



ADVANCED NLP

RAG, GENERATIVE AI, AND USE CASE DEMO



Advanced NLP
Systems



Retrieval-
Augmented
Generation (RAG)



Generative AI



Major Project
Demo



QA / Feedback


ACTIVITY TIME

Join at menti.com | use code 7136 1026

Mentimeter

Instructions

Go to
www.menti.com
Enter the code
7136 1026

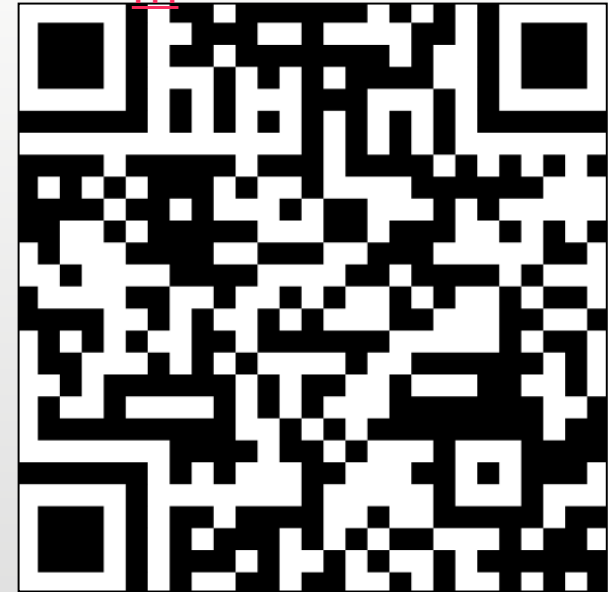


Or use QR code

PART I

Advanced NLP Systems: Deep Dive

[www.menti.co
m](https://www.menti.com)



Code: 7136
1026

ROLE OF TRANSFORMERS IN MODERN NLP

- **Transformers** replaced RNNs and LSTMs as the backbone of NLP.
- **Self-attention mechanism** enables models to weigh words based on context.
- Breakthrough models like **BERT, GPT, and T5** showcase massive scalability.
- Enabled NLP tasks with unparalleled accuracy across **translation, QA, and summarization**.
- **Real-world impact:** GPT-like models generating human-like responses for chatbots, search engines, and content creation.

FINE-TUNING VS. PRE-TRAINING

- **Pre-training** trains a model on massive corpora for general knowledge.
- **Fine-tuning** adapts pre-trained models for domain-specific tasks.
- **Industrial importance:** Reduces training costs and time while improving accuracy.
- Fine-tuned models excel in **legal, healthcare, and sentiment analysis** tasks.
- Example: **Fine-tuned BERT** for legal text summarization or financial analytics.

LARGE LANGUAGE MODELS (LLM)

- LLMs (e.g., GPT-4, PaLM) process billions of parameters for nuanced text generation.
- **Use cases:** Conversational AI, document analysis, creative content generation.
- **Deployment challenge:** Requires massive compute power (e.g., GPUs/TPUs).
- **Ethical concerns:** Biases in training data propagate into real-world outputs.
- **Scaling trends:** OpenAI, Google, and Anthropic are pushing LLM boundaries.

MULTIMODAL TRANSFORMERS

- **Multimodal models** process text, images, audio, and video in a unified framework.
- Advances include **CLIP** (vision-text understanding) and **DALL·E** (text-to-image).
- **Applications:** Automatic video captioning, text-based image generation.
- Transforming **healthcare** (scanning analysis + reports) and **e-commerce** (visual search).
- Emerging trend: **Fusion architectures** for robust cross-modal understanding.

INDUSTRY APPLICATIONS OF LLM

- **Healthcare:** Medical report generation, symptom diagnosis with natural dialogue.
- **Legal tech:** Summarizing legal documents and automating contract analysis.
- **Finance:** Risk analysis, fraud detection using advanced NLP insights.
- **Customer support:** Virtual assistants resolving customer queries with context.
- **Content creation:** Automating marketing copy, blogs, and creative design.

