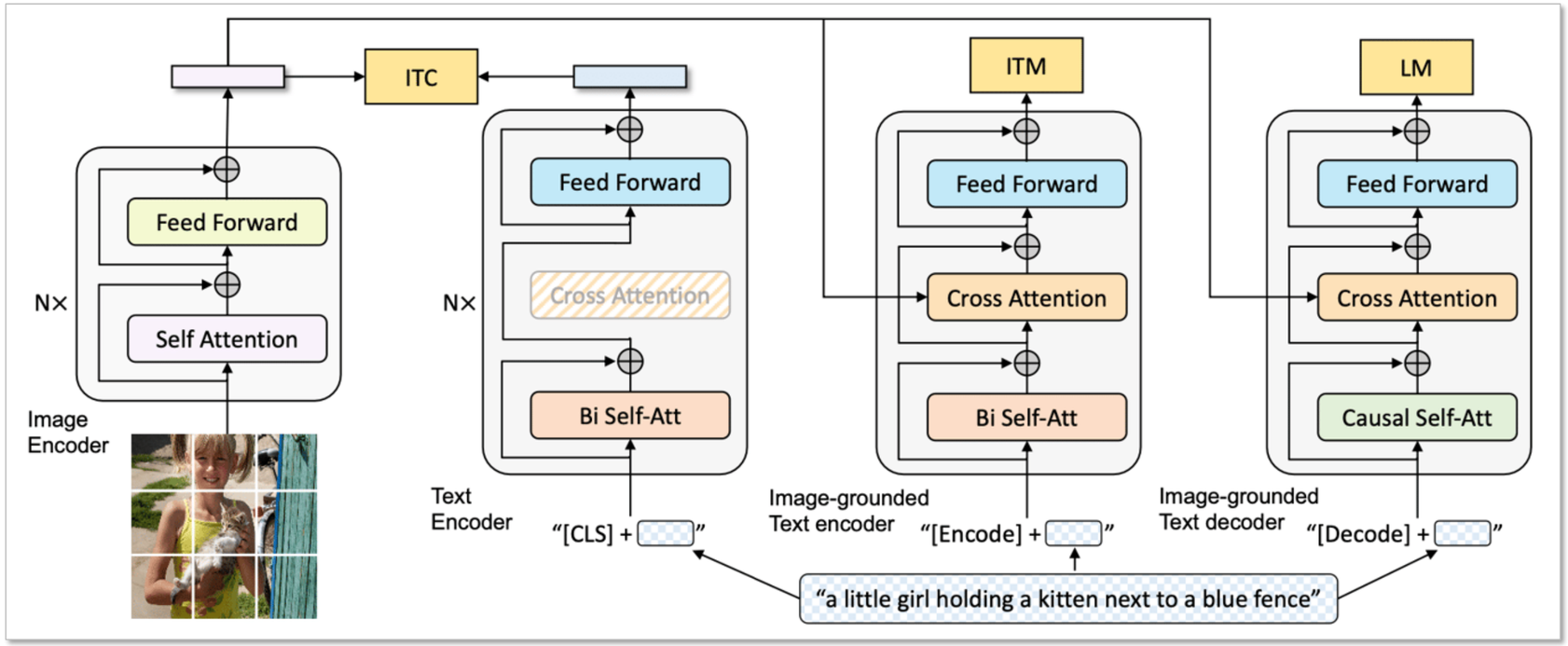
Contrastive Language-Image Pretraining

Contrastive Language-Image Pretraining (CLIP by OpenAI) is a vision-language model that learns to match images and their captions using contrastive learning.



Bootstrapped Language-Image Pretraining

Bootstrapped Language-Image Pretraining (BLIP by Salesforce AI) is a flexible VLM that supports image captioning, question answering, and image-text retrieval using both encoder-decoder and contrastive methods.



Integrating Text and Speech in NLP Applications

Integration of Text and Speech Mean

Integrating text and speech in NLP involves designing systems that can understand, process, and generate both spoken and written language, enabling fluid, human-like interactions.

