# Appendix A Large Language Model Paper Codes

## A.1 Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models

- Lack of annotated data
- Phenotyping in clinical notes
- Use of LLM (Flan- T5)
- Zero- shot prompting
- EHR data utilization
- Positive predictive value
- Subtype identification
- Interpretability of model
- Lack of fine- tuning
- Federated learning potential
- Gynaecology domain
- US hospital data

## A.2 Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models

- Meta- analysis challenges
- Labor- intensive processes
- LLM application (ChatGPT)
- Zero- shot prompting
- Information extraction from RCTs
- GPT- JT model
- Drug repurposing in cancer therapy
- Pharmacovigilance in chronic myeloid leukemia
- Article data from publisher databases
- Systematic error analysis
- Potential inaccuracies
- Variability in RCT quality
- Continuous model updates and refinements

## A.3 Working With AI to Persuade - Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages

- Impact of AI on public health messaging
- Influence on public perception and behavior
- Use of GPT- 3 for pro- vaccination messages
- Persuasive quality of AI- generated messages
- Disruption of conventional methods
- Public Health domain
- Prompt engineering (synonyms, zero- shot, few- shot, automated)

- No dataset for message generation
- Effectiveness and persuasiveness of AI messages
- Preference for human- sourced messages
- Participant biases
- Trustworthiness concerns in AI- generated content
- Need for best practices in AI public health messaging

## A.4 WangLab at MEDIQA-Chat 2023 - Clinical Note Generation from Doctor-Patient Conversations using Large Language Models

- Automatic clinical note generation
- Enhancing medical documentation efficiency
- Fine- tuning PLM on task- specific data
- Few- shot in- context learning (ICL) with GPT- 4
- Dialog to note conversion
- Longformer Encoder Decoder (LED)
- 0, 1, 2, 3- shot prompting
- MEDIQA- Chat 2023 shared task data
- Use of ROUGE and BERTScore for evaluation
- High performance in shared task
- Expert human evaluation preference
- Potential biases in evaluation datasets
- Need for ongoing LLM developments
- Distributed optimal prompting strategy potential

## A.5 VetLLM - Large Language Model for Predicting Diagnosis from Veterinary Notes

- Lack of diagnosis coding in veterinary notes
- Hindrance to medical and public health research
- Use of open- source LLMs
- Zero- shot diagnosis coding
- Fine- tuning for improved performance
- Veterinary Practice domain
- Alpaca- 7B model
- CSU and PP test data
- F1 score evaluation
- VetLLM creation through fine- tuning
- Data efficiency in fine- tuning
- Performance variability across note types
- Dependency on fine- tuning data quality
- Federated learning potential for fine- tuning

## A.6 Validation of a Deep Learning Chest X-ray Interpretation Model - Integrating Large-Scale AI and Large Language Models for Comparative Analysis with ChatGPT

- Diagnostic accuracy in chest X- ray reading
- Validation of AI techniques (KARA- CXR and ChatGPT)
- Evaluation based on qualitative factors
- Medical Imaging domain
- Use of DICOM data
- Zero- shot prompting for LLMs
- KARA- CXR higher diagnostic accuracy
- Comparison of false findings and hallucinations
- Interobserver agreement
- Single institution data bias
- Qualitative evaluation limitations
- Need for improvement in AI systems
- Potential for federated learning deployment

## A.7 Use Of Artificial Intelligence Large Language Models As A Clinical Tool In Rehabilitation Medicine - A Comparative Test Case

- AI application in clinical practice
- Rehabilitation prescriptions formulation
- ICF codes generation for stroke patients
- Use of ChatGPT- 4
- Comparison with standard textbook answers
- Medical Decision Support domain
- GPT- 4 prompted with stroke clinical case
- Broader and general prescriptions by AI
- Rehabilitation therapy approaches proposed
- ICF category error identified
- Sample size limitations
- Lack of prompt diversity and fine- tuning
- Federated learning potential in Medical QA applications

## A.8 Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support

- LLM- powered conversational agents in mental well- being
- Replika as a mental health chatbot
- Use of GPT- 3 in Replika
- On- demand, non- judgmental support
- User feedback from subreddit (r/Replika)
- Benefits in user confidence and self- discovery

- Challenges in filtering harmful content
- Inconsistencies in communication
- Memory retention issues
- Risks of user overdependence
- Social isolation due to stigma
- Need for responsible use and evaluation
- Federated learning opportunities for privacy- preserving support

## A.9 Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention

- LLM- driven chatbots in public health
- CareCall chatbot for socially isolated individuals
- Use of HyperCLOVA LLM
- Few- shot prompting with dialog corpus
- Qualitative data from focus groups and interviews
- Mitigation of loneliness and emotional burdens
- Challenges in meeting health needs
- Stakeholder tensions and system expectations
- Qualitative study limitations
- Federated learning opportunities in chatbot applications

## A.10 Two Directions for Clinical Data Generation with Large Language Models - Data-to-Label and Label-to-Data

- Detection of Alzheimer's Disease- related signs and symptoms
- Use of synthetic clinical data for data augmentation
- Novel taxonomy for AD sign and symptom progression
- Application of GPT- 4 for synthetic dataset generation
- Evaluation using BERT, RoBERTa, and ClinicalBERT
- Longitudinal EHRs, public EHR collections, and synthetic data
- Improvement in detection performance with synthetic data
- Quality and reliability concerns with synthetic data
- Dependence on expert knowledge for taxonomy and datasets
- Potential biases in LLM- generated data
- Federated learning opportunities for centralized validation and data augmentation

## A.11 Trialling a Large Language Model (ChatGPT) in General Practice with the Applied Knowledge Test - Observational Study Demonstrating Opportunities and Limitations in Primary Care

