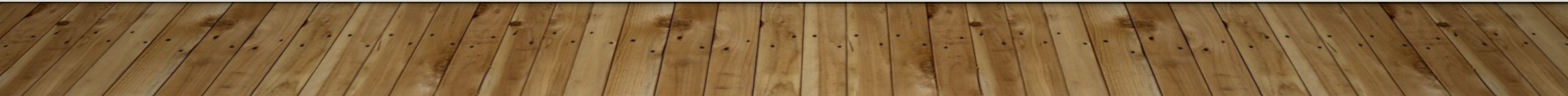# FOUNDATIONAL CONCEPTS & INDUSTRIAL USE CASES

- DEV SOURAV PANDA

Introduction to NLP and Its Industrial Importance

Core NLP Techniques and Applications

NLP for Social Media and Sentiment Analysis

Emerging Trends and Innovations

QA / Feedback

ACTIVITY TIME
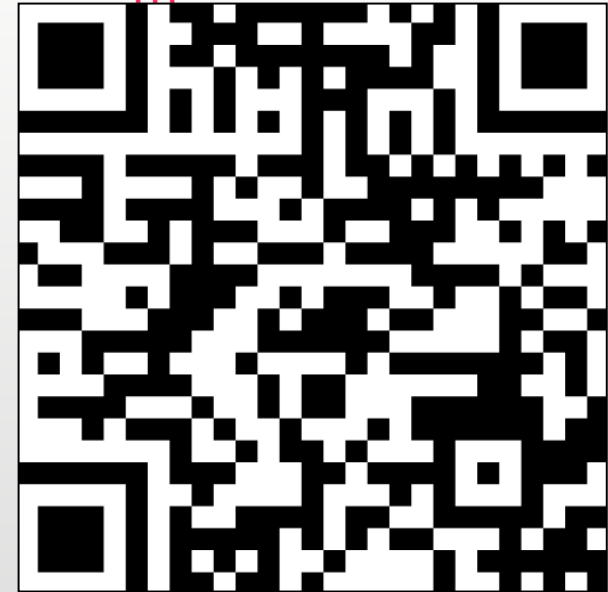
PART I

Introduction to NLP and Its Industrial Importance
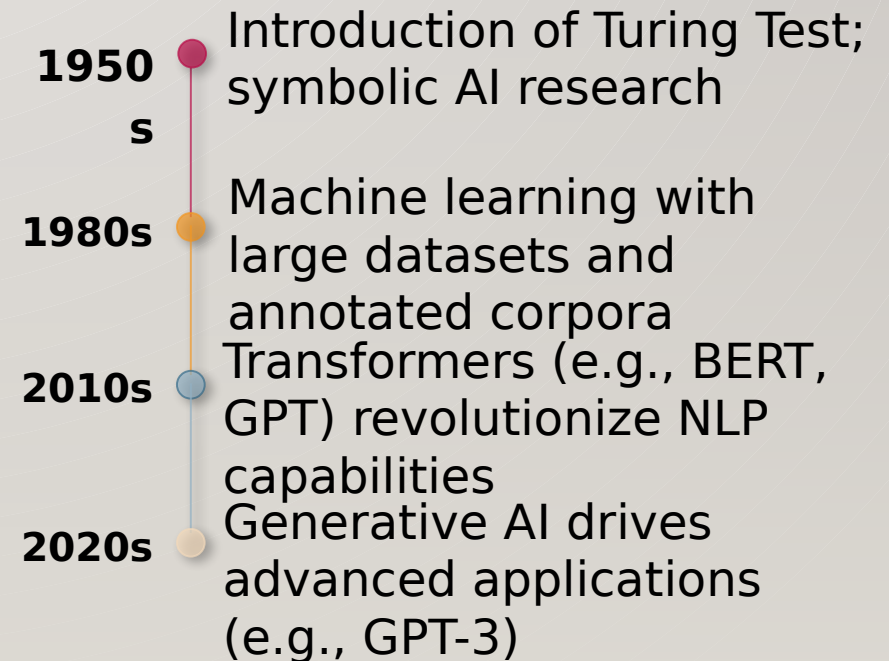
Code: 2725 6137

DEFINITION AND HISTORY OF NLP

- **Definition:** A branch of Artificial Intelligence focusing on interactions between computers and human languages.