Why Integrate Speech and Text?

- Combine the naturalness of voice with the precision of text
- Handle spoken commands, transcription, and voice-based outputs

Modalities Involved

- Speech to Text (ASR) → Converting audio to written form
- Text to Speech (TTS) → Generating spoken audio from text
- NLP Processing → Intent detection, question answering, dialogue systems

Core Components in a Text+Speech System

Automatic Speech Recognition

(ASR)

Natural Language Processing

(NLP Core)

Text-to-Speech

(TTS)

Generative Models for Multimodal Content

Generative Models for Multimodal Content

Generative multimodal models are AI systems capable of creating new content (e.g., text, images, audio, video) by learning the relationships between multiple data modalities like language, vision, and sound.

What is “Multimodal Generation”?

Creating content using more than one modality

- Text → Image (e.g., DALL·E)
- Image → Text (e.g., BLIP captioning)
- Text → Audio (e.g., TTS systems)

Core Generative Models and Their Capabilities

DALL·E / DALL·E 3

BLIP & BLIP-2

AudioLM / MusicLM

GPT-4o (Omni)

Applications of Multimodal Generative Models

- Creative Content Generation
- Education & Accessibility
- Social Media & Marketing
- Virtual Avatars & Agents

Model Compression (Quantization, Pruning, Knowledge Distillation)

Introduction to Model Compression

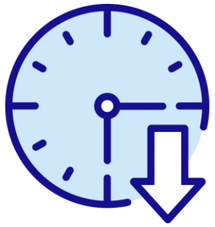
Model compression is a set of techniques used to reduce the size, memory footprint, and latency of deep learning models — while trying to retain most of their original performance.

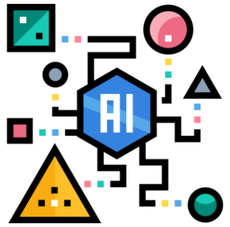
Why It's Important:

Lower Deployment Costs

Efficiency for Edge Devices

Reduced Inference Time





Enabling On-Device AI

Quantization

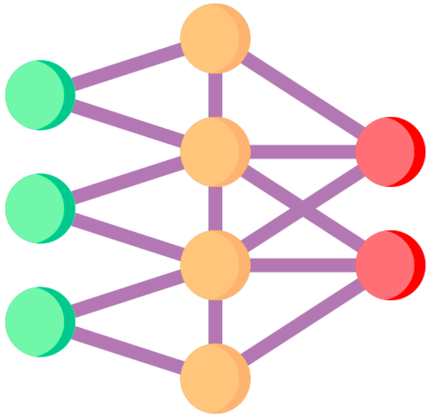
Quantization reduces model size and speeds up inference by converting high-precision (e.g., 32-bit float) weights and activations into lower-precision formats like 8-bit integers.

Methods:

Post-Training Quantization (PTQ) → Quantization-Aware Training (QAT) → Static vs. Dynamic Quantization

Pruning

Pruning removes unnecessary or low-impact weights, neurons, or attention heads from a model to make it smaller and faster.



Neuron/Head Pruning

Iterative vs One-Shot Pruning

Weight Pruning

Knowledge Distillation

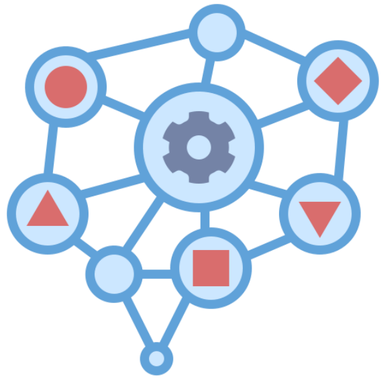
Knowledge distillation is the process of training a smaller “student” model to mimic a larger, well-trained “teacher” model.