- Assessing ChatGPT's performance on a standardized medical test
- Application of ChatGPT for medical QA
- Use of AKT practice questions for evaluation

- Performance metrics: 60.17
- Performance variability across subject categories
- Lack of prompting strategies used in the evaluation
- Potential for Federated Learning in medical QA applications

## A.12 Trends in Accuracy and Appropriateness of Alopecia Areata Information Obtained from a Popular Online Large Language Model, ChatGPT

- Evaluation of LLM in Dermatology
- Accuracy of ChatGPT in Medical QA
- Patient use of AI for Health Information
- Comparison of ChatGPT Versions (3.5 vs 4.0)
- Appropriateness of AI- Generated Health Information
- Expert Evaluation of AI Responses
- Reliability of AI in Patient Education
- Limitations of Non- Fine- Tuned LLMs
- Dermatologist- Created Question Bank
- User Preferences for AI Version
- Agreement Among Medical Reviewers
- Potential Bias in AI Responses
- Role of AI in Dermatology Practice
- ChatGPT in Alopecia Areata Management
- Assessment of AI in Healthcare

## A.13 Towards Interpretable Mental Health Analysis with Large Language Models

- Limitations in Mental Health Studies Using LLMs
- Benchmarking LLMs on Mental Health Tasks
- Use of Prompt Strategies in LLMs
- Explainability in LLMs
- Evaluation of LLMs Across Multiple Datasets
- Performance of ChatGPT in Mental Health Analysis
- Emotion- Enhanced Chain of Thought Reasoning
- Comparison of LLMs (ChatGPT, InstructGPT, LLama)
- Human Evaluation of LLM Outputs
- Performance of Traditional Methods vs. LLMs
- Instability of LLMs to Prompt Changes
- Opportunities for Explainable AI in Federated Learning
- Application of LLMs in Explainable AI
- Role of Prompt Engineering in LLM Performance
- Evaluation of Explainability in Mental Health Tasks

### A.14 The Use of Large Language Models to Generate Education Materials about Uveitis

- Evaluation of LLMs for Medical Information
- Readability of LLM- Generated Health Information
- Comparison of ChatGPT and Bard
- Use of Simple Prompts in LLMs
- Flesch- Kincaid Grade Level Analysis
- Complexity of LLM- Generated Content
- Comparison with Search Engine Results
- Uveitis Health Information
- Limitations of LLMs in Healthcare
- Potential of LLMs in Patient Education
- Opportunities for Federated Learning in Medical QA
- Impact of Prompt Design on LLM Output
- LLM Performance in Ophthalmology
- Readability of Online Health Information
- Challenges in Simplifying Medical Information

### A.15 The promise and peril of using a large language model to obtain clinical information - ChatGPT performs strongly as a fertility counseling tool with limitations

- Evaluation of LLMs in Fertility Counseling
- Hallucination Risk in LLMs
- Accuracy of ChatGPT Responses
- Comparison with CDC FAQs
- Standardized Fertility Knowledge Surveys
- Length and Content Analysis of LLM Outputs
- Sentiment Analysis of LLM Responses
- Subjectivity in LLM- Generated Content
- Incorrect Information in LLM Responses
- Use of General Purpose GPT Model
- Lack of Prompt Engineering in Evaluation
- Potential of LLMs in Medical QA
- Fertility Counseling and LLMs
- Opportunities for Federated Learning in Medical Applications
- Clinical Consensus and LLM Accuracy

### A.16 Team Cadence at MEDIQA-Chat 2023 - Generating, augmenting and summarizing clinical dialogue with large language models

- Automated Summarization of Conversational Data
- MEDIQA- Chat- 2023 Challenge
- Dialogue2Note Summarization

- Clinical Note Generation
- Data Augmentation with Note2Dialogue
- BART Architecture Utilization
- Fine- tuning with SAMSum Dataset
- N- pass Strategy for Large Text Summarization
- Evaluation with MIMIC- IV and MEDIQA- Chat- 2023
- ROUGE Score Improvements
- Section Header Classification
- Resource Constraints in LLM Deployment
- Opportunities for Federated Learning
- Medical Data Summarization Challenges
- Synthetic Data Generation in Healthcare

## A.17 SPeC - A Soft Prompt-Based Calibration on Performance Variability of Large Language Model in Clinical Notes Summarization

- Unstable Quality of LLM- Generated Prompts
- Soft Prompt- Based Calibration (SPeC)
- Clinical Notes Summarization
- Model- Agnostic Approach
- ChatGPT Utilization
- MIMIC- CXR Dataset
- Variance Reduction in Prompt Outputs
- Maintaining Quality in LLM Outputs
- Patient Privacy Challenges
- Data Security Compliance
- Hallucination Risks in LLMs
- Prompt Automation for Federated Learning
- Hospital Setting Applications
- Radiology Report Summarization
- Opportunities for Federated Learning in Prompt Engineering

## A.18 Rule-Augmented Artificial Intelligence-empowered Systems for Medical Diagnosis using Large Language Models

- Hallucinations in LLMs
- Bias in Medical Applications
- Lack of Explainability
- Lack of Validation
- Rule- Based System Integration
- LLM- Enhanced Diagnosis
- ChatGPT API (text- davinci- 003)
- Symptom Pre- Processing
- Initial Diagnosis Generation

- Clinical Data Integration
- Rule- Based Classification
- Iterative Diagnosis Refinement
- Hospital Setting
- Medical Decision Support
- Non- Experimental Study
- Limitations of General- Purpose LLMs
- Opportunities for Federated Learning in Rule- Based Systems
- Diagnosis and Treatment Recommendations