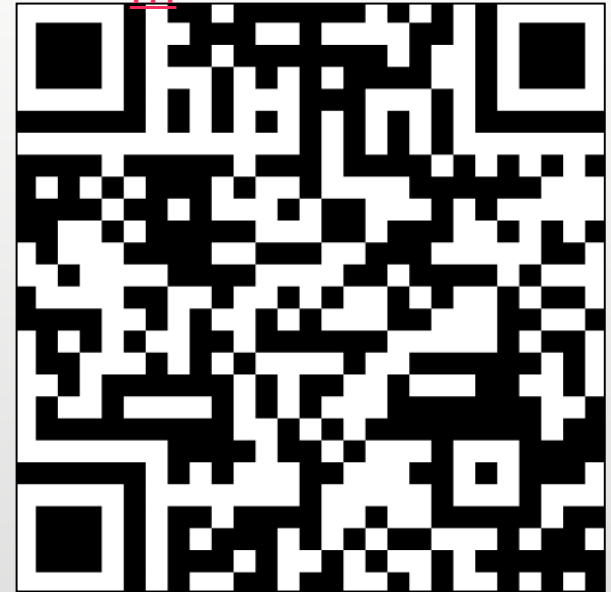
CHALLENGES IN DEPLOYING LLM

- **High computational costs** due to billions of parameters and real-time demands.
- **Scalability issues:** Infrastructure bottlenecks while handling large datasets.
- **Ethical concerns:** Bias, misinformation, and lack of interpretability.
- **Data privacy:** Compliance issues in sensitive domains like healthcare and law.
- **Solutions:** Model pruning, quantization, and hybrid deployment strategies.

ACTIVITY TIME

Test your Understanding!

[www.menti.co](https://www.menti.com)
[m](https://www.menti.com)

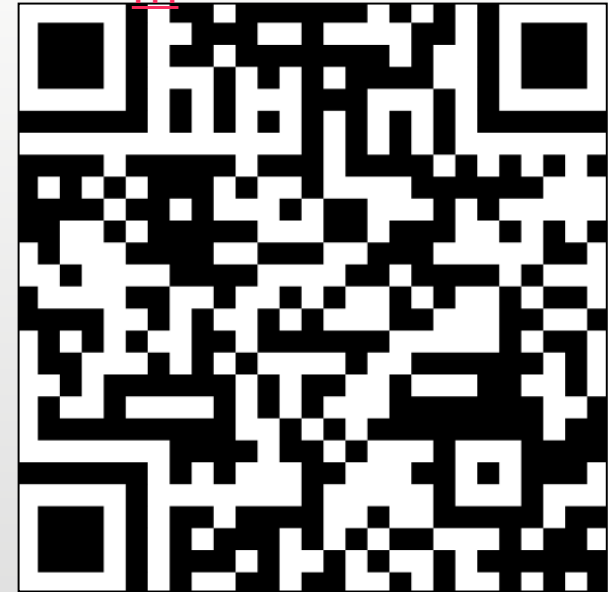


Code: 7136
1026

PART II

Retrieval-Augmented Generation (RAG)

[www.menti.co](https://www.menti.com)
[m](https://www.menti.com)



Code: 7136
1026

INTRODUCTION TO RAG

- Combines **retrieval** with generative AI models.
- Retriever-generator architecture for contextual responses.
- Real-time use of **external data** for accurate generation.
- Reduces hallucinations and enhances factual accuracy.

RAG VS. TRADITIONAL RETRIEVAL

- **Seamless integration** of retrieval and generation workflows.
- Traditional systems rely on search + manual interpretation.
- RAG enables **dynamic, real-time knowledge access**.
- Cost-efficient and adaptable for **domain-specific tasks**.