Teacher Model

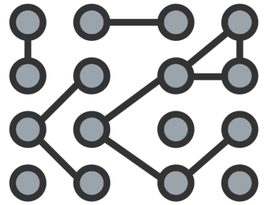
Student Model

Training Strategy

Types of Distillation



Teaching smaller model



Real-Time NLP for Edge Devices

What Is Real-Time NLP on Edge Devices?

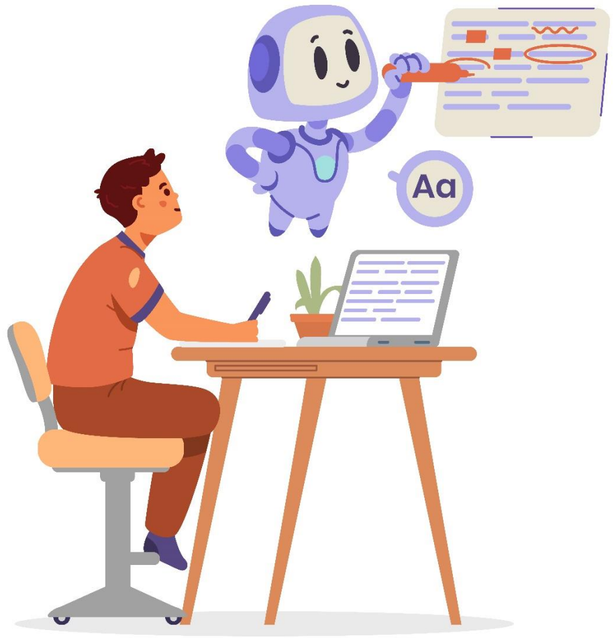
Real-time NLP for edge devices refers to performing natural language tasks (like speech recognition, translation, summarization) directly on local hardware (e.g., phones, wearables), without relying on cloud servers.

What Are Edge Devices?

- Smartphones, Raspberry Pi, wearables, home assistants, automotive systems

Why Use NLP on Edge?

- Privacy: Data stays on-device
- Latency: Instant responses, no network delay
- Offline Functionality: Works without internet



NLP Tasks Commonly Run on Edge Devices

Edge NLP enables basic to moderately complex language functions directly on devices using optimized or compressed models.

Wake Word Detection

Intent Detection

Text Classification

On-Device Translation / TTS / ASR

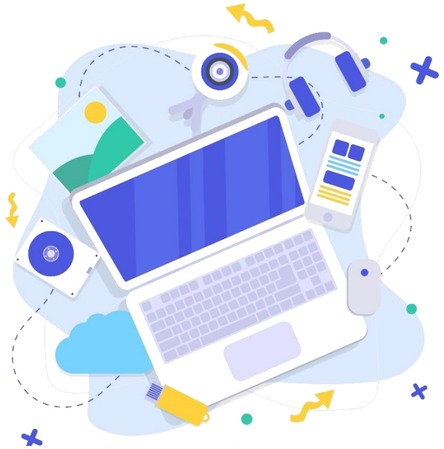
Techniques to Enable NLP on Edge

Deploying NLP on edge requires specialized model compression, optimization, and hardware-aware techniques.

Lightweight NLP Architectures

Model Compression

Hardware Acceleration



Frameworks & Tools

Explainability in Deep NLP Models

What is Explainability in NLP Models?

Explainability refers to the ability to understand, interpret, and trust the decisions made by deep NLP models, especially those using transformers and large neural networks.