- **Objective:** Enables understanding, interpretation, and generation of human language.

- **Applications:** Language translation, sentiment analysis, chatbots, etc.

1950s — Introduction of Turing Test; symbolic AI research

1980s — Machine learning with large datasets and annotated corpora

2010s — Transformers (e.g., BERT, GPT) revolutionize NLP capabilities

2020s — Generative AI drives advanced applications (e.g., GPT-3)

TRANSFORMER REVOLUTION IN NLP

- Introduced in 2017 with "*Attention Is All You Need*", enabling parallel processing.

- Uses self-attention to focus on relevant sequence parts dynamically.

- Faster and more scalable than RNNs, handling large datasets efficiently.

- Foundation for models like BERT, GPT, and T5, revolutionizing NLP tasks.

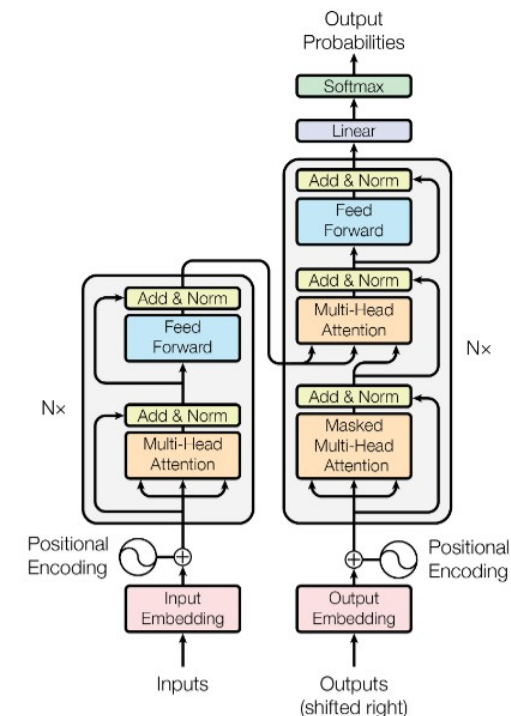- Set new benchmarks in language understanding and generation.

Figure 1: The Transformer - model architecture.

INDUSTRIAL APPLICATIONS OF NLP

STANDARD APPLICATIONS

- Chatbots and Virtual Assistants
- Sentiment Analysis
- Machine Translation
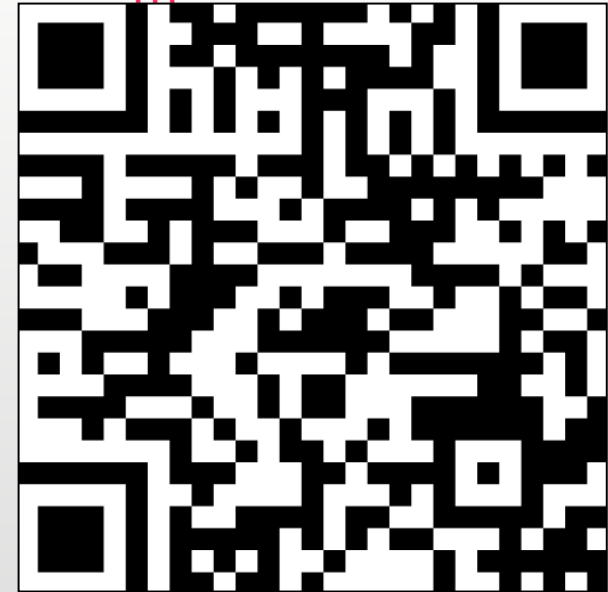- Speech Recognition
- Text Summarization

ADVANCED APPLICATIONS

- AI in Healthcare Diagnostics
- Automated Content Creation
- Intelligent Legal Document Processing
- Personalized Marketing and Advertising
- Financial Report Analysis and Forecasting

ACTIVITY TIME

Guess the NLP Application?