EMBEDDINGS IN RAG

- Embeddings capture semantic similarity over keywords.
- Efficient document retrieval using vector databases (FAISS, Pinecone).
- High-dimensional vector space for precise context matching.
- Enhances accuracy of generated responses with relevant context.



CASE STUDIES IN INDUSTRIES

- **Knowledge assistants** for real-time customer support.
- **Legal tech**: Rapid document scanning and retrieval.
- **Research platforms**: Instant access to scientific literature.
- Personalized **content generation**.
- Real-time **FAQ systems** for enterprise-level tasks.

TOOLS AND FRAMEWORKS

- **Haystack:** End-to-end RAG implementation framework.
- **OpenAI Retrieval Plugin:** Integrates retrieval with GPT systems.
- **Pinecone, Weaviate:** Scalable vector databases for embeddings.
- Simplifies building **production-ready RAG pipelines.**

SCALABILITY AND CHALLENGES

- Scaling retrieval for **large document corpora**.
- Computational costs and latency in real-time systems.
- Optimized indexing and **hybrid search** for scalability.
- Industry examples: Overcoming production bottlenecks.



RAG-BASED WORKFLOWS

- **Pipeline architecture:** Retrieval to generation integration.
- Automating updates for real-time data synchronization.
- Workflow optimization for **speed and accuracy**.
- Real-world production-ready RAG system design.



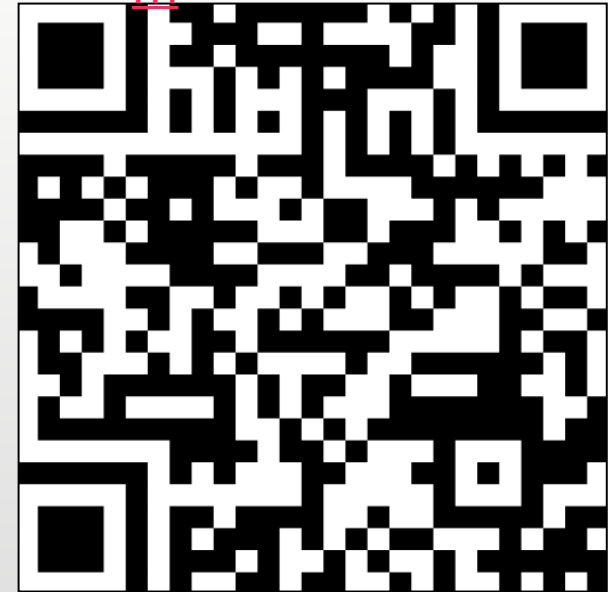
EMERGING TRENDS IN RAG

- **Hybrid RAG:** Symbolic search + deep learning models.
- RAG for **multimodal data:** Text, images, videos.
- **Real-time adaptive retrieval** systems for dynamic queries.
- Advances in **low-latency vector search** and scalability.