## A.19 Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale

- Emergency Medicine
- ChatGPT for ER Triage
- Canadian Triage and Acuity Scale (CTAS)
- Emergency Physicians' Input
- Prompt Development
- Repeatability of Prompts
- Reproducibility of Prompts
- Patient Vignette Simulation
- Triage Accuracy
- Over- Triage and Under- Triage
- Prompt Complexity and Reproducibility
- Consistency Issues in LLMs
- Temperature Settings in LLMs
- Limitations of General- Purpose LLMs
- Federated Learning Potential
- Privacy Concerns in LLM Usage
- Model Optimization
- Classification in Hospital Settings

## A.20 Recommendations for initial diabetic retinopathy screening of diabetic patients using large language model-based artificial intelligence in real-life case scenarios

]

- Ophthalmology
- Diabetic Retinopathy (DR) Screening
- AI in Medicine
- Large Language Models (LLMs)
- Clinical Decision Support

- ChatGPT 3.5
- ChatGPT 4.0
- Bing Chat
- Hypothetical Case Scenarios
- Prompt Engineering
- Inter- Rater Reliability
- Majority Clinician Response
- Majority AI Response
- Kappa Value Analysis
- Systemic Co- Morbidities
- DR Screening Timing
- Medical Non- Ophthalmologists
- Simple Prompting
- General Purpose LLMs
- Dataset Limitations
- Federated Learning (FL) Potential
- Medical QA Applications

## A.21 PULSAR - Pre- training with Extracted Healthcare Terms for Summarising Patients' Problems and Data Augmentation with Black- box Large Language Models

- Hospital Patient Intake
- Data Summarization
- Data Generation
- Large Language Models (LLMs)
- Flan- T5
- BioMedLM
- Medical Data Augmentation
- Data Pipeline Framework
- QuickULMS
- BioNLP 2023 Shared Task 1A
- Rouge Measures
- Problem List Generation
- NER (Named Entity Recognition)
- Resource Intensive Methods
- Optimization Techniques (LoRA)
- Tokenization
- General Purpose LLMs
- Federated Learning (FL)
- Privacy Preservation
- Scaling Solutions

## A.22 Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models - a retrospective cohort study

- Seizure Recurrence Prediction
- Neurology
- Clinical Data
- Large Language Models (LLMs)
- Fine- Tuning
- Longformer
- XGBoost
- Logistic Regression
- Clinical Notes
- Data Augmentation
- MIMIC Dataset
- Boston Children's Hospital Dataset
- Electronic Health Records (EHR)
- Feature Vector Analysis
- Domain- Specific Pre- Training
- Lack of Structured Data
- Performance Metrics (F1- Score, AUC, AUROC)
- Specialized Cohorts
- Privacy Concerns
- Federated Learning (FL)
- Synthetic Data Generation

## A.23 Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries

- Medical QA evaluation
- Ophthalmology queries
- LLM performance comparison
- Response quality assessment
- Data source: National Eye Institute
- Comprehensiveness scores
- Self- awareness capabilities
- General purpose LLMs limitations
- No prompting strategy

## A.24 Performance of large language models on advocating the management of meningitis a comparative qualitative study

- LLM adherence to medical guidelines
- Evaluation of LLM responses

- Meningitis case study
- Comparison of LLM performance
- Clinical practice and international guidelines adherence
- Simulated clinical scenario
- Performance metrics: imaging, lumbar puncture, blood cultures, antibiotic treatment
- Misleading statements by LLMs
- Limitations in medical fine- tuning

## A.25 Performance of Large Language Models on a Neurology Board-Style Examination

- LLM performance in neurology board exams
- Board- style question evaluation
- GPT- 3.5 vs. GPT- 4 performance comparison
- Confidence and classification in responses
- Question bank data: neurology board questions
- Accuracy of LLMs in answering questions
- Lower- order vs. higher- order questions performance
- Confidence in answers and reproducibility
- Limitations due to non- biomedical training
- Opportunities for federated learning

## A.26 Performance of Generative Large Language Models on Ophthalmology Board–Style Questions

- LLM capability for medical information
- Chat interface evaluation: Bing Chat, ChatGPT 3.5, ChatGPT 4.0
- Board- style ophthalmology question performance
- Accuracy of LLM responses
- Comparison of LLMs with human performance
- Performance on workup- type questions
- Difficulty with image interpretation
- Challenges with multi- step reasoning
- Hallucinations and nonlogical reasoning rates
- General- purpose LLMs limitations
- Lack of prompting strategies
- Potential improvements with specific prompts

## A.27 Patients with floaters - Answers from virtual assistants and large language models

- Evaluation of voice assistants and LLMs
- Response to floaters in ophthalmology
- Readability and pertinence of answers
- Comparison of Google Assistant, ChatGPT, Bard, Alexa

- Source: American Academy of Ophthalmology
- Reading comprehension levels
- Word count comparison between LLMs and virtual assistants
- Lack of specificity in LLM versions

## A.28 Leveraging Large Language Models for Decision Support in Personalized Oncology

- LLMs as decision support tools in oncology
- Evaluation of multiple LLMs: BioMedLM, Perplexity, ChatGPT- 3.5, Galactica
- Criteria for LLM comparison: usage, model size, openness, pretraining domain
- Prompt structure for evaluating targeted therapies
- Molecular profiles and fictional patient cases
- Treatment options proposed by LLMs vs. human expert
- F1 scores and precision/recall of LLMs
- Identification of AI- generated treatment options
- Quality and credibility of LLM- generated treatment
- Limitations: small case size, simplicity of prompts, single physician ground truth
- Opportunities for federated learning in decision support systems

## A.29 Learning to Generate Radiology Findings from Impressions Based on Large Language Model

- LLMs for generating clinical findings from radiology reports
- Impressions to findings generation process
- Reverse- prompting for clinical findings
- Use of GLM and ChatGLM architectures
- Parameter- Efficient Fine- tuning: LoRA, P- Tuning
- Evaluation metrics: Rouge, BLEU, BERTScore
- Chinese clinical and radiological dataset
- Focus on knee joint MRI data
- Automated vs. manual testing of LLM performance
- Lack of prompt strategy details
- Privacy and usability concerns
- Opportunities for federated learning in fine- tuning and anonymization

