Lecture 01 - Contents

An overview of the main sections in this lecture.

Part 1
Medical AI Revolution

Part 2
State-of-the-Art Models

Part 3
Real-World Deployments

Hands-on
Environment Setup and Assignment

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Advanced Medical LLMs: Transforming Healthcare with AI

20-Lecture Comprehensive Course

Introduction to Biomedical Data Science

Instructor Name

Medical AI Research Center

Fall 2025

Course Overview and Prerequisites

📚 Learning Objectives

- Understand medical LLM architectures
- Master clinical NLP techniques
- Deploy healthcare AI systems
- Ensure HIPAA compliance
- Evaluate model performance

📋 Prerequisites

- Python programming
- Machine learning basics
- Medical terminology
- Neural networks fundamentals

PyTorch Hugging Face FHIR Docker

📊 Assessment Breakdown

40%
Project

30%
Assignments

30%
Exam

🎯 Course Structure

- 20 comprehensive lectures
- Hands-on coding sessions
- Real-world case studies
- Industry expert guest lectures

Part 1/3:

The Medical AI Revolution

1. Evolution from Rule-Based Systems to LLMs

2. GPT-4, Claude, and Gemini in Healthcare

3. Medical vs General-Purpose LLMs

Evolution From Rule-Based Systems to LLMs

Rule-Based
1970s - 1990s

Technology:
Expert systems
Decision trees
If-then rules

Limitation:
*Cannot handle
uncertainty*

Statistical ML
2000s - 2010s

Technology:
SVM, Random
Forest, Naive
Bayes

Limitation:
*Feature engineering
required*

Deep Learning
2010s - 2020

Technology:
CNN, RNN
LSTM, Attention
mechanisms

Limitation:
*Task-specific
Large labeled data*

Large LLMs
2020s - Present

Technology:
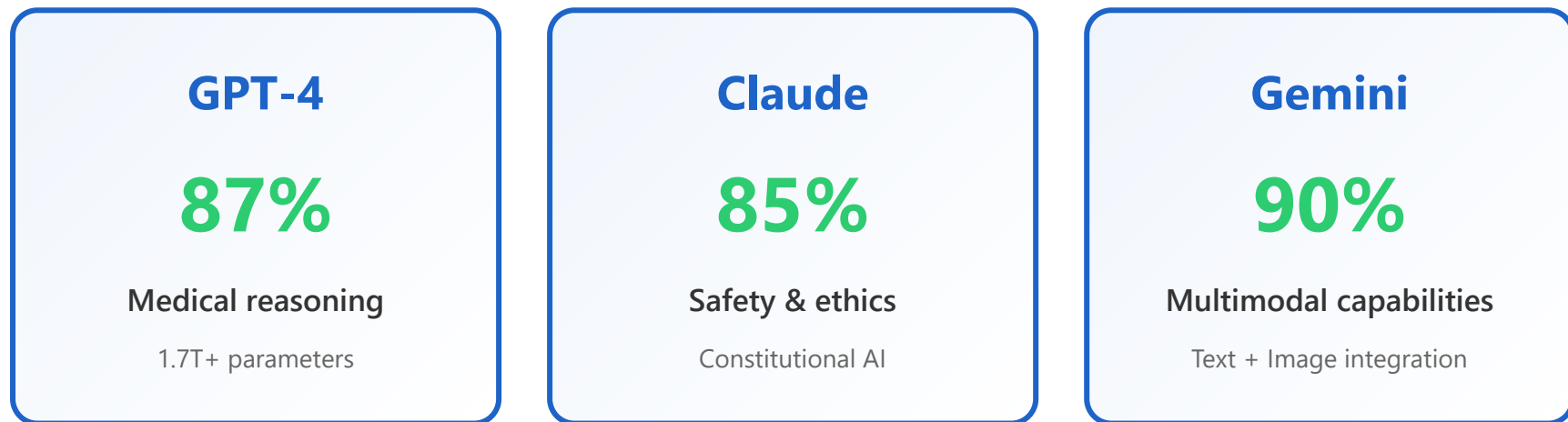Transformers
Self-attention
Pre-train + Fine-tune