ACTIVITY TIME

Brainstorm and present RAG use cases

[www.menti.co](https://www.menti.com)
m

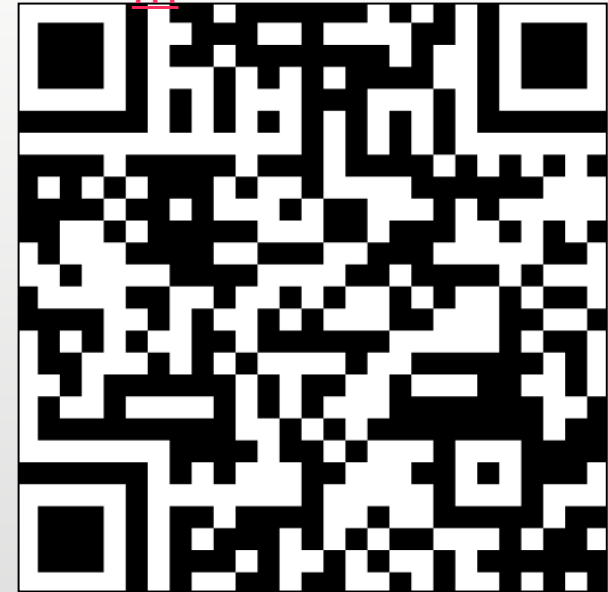


Code: 7136
1026

PART III

Generative AI: Concepts and Applications

[www.menti.co](https://www.menti.com)
m



Code: 7136
1026

INTRODUCTION TO GENERATIVE AI

- Generative AI produces new data resembling training inputs.
- **Key components:** Large language models (LLMs), diffusion models.
- **Transforming industries:** From text generation to multimedia content.
- **Real-world impact:** Efficiency, creativity, personalization.
- **Examples:** GPT models, DALL·E, Stable Diffusion.

TRADITIONAL VS. GENERATIVE SYSTEMS

TRADITIONAL SYSTEMS

- Predictive, classify or label existing data
- Supervised learning, often with labeled data
- Task-specific, limited adaptability

GENERATIVE SYSTEMS

- Produce new, creative data
Unsupervised or self-supervised learning, leveraging large unstructured datasets
- Highly flexible, capable of tackling a wide range of tasks (e.g., text generation, image creation)

DIFFUSION MODELS AND TRANSFORMERS

- **Diffusion models:** Noise-based image generation (Stable Diffusion).
- **Transformer architecture:** Foundation of GPT, BERT, and more.
- **Role of attention mechanism:** "Attention Is All You Need" paper.
- **Diffusion workflow:** Gradual noise removal → High-quality outputs.
- **Latest advancements:** Combined architectures for multimodal outputs.

APPLICATIONS IN INDUSTRY

- **Content creation:** Automated blogs, social media, video scripts.
- **Marketing:** Personalized campaigns, ad creatives, SEO content.
- **Design:** Image generation for fashion, architecture, e-commerce.
- **Healthcare:** Synthetic medical data, medical image generation.
- **Research:** Generating hypotheses, summarizing papers.

ETHICAL CONCERNS

- **Bias in training data:** Reinforcing stereotypes, misinformation.
- **Misuse of AI-generated content:** Deepfakes, fake news, plagiarism concerns.
- **Safeguards and regulations:** Model audits, regulatory frameworks, responsible AI.
- **Data Privacy Concerns:** AI models training on sensitive data, raising privacy issues.
- **Accountability and Transparency:** Opaque AI decision-making processes.
- **Industry initiatives:** OpenAI guidelines, ethical model training.



CASE STUDY: AI FOR CONTENT CREATION

- **E-commerce:** Automating product descriptions.
- **Social media:** Generating targeted, high-performing ad copy.
- **Success story:** Shopify's AI-powered content tools.
- **Efficiency:** Reduces time and cost significantly.
- **Personalization:** Tailored to audience preferences at scale.



INDUSTRY IMPACT

- **Cost efficiency:** Automating repetitive content tasks.
- **Speed:** Accelerating product design and content creation.
- **Personalization:** Targeting niche audiences effectively.
- **Scaling:** Enabling small businesses to compete with giants.
- **Real-world example:** AI-driven personalized marketing at scale.