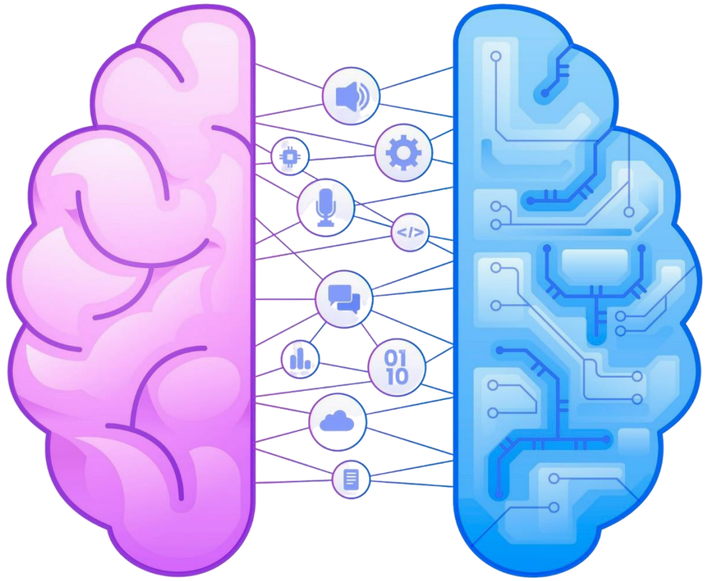
Why Explainability Matters:

Transparency in AI Decisions

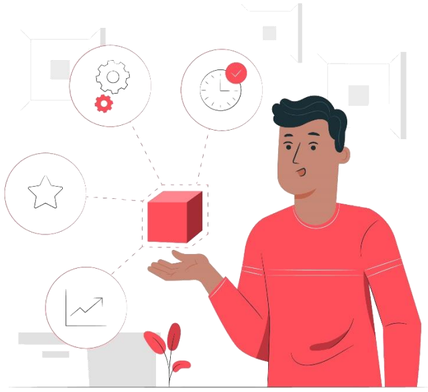
Debugging and Improvement

Regulatory Compliance

User Trust



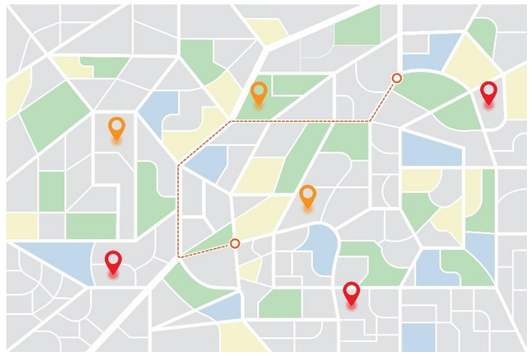
Methods for Explaining NLP Models



Feature Importance

Counterfactual Examples

Attention Visualization



Saliency Maps

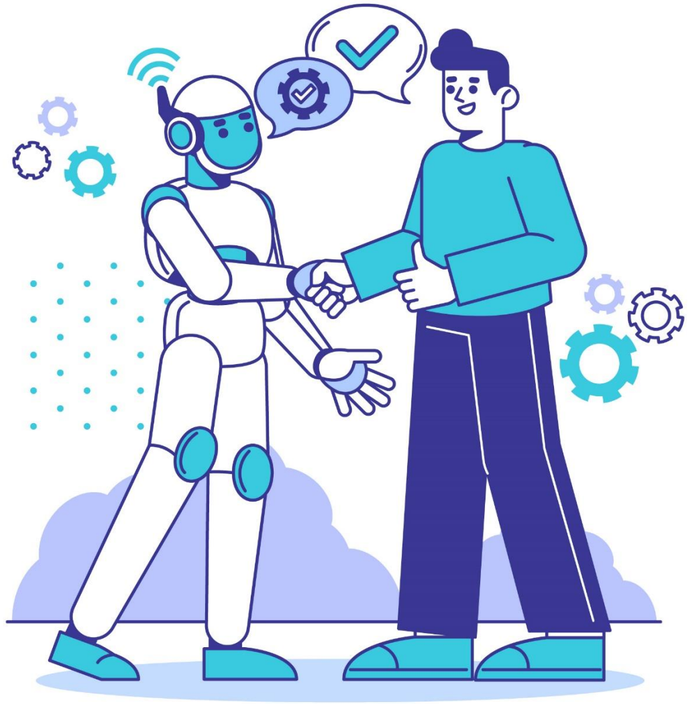
Tools and Libraries for Explainable NLP

ELI5

Ethical AI in NLP

What is Ethical AI in NLP?