Strength:
*Few-shot learning
Broad knowledge*

GPT-4, Claude, and Gemini in Healthcare

GPT-4

87%

Medical reasoning

1.7T+ parameters

Claude

85%

Safety & ethics

Constitutional AI

Gemini

90%

Multimodal capabilities

Text + Image integration

Medical vs General-Purpose LLMs

🏥 Medical LLMs

✓ Domain-specific knowledge

✓ Clinical safety protocols

✓ HIPAA/regulatory compliance

✓ Medical terminology precision

✓ Evidence-based responses

✓ Rare disease understanding

🌐 General-Purpose LLMs

✓ Broad knowledge coverage

✓ Creative problem-solving

✓ Multi-domain versatility

✓ General reasoning ability

✓ Larger training datasets

✓ Conversational flexibility

🔄 Common Ground

Language understanding • Contextual reasoning • Pattern recognition • Natural language generation

Key Medical AI Challenges

🔒 Data Privacy

HIPAA compliance, patient data protection, secure data handling and storage

Critical Risk

⚠️ Medical Errors

Hallucinations, incorrect diagnoses, liability concerns, patient safety

High Risk

📋 Regulatory Approval

FDA clearance process, clinical validation, compliance documentation

High Complexity

🔍 Explainability

Model interpretability, decision transparency, clinical reasoning clarity

Medium Risk

⚖️ Bias & Fairness

Demographic bias, training data representation, equitable healthcare access

High Risk

💰 Implementation Cost

Infrastructure investment, training costs, maintenance and updates

Medium Complexity

HIPAA Privacy and Data Security

🔒 18 PHI Identifiers

👤 Names	🔢 SSN
🏠 Addresses	📋 Medical Records
📅 Dates	📞 Phone Numbers
📧 Email	🌐 IP Address
👆 Biometric IDs	📷 Photos

🛡️ Security Requirements

✓ End-to-end encryption (AES-256)

✓ Access control & authentication

✓ Audit logging & monitoring

✓ Data de-identification

✓ Secure data transmission (TLS)

✓ Regular security assessments

Multi-layer Security Architecture

⚖️ Compliance Checklist

📄	💾	🚨
Privacy Notice	**Data Backup**	**Incident Response**
👨‍🏫	🤝	📊

| Staff Training | Business Associate Agreements | Risk Analysis |

Clinical NLP Task Taxonomy

Named Entity Recognition
EXTRACTION

Identify diseases, medications, symptoms, procedures from clinical text

EXAMPLE
Patient has diabetes and takes metformin 500mg daily

94% F1

Relation Extraction
RELATION

Find relationships between medical entities (drug-disease, symptom-cause)

treats

EXAMPLE
Aspirin →treats→ Headache

91% F1

Temporal Expression
TEMPORAL

Normalize time expressions and sequence medical events chronologically

T-14d   T-7d   Today

EXAMPLE
"2 weeks ago" → 2024-01-01
"after surgery" → T+1d

88% Acc

Negation Detection
LOGIC

Identify negated medical concepts and uncertainty modifiers

EXAMPLE
"No signs of infection" ✓
"Possible pneumonia" ?