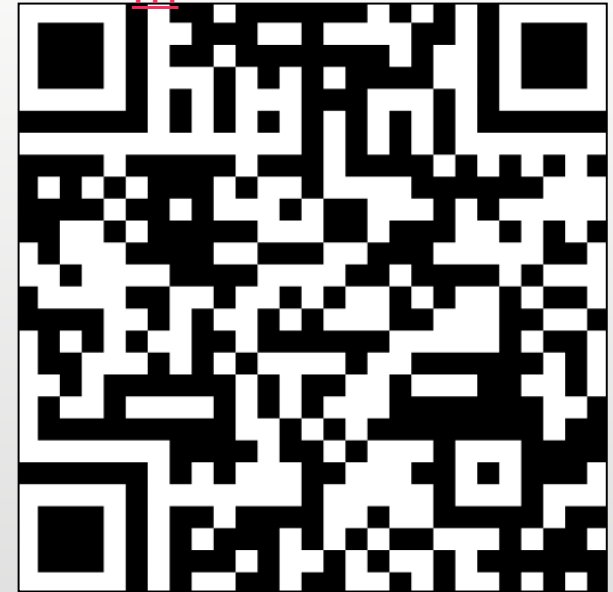
TOOLS AND FRAMEWORKS

- **Hugging Face:** Open-source hub for NLP models.
- **OpenAI APIs:** GPT-powered solutions for businesses.
- **Stable Diffusion:** State-of-the-art image generation.
- **Google Vertex AI:** Cloud platform for generative solutions.

ACTIVITY TIME

Generate real-time content using a GPT API

[www.menti.co
m](https://www.menti.com)

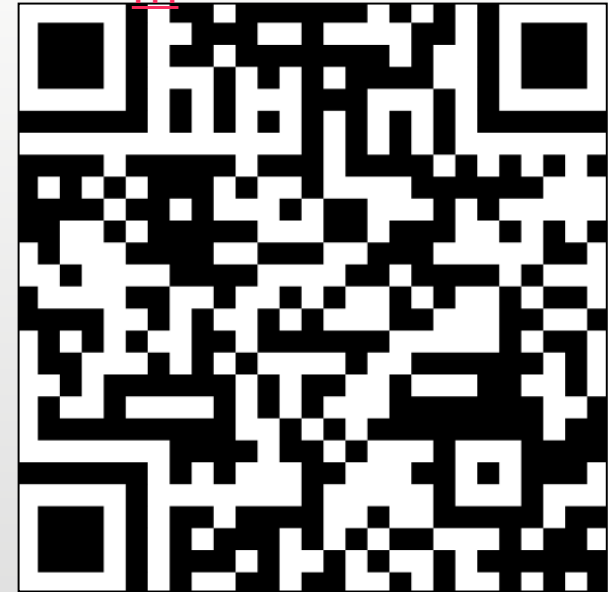


Code: 7136
1026

PART IV

Major Project Demo: Retrieval-Augmented QA
System

[www.menti.co
m](https://www.menti.com)



Code: 7136
1026

HOW TO CREATE A PROBLEM STATEMENT

- **Business Needs Document:** What needs to be created, High-Level Deliverables, needs of the project will be analyzed.
- **Business Case:** Economic Feasibility Study
 - Check for **Business Feasibility**
 - Check for **Technical Feasibility**
 - Check for **User Feasibility**
- **Benefits Management Plan:** How and when Benefits of the project will be derived and measured.

PROBLEM STATEMENT AND PROPOSED SOLUTION

- **Problem Statement:** Organizations struggle to retrieve relevant, context-aware answers from large volumes of unstructured technical documentation. Traditional keyword-based search engines fail to deliver accurate, context-rich responses, causing delays and inefficiencies.
- **Proposed Solution:** A Retrieval-Augmented QA System integrates advanced document retrieval and generative AI. By embedding technical documents into vector representations and using FAISS for efficient indexing, the system retrieves relevant documents and generates contextually accurate responses using models like T5 or GPT.

KEY TECHNICAL REQUIREMENTS

- **Data preparation** and embedding generation for large technical documents.
- **Embedding-based retrieval** using vector search libraries like FAISS.
- **Generative model integration** for natural language answer generation.
- **Deployment and scalability**, ensuring the system can handle thousands of queries simultaneously.