Ethical AI in NLP involves designing and deploying natural language models that are fair, unbiased, accountable, transparent, and respect human rights.



Common Ethical Issues in NLP

Bias in Training Data

- Models trained on internet text often inherit stereotypes
- May associate professions with gender or race

Offensive or Harmful Outputs

- Chatbots may produce toxic responses if not filtered
- Hate speech detection can be too aggressive or too lenient

Privacy Violations

- Language models may memorize and leak sensitive data (e.g., phone numbers)

Misinformation and Hallucination

- Generative models can confidently produce incorrect or fake facts

Future of Ethical AI in NLP

Transparent Benchmarking

Ethical AI by Design

Cross-Disciplinary Teams

Global Standards and Regulations

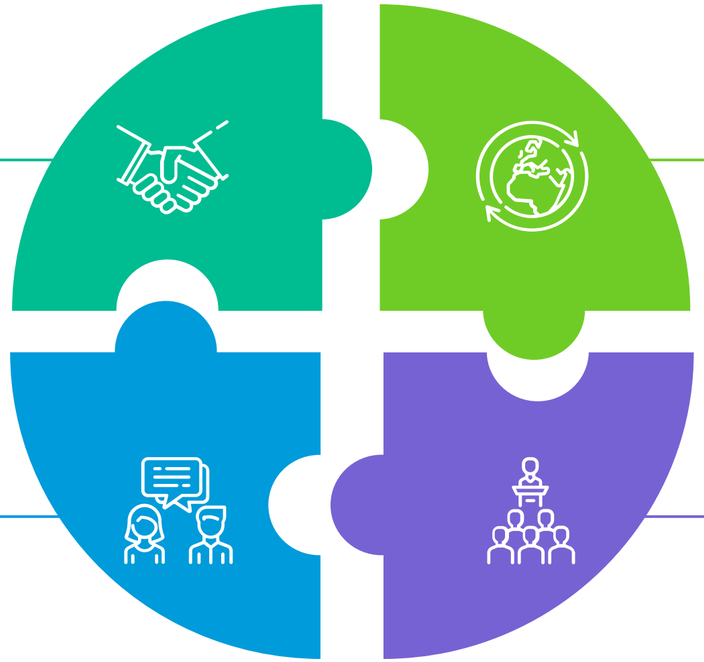
Emerging Trends (GPT-5, HyperNetworks, Autonomous AI Agents)

What’s Driving Emerging Trends in NLP?

Emerging trends in NLP focus on building more intelligent, adaptive, and autonomous systems using advancements in foundation models, network architectures, and agent-based AI.

Scale and Generalization

Efficiency and Modularity

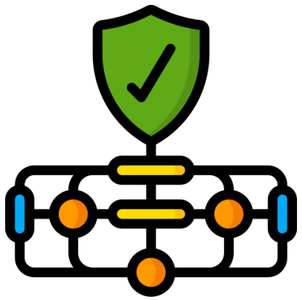


Cross-Modality Integration

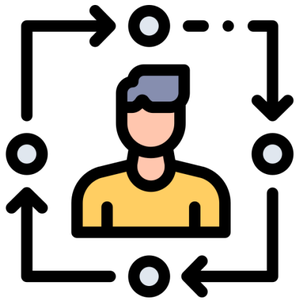
Real-World Task Automation

GPT-5- The Next Gen of Foundation Models

GPT-5 (upcoming) represents the next evolution in large language models, expected to be more multimodal, grounded, and autonomous.

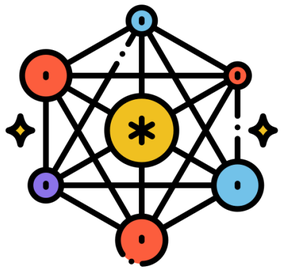


Safer, More Controlled Generation



Long-Term Memory & Personalization

Better Reasoning & Planning



Deeper Multimodal Capabilities

HyperNetworks – Smarter Parameter Sharing

HyperNetworks are neural networks that generate the weights for another network, enabling faster training and task adaptability with fewer parameters.