Code: 2725 6137
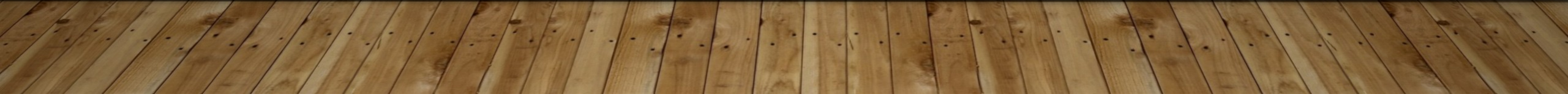
PART II

Core NLP Techniques and Challenges
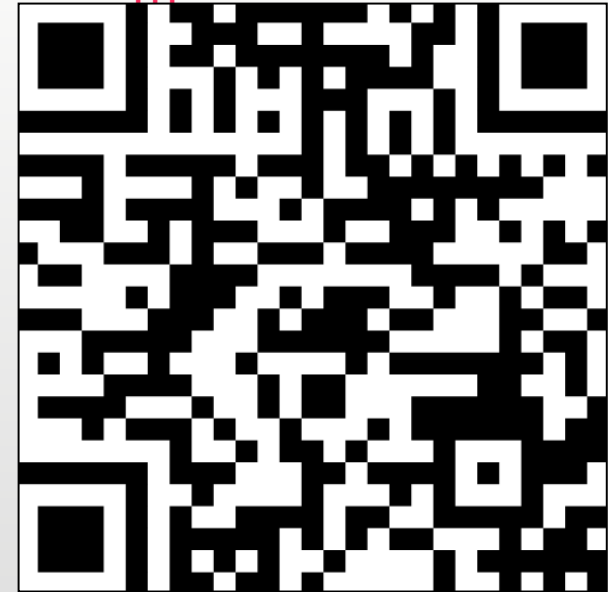
Code: 2725 6137

DATA ACQUISITION

- **Multimodal data acquisition** for training advanced generative AI models across diverse modalities.

- **Data availability spectrum:** Extensive public datasets vs. limited domain-specific corpora.

- **Data collection challenges:** Accessibility, privacy compliance, and data heterogeneity.

- **Mitigation strategies:** Data augmentation, synthetic data generation, and transfer learning.

- **Multimodal fusion** enhances model robustness, generalization, and generative capabilities.

DATA EXTRACTION

- **Text Extraction**: Web pages, documents, and databases.

- **Multimedia Extraction**: Features from images, audio, and video via OCR and speech recognition.

- **Structured Data Extraction**: Tables, spreadsheets, and databases.

- **Data Provenance**: Tracking data origin, transformations, and compliance.

- **Cross-Domain Extraction**: Extracting from specialized sources like legal or medical texts.

DATA PREPARATION

- **Text Normalization**: Standardizing text (e.g., lowercasing, removing special characters)

- **Tokenization**: Splitting text into words or sub-words for further analysis.

- **Stop-word Removal**: Removing common words (e.g., "the", "is") that don't add value to analysis, typically used in text classification and topic modeling.

- **Named Entity Recognition (NER)**: Identifying and categorizing entities like names, dates, and locations.

- **Text Segmentation**: Dividing large documents into meaningful chunks, used in tasks like document classification or summarization, where context within sections matters.

WORD EMBEDDINGS

- **Vector Representation**: Converting words into numeric vectors to capture semantic meaning.

- **Dimensionality Reduction**: Reducing the size of word vectors to optimize models.

- **Pre-trained Embeddings**: Using pre-trained models like Word2Vec, GloVe, or FastText to save time on training and improve performance in downstream tasks.

- **Contextual Embeddings**: Word representations change based on context (e.g., BERT), enhancing performance in complex tasks like question answering or translation.

- **Transfer Learning for Embeddings**: Fine-tuning pre-trained embeddings for specific tasks, useful in domain-specific applications like legal or medical NLP.

MODEL TRAINING

- **Model Architecture Selection**: Choosing the right model (e.g., RNN, CNN, Transformer) based on the task and data.

- **Supervised Learning**: Training the model with labeled data to learn patterns and make predictions.

- **Hyperparameter Tuning**: Optimizing learning rate, batch size, and model depth to enhance model performance.