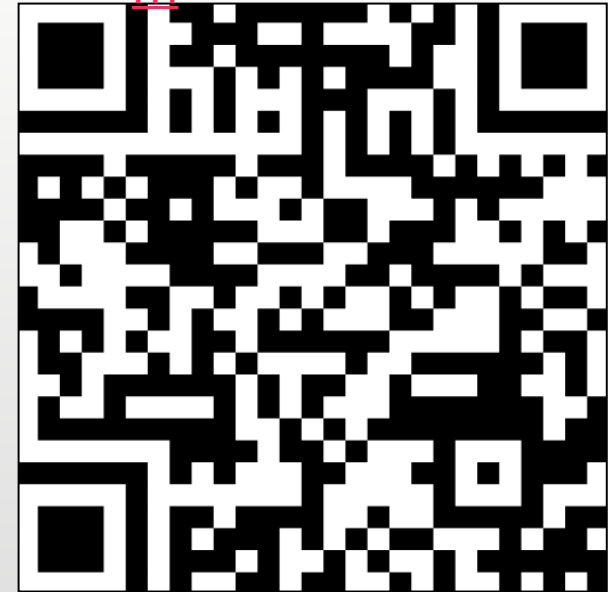
LIVE DEMO

- Data preparation and indexing
- Retrieval pipeline implementation
- Response generation and testing
- Deployment and Q&A

PART V

Closing and Key Takeaways

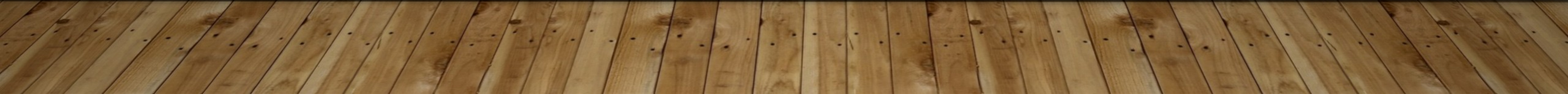
[www.menti.co
m](https://www.menti.com)



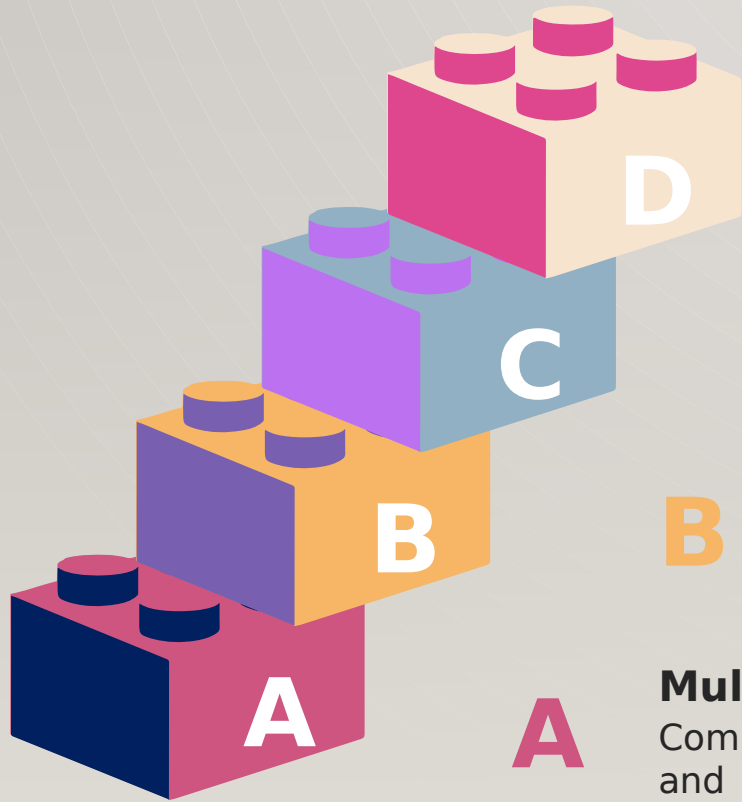
Code: 7136
1026



RECAP OF CONCEPTS COVERED

- **Advanced NLP systems:** Transformers and LLMs in industries.
 - **RAG** architecture and industrial applications.
 - **Generative AI** principles, tools, and ethical considerations.
 - **Live demo:** Retrieval-Augmented QA system implementation.
 - Deployment strategies and real-world challenges.
- 

FUTURE TRENDS IN NLP



A

Multimodal NLP Integration

Combining text with other data modalities like images, videos, and audio to build more comprehensive AI systems for applications such as healthcare diagnostics and content generation.