Benefits & Use Cases:

1. Efficient Fine-Tuning
2. Modular & Transferable
3. Lower Training Cost
4. Supports Continual Learning

Rise of Autonomous AI Agents

Autonomous AI agents are systems that can perceive, plan, and act in real-world or digital environments, often powered by large LMs + tools.

Capabilities & Components:

Self-Critique and Planning

Goal-Driven Behavior

Memory and Learning

Tool Usage

Building a Voice Narration QA Assistant for Audiobooks (Demonstration)

Note: Refer to Module 9: Demo 1 on LMS for detailed steps.

Autonomous Warehouse Object Classification via Text–Image Fusion (Demonstration)

Note: Refer to Module 9: Demo 2 on LMS for detailed steps.

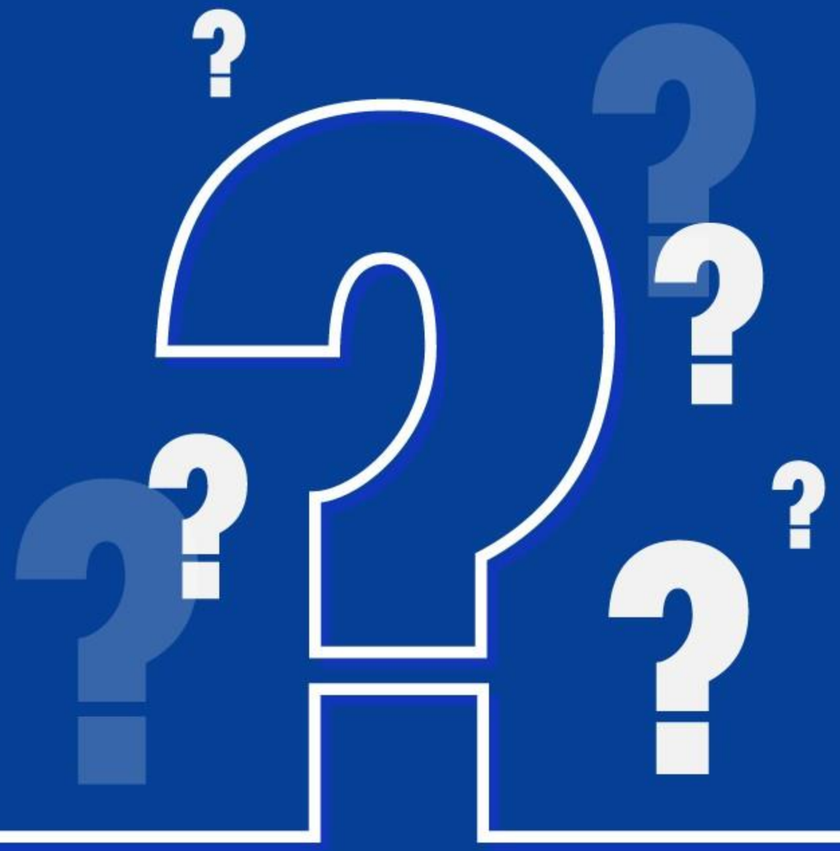
Summary

In this lesson, you have learned to:

- Explain how speech and text are processed in NLP
- Perform comparison between ASR models
- Understand challenges in speech processing and real-time NLP for edge deployment
- Apply multimodal and generative models to integrate and generate text, audio, and images
- Evaluate the use of model compression, explainability, and ethical principles in NLP systems
- Summarize emerging trends in NLP such as GPT-5, HyperNetworks, and autonomous AI agents

edureka! @ Veranda ENTERPRISE

Questions



edureka! a Veranda ENTERPRISE

Feedback

edureka! a Veranda ENTERPRISE

Thank You

For information, Please Visit our Website
www.edureka.co