- **Transfer Learning**: Leveraging pre-trained models and fine-tuning them for specific tasks to save time and resources.

- **Cross-Validation**: Using techniques like k-fold cross-validation to ensure the model's robustness and generalization across different datasets.

MODEL EVALUATION

- **Accuracy**: Measures the percentage of correct predictions made by the model compared to the total number of predictions.

- **Precision & Recall**: Precision evaluates the correctness of positive predictions, while recall measures the model's ability to identify all relevant instances.

- **Confusion Matrix**: Provides a comprehensive overview of the model's performance by showing true positives, true negatives, false positives, and false negatives.

- **F1-Score**: The harmonic mean of precision and recall, offering a balanced measure of the model's performance, especially in imbalanced datasets.

- **ROC-AUC Curve**: Evaluates model performance across different classification thresholds, providing a clear view of the model's ability to distinguish between classes.

MODEL DEPLOYMENT

- **Deployment Readiness**: If evaluation metrics are justified, the model is deployed through APIs, applications, or cloud-based platforms. Else it retrains.

- **Model Integration**: Integrating the model into existing systems or platforms for real-time predictions and decision-making.

- **Scalability**: Ensuring the model can handle large volumes of requests and data seamlessly during deployment.

- **Real-time Monitoring**: Continuously tracking model performance in production to ensure it delivers accurate predictions.

- **Model Maintenance**: Updating the model regularly with new data and retraining to address data drift and evolving business needs.