92% Prec

Medical Code Mapping
MAPPING

Map clinical text to ICD-10, CPT, SNOMED-CT codes

Text → ICD-10

EXAMPLE
"Hypertension" →
ICD-10: I10

87% MAP

Clinical Text Summarization
GENERATION

Generate concise summaries from lengthy clinical notes

EXAMPLE
5-page note → 3-line summary
Compression: 95%

ROUGE 0.82

Diagnosis Generation Systems

Data Input
Symptoms, Lab Results, History

JSON

Preprocessing

AI Processing
LLM Reasoning Engine
Knowledge Base Retrieval

BERT

Analysis

Diagnosis Output
Ranked Differential Diagnosis

API

⚡ **Processing Time: <2 seconds**

Accuracy: 92% | Precision: 89%
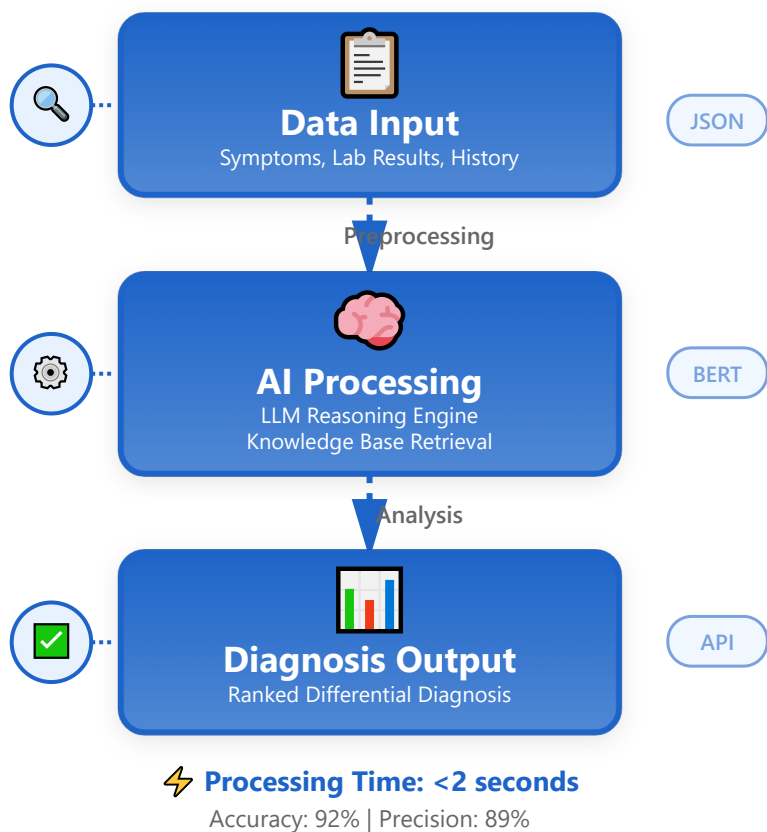
🎯 Example Output: Differential Diagnosis

1 Community-Acquired Pneumonia **78%**

Evidence: Fever (38.5°C), productive cough, chest X-ray infiltrates

HIGH CONFIDENCE

2 Acute Bronchitis **15%**

Evidence: Cough duration <3 weeks, no infiltrates   MODERATE

3 Viral URI **7%**

Evidence: Mild symptoms, no bacterial markers   LOW

Total Conditions Analyzed: 147   🔄 Real-time Updates

Medical Knowledge Graph Integration
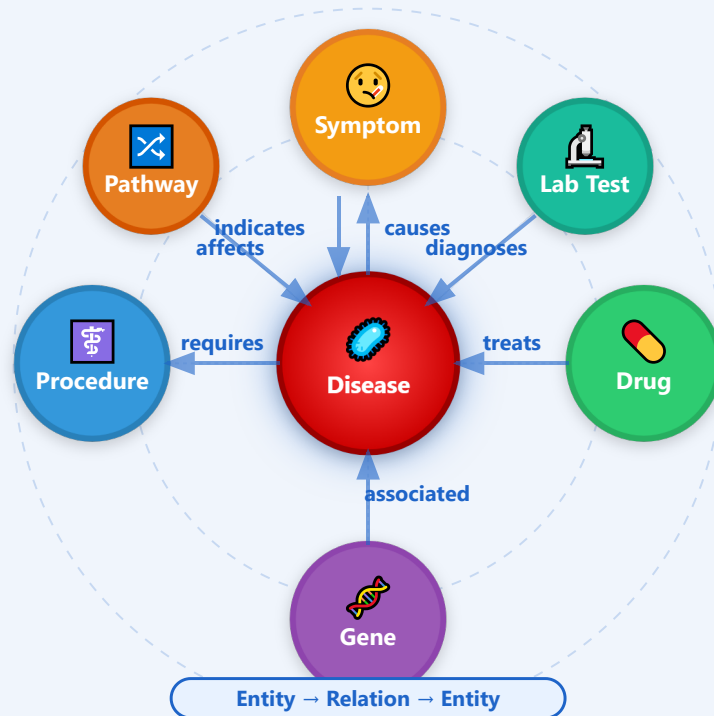
2.8M
UMLS Concepts