B

Low-Resource and Multilingual NLP

Developing models that can perform efficiently in low-resource languages and across multiple languages, ensuring inclusivity and global accessibility.

C

Efficient and Scalable NLP Models

Advancements in lightweight and efficient models like DistilBERT and quantization techniques to address computational and energy efficiency challenges in real-world deployments.

D

Ethical AI and Bias Mitigation

Research and tools to detect, address, and prevent biases in NLP systems, ensuring fairness, transparency, and trustworthiness in AI applications.

RESOURCES FOR FURTHER LEARNING

Research

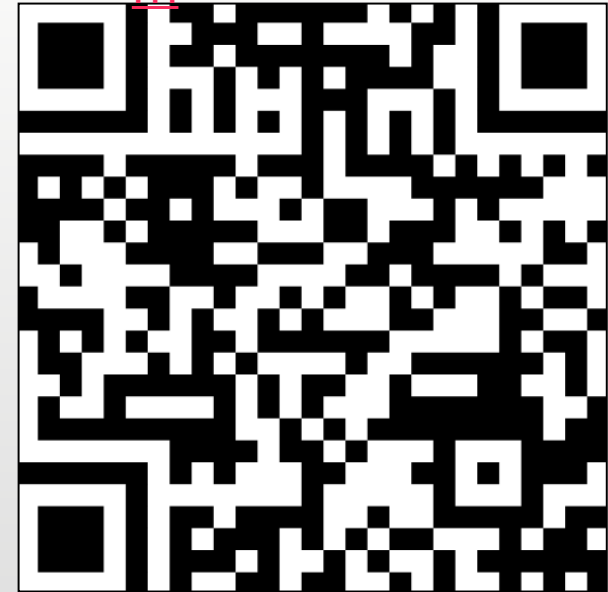
- "*Language Models are Few-Shot Learners*" by Tom B. Brown, et al.
- "*Retrieval-augmented Generation for Knowledge-Intensive NLP Tasks*" by Patrick Lewis, et al.
- "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*" by Jacob Devlin, et al.

Books

- "*Speech and Language Processing*" by Daniel Jurafsky and James H. Martin
- "*Transformers for Natural Language Processing*" by Denis Rothman

OPEN DISCUSSION

[www.menti.co
m](https://www.menti.com)



Code: 7136
1026

LET'S STAY CONNECTED !

KEY ACHIEVEMENTS

- Panda, D.S., Dixit, R., Dixit, A. et al. '*Mathematical Model and AI Integration for COVID-19: Improving Forecasting and Policy-Making*'. SN COMPUT. SCI. 5, 246 (2024) [Link](#)
- Dixit, R., Panda, D.S. & Panda, S.S. '*An Advanced Susceptible-Exposed-Infectious-Recovered model for quantitative analysis of COVID-19*'. Sādhana 46, 85 (2021) [Link](#)
- Panda, S.S., Panda, D.S., Dixit, R. (2023) '*Revolutionary Solutions for Comprehensive Assessment of COVID-19 Pandemic*'. In: Tiwari, R., Pavone, M.F., Ravindranathan Nair, R. (eds) Proceedings of International Conference on Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. [Link](#)

LET'S COLLABORATE

- Follow-up sessions or workshops on trending **Advanced NLP** concepts.
- Collaborative research or **industry-academia projects**.
- Mentoring students interested in NLP.

CONTACT DETAILS

- **Mobile:** (+91) 6370 777 150
- **Email:** pandadevsourav@gmail.com
- **LinkedIn:** [/in/devsouravpanda](https://in/devsouravpanda)
- **Google Scholar:** [/citations?user=D-upDdYAAAAJ](https://citations?user=D-upDdYAAAAJ)

Thank You
For your Attention