## A.30 Large Language ModelBased Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures

- LLMs for generating informed consent forms
- Evaluation of Risks, Benefits, and Alternatives (RBA)
- ChatGPT GPT- 3.5 architecture
- Readability metrics: Flesch- Kincaid, Gunning Fog, SMOG, Coleman- Liau
- Comparison of LLM- generated vs. surgeon- generated RBAs
- Assessment of completeness and accuracy of RBAs

- Readability scores: LLM (12.9) vs. surgeon (15.7)
- Completeness and accuracy scores: LLM (2.2) vs. surgeon (1.6)
- Strengths in benefits and alternatives descriptions by LLM
- No significant difference in risk descriptions
- Shortcomings: Older GPT version, general purpose model, lack of prompt strategies
- Opportunities for iterative improvement with physician intervention

## A.31 Large language models propagate race- based medicine

- Outdated medical data in LLMs
- Evaluation of race- based outdated information
- Medical ethics: race differences in medical practice
- Evaluation of GPT- 4, ChatGPT- 3.5, Google Bard, Claude
- Questions based on debunked race- based formulas
- Arbitrary questions
- Absence of fine- tuned LLM
- No prompt strategy presented

## A.32 Large Language Models for Therapy Recommendations Across 3 Clinical Specialties Comparative Study

- Evaluation of LLM treatment recommendations
- Performance of Claude- instant- v1.0, GPT- 3.5- Turbo, Command- xlarge- nightly, Bloomz
- Medical content across ophthalmology, orthopedics, dermatology
- mDISCERN score, correctness, harmfulness
- 60 diseases from three specialties
- Significant differences in model quality
- Error patterns: confusing diagnoses, vague advice, missing treatments
- GPT- 3.5- Turbo: lowest harmfulness rating
- Lack of detail on treatment risks and quality of life effects
- Strong alignment between physician assessments and GPT- 4
- Hallucinations and incomplete information concerns
- Non- biomedical LLMs used

## A.33 Large Language Models for Healthcare Data Augmentation - An Example on Patient-Trial Matching

- Matching patients with clinical trials
- LLM embeddings for patient- trial matching
- Clinical studies domain
- Pre- trained BERT (Clinical BERT) and Memory Network (Mem)
- Highway network for embedding evaluation
- GPT for rephrasings/augmentations of criteria
- Patient records and clinical trial criteria data
- 7.32
- 12.12

- Concerns about data privacy
- LLM embeddings for anonymizing patient data

## A.34 Large language models assisted multi-effect variants mining on cerebral cavernous malformation familial whole genome sequencing

- Feature extraction from genetic annotation texts
- BioBERT for text vectorization
- ResNet for pathogenicity classification
- Model BRLM for SNV classification
- Cerebral Cavernous Malformation (CCM) risk identification
- SNV data from The Cancer Genome Atlas
- Classification into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign, benign
- Accuracy exceeding 99
- FGF1 identified as significant in CCM development
- Secondary role of LLM in the study
- Potential for FL pipelines in clinical genetics

## A.35 Large Language Models as Instructors - A Study on Multilingual Clinical Entity Extraction

- Challenges with proprietary LLMs and confidentiality
- Weak supervised model distillation for healthcare
- InstructionGPT- 3 for annotation and BERT fine- tuning
- Clinical notes domain
- Gold, Silver, and Unannotated scenarios for evaluation
- E3C multilingual dataset for clinical entities
- Biases from UMLS modifications in annotations
- Opportunities for FL with local models and feature extractors

## A.36 Large Language Models are Few-Shot Clinical Information Extractors

- Challenges with small and fragmented clinical NLP datasets
- Benchmarking LLMs for clinical NLP tasks
- New annotated datasets for clinical data extraction
- Tasks: Clinical Sense Disambiguation, Biomedical Evidence Extraction, Coreference Resolution, Medication Status Extraction, Medication Attribute Extraction
- Prompting strategy for LLM data extraction
- Datasets derived from Clinical Acronym Sense Inventory (CASI)
- Handcrafted prompts and limitations of non- biomedical LLMs
- Opportunities for FL to address clinical data privacy concerns

## A.37 Large language model (ChatGPT) as a support tool for breast tumor board

- Evaluation of ChatGPT for breast tumor board decision support
- Use of ChatGPT- 3.5 for clinical management recommendations
- Comparison of ChatGPT recommendations with tumor board decisions
- Clinical vignettes with patient data for analysis
- Mean scores for summarization, recommendation, and explanation
- Privacy concerns with using internal hospital data
- Lack of prompt engineering and use of non- biomedical LLM
- Opportunities for FL to enhance privacy and scalability

## A.38 Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting

- LLMs for cancer disease response classification from radiology reports
- Comparison of LLMs with traditional NLP methods
- Data augmentation and prompt- based fine- tuning effects
- Accuracy as the performance metric
- Comparison of transformer models: BERT, BioBERT, BioClinicalBERT, BioMegatron, DeBERTa, GatorTron, PubMedGPT, RadBERT, RoBERTa, XLNet
- GatorTron achieved highest accuracy: 0.8916 to 0.8976
- Prompt- based fine- tuning reduced training sample needs
- Single- center study limits generalizability
- Opportunities for FL to enhance privacy and multi- center collaboration

## A.39 Improving the Transferability of Clinical Note Section Classification Models with BERT and Large Language Model Ensembles