CHALLENGES IN DEPLOYING NLP SYSTEMS

- Data Quality and Availability

- Real-time Performance

- Scalability

- Model Interpretability

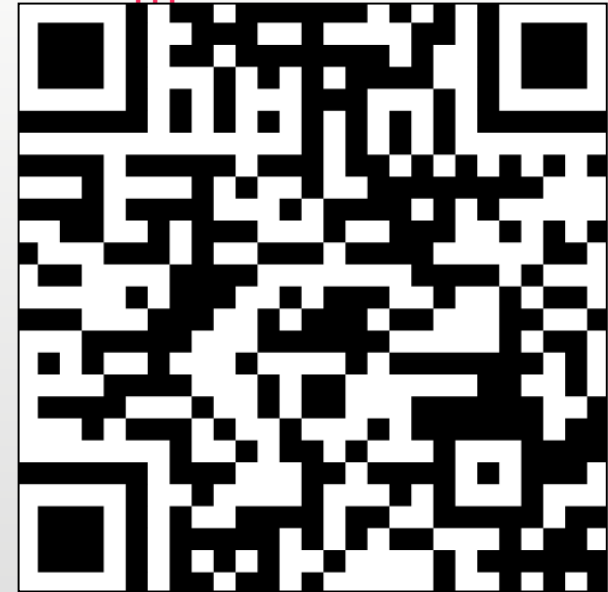- Integration with Existing Systems

TRANSFORMERS AND SELF-ATTENTION

Parallel Processing

Self-Attention Mechanism

Contextual Understanding

Positional Encoding

Scalability
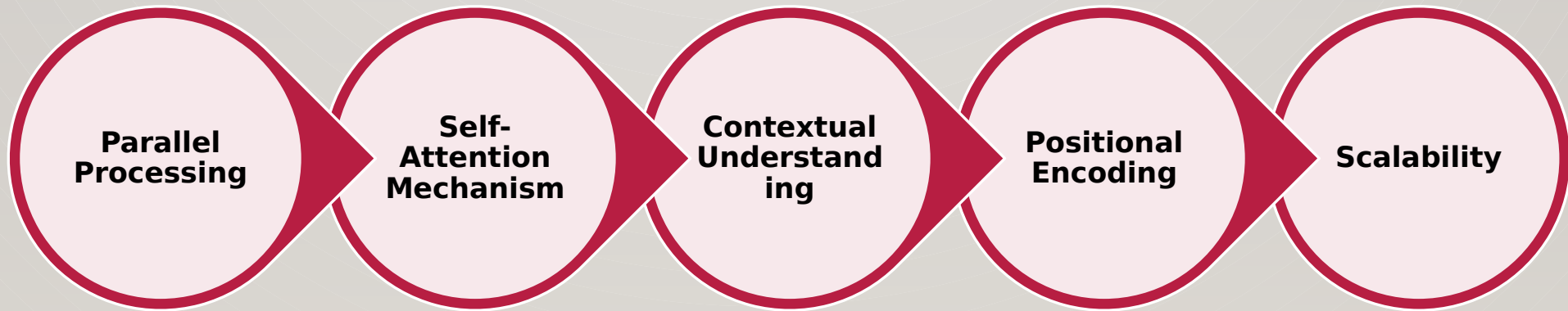
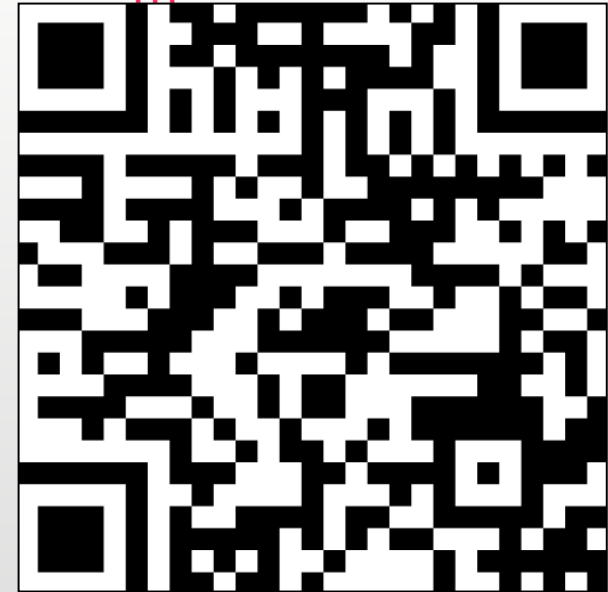# ACTIVITY TIME

Quiz on core NLP techniques

www.menti.com

Code: 2725 6137

PART III

NLP for Social Media and Sentiment Analysis

Code: 2725 6137

INTRODUCTION TO SENTIMENT ANALYSIS

- **Purpose:** Identifies the emotional tone (positive, negative, or neutral) in text.

- **Techniques:** Uses machine learning, deep learning, and lexicon-based methods.

- **Popular Tools:** *VADER* and *TextBlob*

- **Applications:** Common in customer feedback, social media monitoring, and market research.

- **Challenges:** Handling sarcasm, context, and multilingual data.

- **Outcome:** Provides insights into public opinion, brand reputation, and customer satisfaction.

CASE STUDY: SOCIAL MEDIA SENTIMENT TRACKING

- **Building on Sentiment Analysis:** Extends the principles of sentiment detection to social media platforms.

- **Real-Time Insights:** Monitors public opinion trends, enabling immediate responses to feedback or crises.

- **Coca-Cola Example:** Used sentiment tracking to evaluate campaign success and protect brand reputation.

- **Impact on Strategy:** Helps refine marketing strategies, improve customer engagement, and enhance overall brand perception.
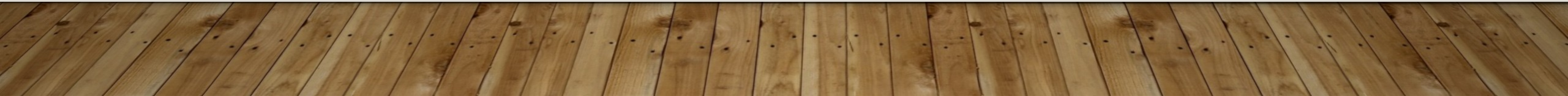
SENTIMENT TOOLS IN ACTION

STANDARD MODELS

- **VADER:** Great for social media sentiment analysis.