15M+
Relationships

🦴 **Graph Embedding**

TransE, TransR relations

Node2Vec embeddings

Graph Attention Networks

Knowledge Graph Embeddings

🔗 **LLM Integration**

RAG	K-Prompting
Graph Reasoning	Entity Linking
Hybrid Retrieval	Normalization

Diagram labels:

- Symptom
- Pathway
- Lab Test
- Procedure
- Disease
- Drug
- Gene

Relations: indicates, affects, causes, diagnoses, requires, treats, associated

Entity → Relation → Entity

Part 2/3:

State-of-the-Art Medical Models

🏥 MedPaLM 2

🧬 BioGPT

☐ RadBERT

🎭 Multimodal Systems

🐊 GatorTron

📚 PubMedGPT

🔬 PathLLM

MedPaLM 2 Architecture

540B
Parameters

86.5%
MedQA Score

Flan-PaLM
Base Model

64
Layers

SOTA Performance

Input	📋 **Medical Query Input** Tokenization & Embedding
	[Tokens]
Encode	🔤 **Embedding Layer** 50,257 vocab × 8,192 dim
Process	🧠 **Transformer Blocks** 64 Layers × 128 Attention Heads Feed-Forward: 32,768 dim
	[Context]
Adapt	🏥 **Medical Adaptation** PubMed + MIMIC-III + Clinical Notes 90B medical tokens
Output	✨ **Clinical Output** Diagnosis + Confidence + Evidence

Top-1 **87%** Accuracy

Memory: 1.08TB (80% utilization)

🎯 **Instruction Tuning**
Medical task-specific fine-tuning

💡 **Chain-of-Thought**
Step-by-step clinical reasoning

🚀 **Few-Shot Learning**
5-shot prompting for rare diseases

🛡️ **Safety Alignment**
Medical ethics & harm reduction

📚 **Knowledge Base**
UpToDate, PubMed integration

🔄 **Continuous Learning**
Weekly medical literature updates

GatorTron - Clinical BERT

8.9B Parameter Clinical Language Model

Trained on 90 billion words from UF Health clinical notes

NER Performance

96%

F1 Score

Relation Extract

94%

F1 Score

Parameters

8.9B

Largest clinical

BioGPT vs PubMedGPT Comparison

BioGPT

1.5B Parameters

Focus: General biomedical text generation

Training: PubMed abstracts (15M documents)

Strengths: Question answering, summarization

Use Case: Research assistance, literature review

Speed: Faster inference time

PubMedGPT

2.7B Parameters

Focus: Medical literature specialization

Training: Full PubMed papers (3M+ full text)

Strengths: Scientific writing, detailed analysis

Use Case: Literature review, paper generation

Depth: More comprehensive knowledge

📊 Performance Comparison

	BioGPT	PubMedGPT
Model Size	1.5B	2.7B
Training Data	15M abstracts	3M full papers
Inference Speed	Fast	Moderate
Knowledge Depth	Good	Excellent

Reserved Slot (L01_16)

추후 내용이 추가될 자리입니다. 강의 흐름의 연속성을 위해 번호를 보존합니다.

Multimodal Medical Models

Text + Image
Radiology reports with X-rays, CT scans, MRI

Text + Signal
ECG, EEG, vital signs waveform analysis

Text + Video
Surgical procedures, endoscopy, ultrasound
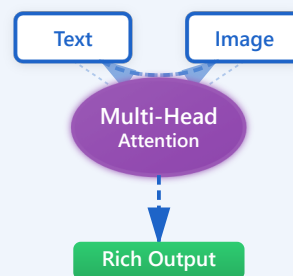
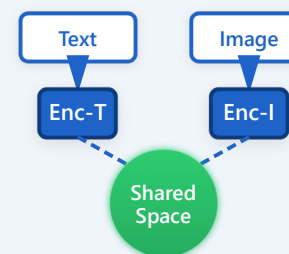## Fusion Strategies Comparison