- Section Classification
- LLM Performance
- Comparison with BERT
- Ensemble Methods
- Dataset Specific Knowledge
- Transferability of Models
- Token Size Restrictions
- Few- Shot Prompting
- Model Hallucinations
- Clinical Notes Domain
- Data Sources (discharge, thyme, progress)
- Opportunities for Federated Learning

## A.40 Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models

- Prompt Learning for Rare Diseases
- Phenotype Extraction
- ChatGPT and BERT (BioClinicalBERT)
- Zero and Few- Shot Learning
- Structured vs. Non- Structured Prompts
- RareDis Corpus
- NER Performance
- Fine- Tuning vs. Prompt Engineering
- Accuracy of Rare Diseases and Signs
- Annotation Consistency
- Limitations in Dataset Access
- Handcrafted Prompts
- Lack of Clinical Data

## A.41 Generating medically-accurate summaries of patient-provider dialogue A multi-stage approach using large language models

- Medical Conversation Summarization
- MEDSUM- ENT Model
- Multi- Stage Summarization
- GPT- 3 Backbone
- Medical Entity Extraction
- Prompt Chaining
- Few- Shot Prompting
- In- Context Example Selection
- Dialogue- Summary Pairs
- Pertinent Positives, Negatives, and Unknowns
- Entity Resolver
- Sample Size Limitations
- General Purpose Transformer
- Opportunities for Chatbot- Based Studies

## A.42 From language models to large-scale food and biomedical knowledge graphs

- Knowledge Graphs in Healthcare
- LLM for Named- Entity- Recognition and Linking
- Cause- Effect and Treat- Effect Relations
- SAFFRON Model (BERT, RoBERTa, BioBERT)
- Food, Chemical, and Disease Entities
- Cardiovascular Diseases and Milk
- Precision of Extracted Relations

- Pipeline Efficiency
- Limitations in Prompting Information
- Federated Learning for Privacy in Knowledge Graphs

## A.43 From jargon to clarity - Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model

- Readability Improvement in Radiology Reports
- LLM for Report Simplification and Summarization
- GPT- 3.5 Architecture
- Flesch Reading Ease Score (FRES)
- Flesch- Kincaid Reading Level (FKRL)
- Radiology Report Data (X- Ray, CT, MRI)
- Statistical Improvement in Readability
- Hallucination Rates
- Limitations in Prompt Details and Model Fine- Tuning
- Federated Learning for Summarization Adaptation

## A.44 Feasibility of Using the Privacy- preserving Large Language Model Vicuna for Labeling Radiology Reports

- Local Model for Radiology Reports
- Vicuna- 13B Architecture
- Comparison with CheXpert and CheXBERT
- Prompt Engineering
- Report Findings Categorization
- MIMIC- CXR and NIH Datasets
- Agreement with Labelers ( and AUC Scores)
- Privacy Concerns with Public LLMs
- Federated Learning Opportunities

## A.45 Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model

- Differential Diagnosis Generation
- GPT- 4 Architecture
- Top- 5 Differential Diagnoses
- Concordance with Expert Consensus
- Acceptable Alternatives
- General Purpose GPT Model
- Minimal Prompting
- Federated Learning for Local Knowledge

## A.46 Exploring Capabilities of Large Language Models such as ChatGPT in Radiation Oncology

- ChatGPT in Radiation Therapy
- GPT- 3.5- Turbo Architecture
- Multiple- Choice and Open- Ended Questions
- Correctness and Helpfulness Evaluation
- Radiation Oncology Knowledge
- General Purpose GPT Model
- Lack of Prompt Strategies
- No Repeated Questions

## A.47 Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry - Comparative Mixed Methods Study

- Comparative Evaluation of LLMs
- Generative AI in Dentistry
- ChatGPT- 3.5 Performance
- ChatGPT- 4 Performance
- Bard Model Performance
- Bing Chat Performance
- Open- Type Questions
- Scoring Against Dental Consensus
- Inaccuracies in LLM Responses
- General Purpose LLMs
- Lack of Source References
- Outdated Content
- Irrelevant or Vague Answers
- No Fine- Tuning
- Minimal Prompting Strategy

## A.48 Evaluation of Large language model performance on the Multi-Specialty Recruitment Assessment (MSRA) exam

- LLM Performance in Medical QA
- Clinical Problem Solving Exam
- MSRA Exam Question Analysis
- Llama 2 Evaluation
- Google Bard Evaluation
- Bing Chat Evaluation
- ChatGPT- 3.5 Evaluation
- Practice Questions from Qbank
- Medical Knowledge Question Answering
- Comparison with Human Users

- Bing Chat Superior Performance
- Statistical Performance Differences
- No Fine- Tuning Applied
- Common Prompt Usage

## A.49 Evaluation High-Quality of Information from ChatGPT (Artificial Intelligence—Large Language Model) Artificial Intelligence on Shoulder Stabilization Surgery

- Shoulder Instability Information Quality
- Readability of Medical Information
- ChatGPT for Patient Queries
- Orthopaedic Surgery Information
- JAMA Benchmark Criteria Evaluation
- DISCERN Score Assessment
- Flesch- Kincaid Readability Scores
- Medical Accuracy of AI Responses
- Lack of Source Citing
- General- Purpose Model Evaluation
- Prompting Strategies in AI
- Importance of Professional Consultation
- Federated Learning in Medical QA
- AI in Orthopaedic Patient Education
- Quality of Online Health Information

## A.50 Evaluating the use of large language model in identifying top research questions in gastroenterology

- ChatGPT in Gastroenterology Research
- Inflammatory Bowel Disease Questions
- Microbiome Research Queries
- AI Applications in Gastroenterology
- Advanced Endoscopy Research
- LLM Performance Evaluation
- Research Question Relevance
- Clarity and Specificity of AI Responses
- Originality of LLM- Generated Questions
- Panel Rating of AI- Generated Questions
- Inter- Rater Reliability in Evaluation
- General- Purpose LLMs in Medical Research
- Limitations of LLM in Health Research
- Impact of Non- Specific Questions
- Use of Older LLM Versions