- **TextBlob:** Simple and easy-to-use for sentiment scoring.
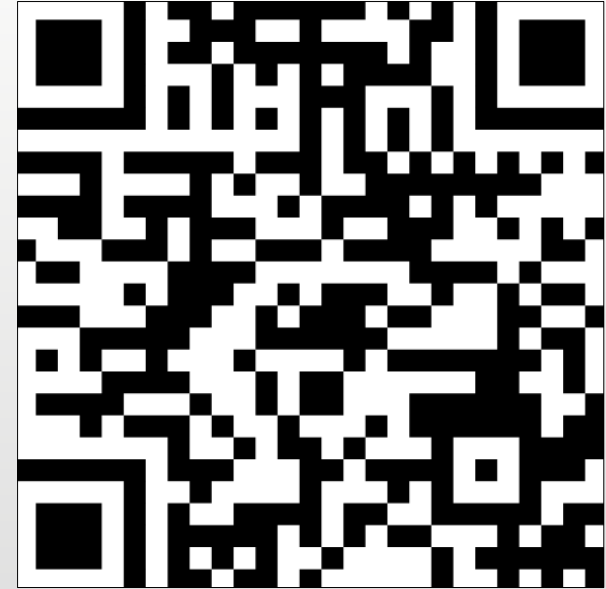
ADVANCED MODELS

- **BERT-Based Models:** Understands context and nuances like sarcasm and ambiguity.

- **Google Cloud NLP API:** Scalable sentiment analysis via cloud service.

- **Custom ML Pipelines:** Tailored sentiment models for specific industries.

ACTIVITY TIME

Classify sample tweets into positive, neutral, or negative sentiments using Python (Hands-on)
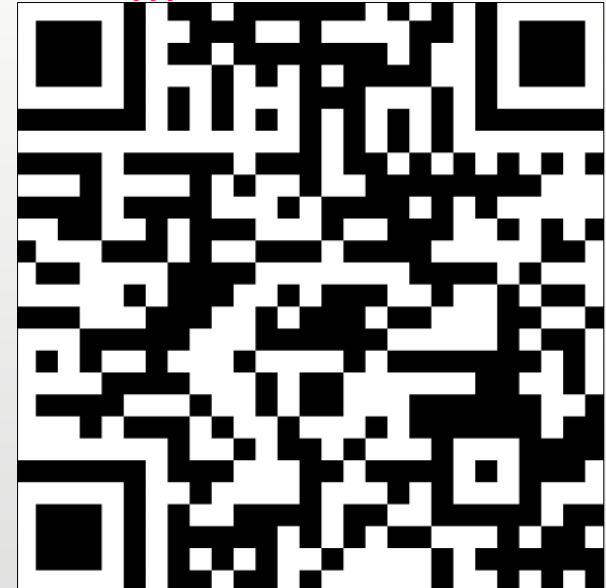
www.menti.com

Code: 2725 6137

PART IV

Emerging Trends and Innovations

www.menti.com

Code: 2725 6137

ROADMAP TO ADVANCED NLP CONCEPTS

- Multimodal NLP

- Cross-Language Models

- Zero-Shot/Few-Shot Learning

- **Generative AI and RAG**

- Explainability and Interpretability

- Ethics and Fairness

- Personalized NLP

COLLABORATION OPPORTUNITIES BETWEEN ACADEMIA AND INDUSTRY

Research & Development Partnerships

Collaborating on cutting-edge research to drive innovation in NLP, enabling academia to work on real-world problems and industry to access novel solutions

Internships and Industry Projects

Providing students with opportunities to work on live industry projects, gaining practical experience while contributing fresh insights to companies
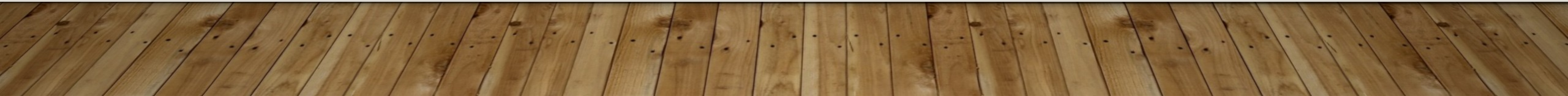
Knowledge Sharing & Training

Facilitating workshops, webinars, and joint conferences to bridge the gap between theoretical advancements in academia and practical applications in the industry.

OPEN DISCUSSION

INDUSTRY-ACADEMIA RELATIONSHIPS

Thank You
For your Attention