Early Fusion

Text | Image

Concatenate
Joint Processing

Output
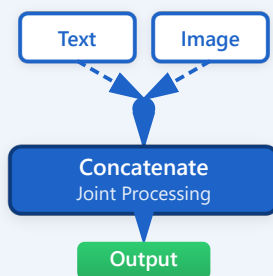
Speed: ⚡⚡⚡
Simple, Fast

Late Fusion

Text | Image

BERT | CNN

Output
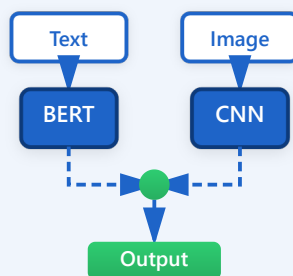
Accuracy: ⭐⭐⭐
Flexible, Modular

Cross-Attention

Text | Image

Multi-Head Attention

Rich Output

Quality: ⭐⭐⭐⭐⭐
Best Performance

Joint Space

Text | Image

Enc-T | Enc-I

Shared Space

Retrieval: ⭐⭐⭐⭐
Cross-Modal Search

RadBERT

Radiology Applications

📊 4.7M radiology reports

✓ Chest X-ray abnormality detection: 94% AUC

✓ Automated report generation

✓ Finding classification & localization

✓ Report quality assessment

🔬
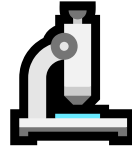
PathLLM

Digital Pathology

📊 **Whole Slide Image (WSI) processing**

✓ Cancer grading & classification

✓ Cell counting & detection

✓ Tissue segmentation

✓ Histopathology report generation

Part 3/3:

Real-World Clinical Deployments

🏥 Hospital Case Studies

🚨 Emergency Department Applications

📊 Clinical Decision Support

✅ FDA Approved Systems

💰 ROI Analysis

🏥 Mayo Clinic Case Study

2021

Deployment Year

3M annually

Scale

15% diagnosis accuracy ↑

Key Result

✨ Key Achievements

✓ 30% physician time saved

✓ 3-year payback period

✓ Enhanced clinical workflows and patient outcomes

🏥 Stanford Healthcare Case Study

📅 **2022**

Deployment Year

Epic EHR integration

System
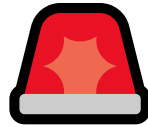
47% medication error reduction

Key Result

✨ Key Achievements

✓ Real-time clinical decision support

✓ Improved patient safety

✓ Enhanced clinical workflows and patient outcomes

🚨

Emergency Department Triage

23 min average wait time reduction

92% priority accuracy

ESI level prediction

Real-time risk assessment

Clinical Decision Support Systems

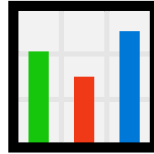| Drug interaction alerts	Evidence-based guidelines
Lab test optimization	Risk score calculation

FDA Approved AI Systems

520+ FDA cleared devices	IDx-DR (diabetic retinopathy)
Caption Health (echocardiography)	Rapid regulatory growth

Performance Metrics & Benchmarks

MedQA: 87% accuracy

PubMedQA: 78% F1

MMLU-Medical: 91%

AUROC, Sensitivity, Specificity

Cost-Benefit Analysis

Initial: $2M investment	Operating: $500K/year
Savings: $3M/year (error reduction)	ROI: 250% over 3 years

🚀 Future Directions & Opportunities

Personalized Medicine
Genomic data integration, tailored treatment plans

Real-time Monitoring
Wearable integration, continuous health tracking

Drug Discovery
Accelerated compound identification, clinical trials

Precision Surgery
AI-assisted robotics, real-time guidance

Preventive Medicine
Early disease detection, risk prediction

Global Health Access
Telemedicine, underserved populations

⚙️ Hands-On Environment Setup

📦 **Required Software**

✓ Python 3.9+	✓ PyTorch 2.0
✓ Transformers 4.30	✓ Docker Desktop
✓ CUDA 11.8+ (GPU)	✓ Jupyter Notebook

```
# Installation Commands pip install torch transformers datasets pip install huggingface_hub
accelerate pip install pandas numpy scikit-learn pip install jupyter notebook # Verify
Installation python -c "import torch; print(torch.cuda.is_available())"
```

📝 Assignment: Medical QA System

🎯 **Task:** Build a medical question-answering system

📊 **Dataset:** MedQA or PubMedQA (provided)

⏰ **Deadline:** 2 weeks

📋 **Evaluation:** Accuracy, code quality, documentation

Thank You! 🎉

Questions? Contact your instructor
📧 instructor@university.edu

Next Lecture: Fine-tuning Medical LLMs