### A.51 Evaluating the performance of large language models - ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders

- Neuropathology AI Diagnosis
- ChatGPT vs. Google Bard in Neuropathology
- Clinical Summary Analysis for Neuropathology
- Predicting Neuropathological Diagnoses
- LLM Performance in Neuropathology
- Correct Diagnosis Rates by LLM
- Differential Diagnoses in Neurology
- Prompt Engineering for Neuropathological Evaluation
- General- Purpose LLMs in Medical Diagnosis
- Federated Learning Opportunities in Clinical Notes
- Mayo Clinic Brain Bank Case Analysis
- Clinical vs. Neuropathological Diagnoses
- Accuracy of AI in Neurodegenerative Disorders
- Rationale and Diagnosis by LLMs
- Limitations of Current LLM Prompts

### A.52 Evaluating the performance of large language models in haematopoietic stem cell transplantation decision-making

- LLMs in Haematology Decision Making
- Stem Cell Transplantation with LLMs
- Comparison of GPT- 4, PaLM, Llama in Haematology
- LLM Calibration for Medical Decision Support
- Haematology Patient Histories Analysis
- Decision- Making Capabilities of LLMs
- LLM vs. Resident Decision Agreement
- Transplant Recommendations by LLMs
- LLMs in Estimating Transplant- Related Mortality
- Zero- Shot vs. Engineered Prompts in LLMs
- LLM Responses and Medical Uncertainty
- Federated Learning for Privacy- Preserving Transplant Support
- Accuracy of LLMs in Haematological Contexts
- Medical Application of LLMs in Stem Cell Transplantation
- Implications of 'I Don't Know' Responses in LLMs

### A.53 Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis

- LLMs in Urology Consultations

- Evaluating LLMs for Urolithiasis Education
- Performance Comparison of Bard, Claude, ChatGPT- 4, NewBing
- Accuracy and Comprehensiveness of LLMs in Urology
- Patient Education with LLMs in Urology
- Human Caring and Case Analysis in Urology LLMs
- Claude vs. Other LLMs in Urology Consultation
- ChatGPT- 4 and Claude in Clinical Case Analysis
- Limitations of General- Purpose LLMs in Medical Contexts
- Fine- Tuning LLMs for Improved Medical Performance
- Federated Learning for Enhanced Urology Patient Support
- Model Performance in Simulated Clinical Scenarios
- Urology Questionnaires and LLM Responses
- Evaluation Metrics for Medical LLMs
- LLMs and Patient Information Accuracy in Urology

## A.54 Evaluating large language models on medical evidence summarization

- LLMs for Clinical Evidence Summarization
- Zero- Shot Summarization with GPT- 3.5 and ChatGPT
- Medical Evidence Summary Evaluation
- Performance Metrics: ROUGE, METEOR, BLEU
- Comparison of LLM Summaries to Author Conclusions
- Abstract vs. Main Results Summarization in GPT Models
- Extractiveness and Novelty in LLM Summaries
- Human Evaluation of LLM Summaries
- Limitations of General- Purpose GPT for Summarization
- ChatGPT vs. GPT- 3.5 in Clinical Summarization
- Improving Summarization Quality with Advanced Prompting
- Coherence and Comprehensiveness of GPT Summaries
- Challenges in LLM- Based Medical Evidence Summarization
- Older GPT Models in Clinical Evidence Review

## A.55 Evaluating Computer Vision, Large Language, and Genome- Wide Association Models in a Limited Sized Patient Cohort for Pre- Operative Risk Stratification in Adult Spinal Deformity Surgery

- Automated Risk Stratification for ASD Surgery
- CNN vs. LLM in Predicting Surgery Complications
- GatorTron for Clinical Notes Analysis
- Genomic Data Integration with GWAS in Surgery Risk
- Performance Metrics: F1 Score and AUC in Risk Prediction
- Pulmonary, Neurological, and Sepsis Complications in ASD
- GWAS SNPs and ASD Surgery Risk
- Specificity and Predictive Value of LLM for Complications

- CNN Performance on Radiographs vs. LLM on Clinical Notes
- Privacy- Preserving Federated Learning for Rare Diseases
- Impact of Prompting on LLM Performance
- Ectoderm Differentiation Gene (LDB2) and ASD Risk
- Combining Imaging, Clinical, and Genomic Data for Risk Assessment
- Enhancing Risk Prediction with Advanced AI Models

## A.56 Evaluating capabilities of large language models - Performance of GPT-4 on surgical knowledge assessments

- Evaluation of ChatGPT in Medical QA for Surgery
- Performance of GPT- 4 in Surgical MCQs and Open- Ended Questions
- Accuracy and Reliability of ChatGPT Responses
- Analysis of Error Types in ChatGPT Medical Responses
- Repeat Query Variation and Response Consistency
- Assessment of ChatGPT Using SCORE and Data- B Question Banks
- Insights and Discrepancies in ChatGPT Medical Answers
- Potential for Fine- Tuned Models in Biomedical Data
- Opportunities for Synthetic Data in Privacy- Preserving Training
- Challenges in AI- Assisted Medical Education
- Impact of General Purpose LLMs in Surgical Learning
- Data Privacy and Governance in AI Medical Applications
- Variation in ChatGPT Answers on Repeated Queries

## A.57 Enhancing phenotype recognition in clinical notes using large language models - PhenoBCBERT and PhenoGPT

- Phenotype Recognition Models
- HPO Dictionary Limitation
- Transformer Decoder Models
- PhenoCBERT
- PhenoGPT
- Fine- Tuning Pre- trained LLMs
- Clinical Notes
- Pretrained BERT
- Bio+Clinical BERT
- GPT Models
- CHOP Database
- BiolarkGSC+ Dataset
- ID- 68 Dataset
- ICD- 10 Codes
- Phenotypic Concepts
- Model Performance (F1 Scores)
- Privacy and Anonymity Concerns

## A.58 Empowering Psychotherapy with Large Language Models - Cognitive Distortion Detection through Diagnosis of Thought Prompting

- LLMs in Psychology
- Zero- Shot Approaches
- Chain of Thought (CoT)
- Diagnosis of Thought (DoT)
- Subjectivity Assessment
- Contrastive Reasoning
- Schema Analysis
- Cognitive Distortions
- GPT- 3.5- Turbo
- Vicuna
- GPT- 4
- Therapist QA Dataset
- Annotated Patient Speech
- Cognitive Distortion Detection
- Model Performance Comparison
- Human Evaluation
- Dataset Volume and Quality
- Patient Demographics
- Privacy Concerns

## A.59 Doctor Versus Artificial Intelligence - Patient and Physician Evaluation of Large Language Model Responses to Rheumatology Patient Questions in a Cross-Sectional Study

- LLMs in Medical QA
- Real Patient Questions
- Comparison with Physician Responses
- Patient Perception of LLM Responses
- GPT- 4
- Bing Chat ”More Precise” Mode
- Alberta Rheumatology Website
- Dataset of Patient Questions
- Readability and Comprehensiveness Evaluation
- Physician vs AI Response Ratings
- Accuracy of AI Responses
- Patient and Physician Preferences
- AI Response Identification
- LLM Availability vs Physician Availability
- Lack of Structured Prompts
- No Fine- Tuned Model

## A.60 Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine

- Medical Reasoning
- Chain of Thought Prompting
- Diagnostic Reasoning Prompts
- Cognitive Processes
- Intuitive Reasoning
- Analytical Reasoning
- Bayesian Inference
- LLM Interpretability
- GPT- 3.5 (Davinci- 003)
- GPT- 4
- USMLE Questions Dataset
- NEJM Case Series
- CoT Baseline
- Analytical Reasoning CoT
- Differential Diagnosis CoT
- Performance Metrics
- Perception of Black- Box Nature
- XAI (Explainable AI) Integration
- Biomedical Corpora
- Model Reasoning vs. Post- Hoc Explanations

## A.61 Diagnostic accuracy of a large language model in rheumatology - comparison of physician and ChatGPT-4

- Diagnostic Accuracy
- ChatGPT- 4
- Physician Comparison
- Symptom Checker Suggestions
- Rheumatic and Musculoskeletal Diseases (RMDs)
- Top- 5 Differential Diagnosis
- Clinical Notes Dataset
- Ada Symptom Checker
- Comparison Metrics
- IR- Disease (IRD) Cases
- Top Diagnosis Accuracy
- Top- 3 Diagnosis Accuracy
- General- Purpose LLM
- Prompt Strategies
- Federated Learning Opportunities

## A.62 Computational screening of biomarkers and potential drugs for arthrofibrosis based on combination of sequencing and large nature language model

- Biomarker Gene Extraction
- ChatGPT
- Fibrosis Diseases
- Scientific Literature Analysis
- Gene Expression Chip Dataset
- GPR17 Regulator
- Therapeutic Targets
- Pharmaceutical Agents
- Data Mining
- Clinical Genetics
- NER (Named Entity Recognition)
- Lack of Prompt Strategies

## A.63 Complications Following Facelift and Neck Lift - Implementation and Assessment of Large Language Model and Artificial Intelligence (ChatGPT) Performance Across 16 Simulated Patient Presentations

- Post- Operative Care
- Plastic Surgery
- ChatGPT
- Differential Diagnosis
- Simulated Patient Presentations
- History- Taking
- Provisional Diagnosis
- Patient Disposition
- Home Interventions
- Red Flags
- GPT- 3
- Accuracy Metrics
- Comparison with Bard
- Limitations of LLM Usage

## A.64 Complications Following Body Contouring - Performance Validation of Bard, a Novel AI Large Language Model, in Triaging and Managing Postoperative Patient Concerns

- Bard LLM
- Differential Diagnosis Generation
- Post- Operative Concerns
- Prompt Evaluation

- Accuracy, Relevance, Safety
- Acute, Early, Late Complications
- History Taking
- Diagnosis Identification
- Patient Disposition
- First Aid and Treatments
- Red Flags
- Simulated Questions
- Accuracy Metrics
- Response Failure Rate
- Impact of Chat History
- No Fine- Tuned LLM
- Comparison with GPT

## A.65 Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions

- Ophthalmology
- ChatGPT
- GPT- 3.5
- Question Quality Evaluation
- Human vs. Chatbot Responses
- Instruction Prompt Engineering
- Few- Shot Prompting
- Online Forum Data
- Eye Care Forum
- Dataset of Questions and Answers
- Masked Panel Review
- Response Accuracy
- Inclusion of Incorrect Information
- Likelihood and Extent of Harm
- Deviation from Standards
- General- Purpose LLM
- Fine- Tuned Model

## A.66 CHiLL - Zero- shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models

- LLM Feature Extraction
- EHR Notes
- Flan- T5
- ICD Code Extraction
- Readmission Prediction
- Mortality Prediction

- Phenotype Prediction
- X- Ray Report Classification
- MIMIC Dataset
- MIMIC- CXR
- Downstream Model Training
- SGDClassifier
- AUROC Performance
- F1 Score
- Querying Instructions
- Feature Representation
- Open Questions
- Federated Learning Opportunities
- Close- Source LLMs
- Privacy Concerns

## A.67 ChatGPT sits the DFPH exam - large language model performance and potential to support public health learning

- LLM Performance Evaluation
- Public Health Exam
- Zero- Shot Learning
- Question Bank Analysis
- Examiner Assessment
- Pass Rate Comparison
- Research Methods Performance
- Answer Accuracy
- Insight Generation
- Model Improvement
- Non- Fine- Tuned Model
- Unique Insights

## A.68 Capacity for large language model chatbots to aid in orthopaedic management, research, and patient queries

- Evaluation of how accurately LLMs respond to orthopaedic- related questions
- Assessment of performance differences between ChatGPT, Bard, and BingAI
- Specific comparison of ChatGPT's superior performance relative to Bard and BingAI
- The three categories used for evaluating responses: bone physiology, referring physician, and patient queries
- Instances of responses being complete or incomplete
- Analysis of the chatbots' ability to provide appropriate clinical management advice
- Performance of chatbots in responding to patient- related queries
- Identification of outdated or incorrect information in chatbot responses
- Instances where Google Bard refused to answer questions

- Consistency of chatbot responses to patient queries
- Limitations due to lack of multimodal capabilities in chatbots
- Challenges in providing accurate information without multimodal inputs

## A.69 Can Large Language Models Safely Address Patient Questions Following Cataract Surgery

- Evaluation of the safety and quality of answers presented by patients
- Assessment of responses to patient questions collected from a telemedicine platform
- Evaluation of responses to cataract surgery- related questions by ChatGPT
- Application of GPT with zero- shot prompts to provide scientifically grounded answers
- Use of a dataset of 131 unique questions from patients following cataract surgery
- Assessment of response clarity and helpfulness from patient- generated questions
- Percentage of responses rated as ‘helpful’ or ‘somewhat helpful’
- Likelihood of harm based on response ratings
- Evaluation of responses against clinical or scientific consensus
- Recognition of the impact of minimal prompting on response quality
- Potential improvement with additional context in prompts

## A.70 Benchmarking large language models’ performances for myopia care a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard

- Evaluation of LLM performance in answering myopia-related questions
- Presentation of myopia questions to LLM chat front-ends as a typical user would
- Blind expert evaluation of responses from different LLMs
- Use of GPT-3.5, GPT-4.0, and Bard for answering myopia questions
- Definition of a benchmark of 31 commonly asked myopia questions
- Comparison of accuracy ratings between GPT-4.0, GPT-3.5, and Google Bard
- Assessment of response comprehensiveness across different LLMs
- Evaluation of self-correction capabilities in LLM responses
- Performance consistency across different myopia-related domains
- Potential for improvement with specialized biomedical models
- Impact of fine-tuning or specialized prompts on answer quality
- Possibility of federated learning to enhance myopia-related responses

## A.71 Battle of the (Chat)Bots - Comparing Large Language Models to Practice Guidelines for Transfusion-Associated Graft-Versus-Host Disease Prevention

- Investigation of LLM suitability for medical topics with uncertainty
- Use of an engineered prompt to list indications for blood component irradiation
- Presentation of prompts to chatbot versions of GPT and Bard
- Evaluation of LLMs in transfusion indication knowledge retrieval

- No fine-tuning applied to LLMs used in the study
- Use of the acronym in the prompt to test LLM comprehension
- Assessment of relevance in LLM-generated lists for TA-GVHD indications
- Comparison of chatbot responses to BSH guidelines on irradiation indications
- Identification of erroneous indications provided by chatbots
- Potential improvements with more detailed prompts or fine-tuned models
- Challenges in replicating medical community consensus within LLM outputs
- Variability in LLM responses related to foetuses and neonates

### A.72 AutoCriteria - a generalizable clinical trial eligibility criteria extraction system powered by large language models

- Introduction of a LLM-powered approach for eligibility identification in clinical trials
- Use of a specifically tuned prompt for identifying inclusion and exclusion criteria
- Evaluation of the approach on clinical trials from five disease groups
- Utilization of GPT-4 for data eligibility tasks in clinical studies
- Modeling of prompts based on information tokens for eligibility assessment
- Collection of clinical trial documents from ClinicalTrials.gov for nine diseases
- Design and calibration of prompts using selected clinical trials
- High extraction F1-scores across disease groups for eligibility criteria
- Best accuracy results achieved with Attribute + Value pairs in prompts
- Limitations due to small and specialized sample size
- Challenges with large prompt length for the extraction tasks
- Potential for applying this strategy to diagnosis criteria identification under FL

### A.73 Autocompletion of Chief Complaints in the Electronic Health Records using Large Language Models

- Investigation of LLMs' ability to identify and autocomplete patient chief complaints
- Comparison of LSTM and fine-tuned GPT models for chief complaint identification
- Application of prompt engineering to optimize GPT outputs for EHR notes
- Use of BioGPT and GPT-4 models in hospital EHR settings
- Utilization of de-identified EHR datasets from hospital patient reception
- Analysis of chief complaint reports from the GOUT-CC-2019-CORPUS and GOUT-CC-2020-CORPUS datasets
- BioGPT-Large's superior performance in generating chief complaints with low perplexity
- Significance of fine-tuning and prompt engineering in improving model performance
- Limitation of relying solely on perplexity and BERTScore for evaluating outputs
- Potential for federated learning to improve model scalability and privacy preservation

### A.74 Augmenting Large Language Models with Rules for Enhanced Domain-Specific Interactions The Case of Medical Diagnosis

- Introduction of a rule-augmented approach to GPT with physician-provided domain knowledge
- Use of a prompt with defined dialogue rules for GPT performance enhancement
- Embedding of domain knowledge using NLP techniques such as vectorization
- Definition of the domain space with specific question types and reply categories
- Application of the architecture in medical QA within a hospital setting
- Utilization of GPT-4.0 as the LLM in a multimodal architecture
- Integration capability of the LLM with vectorized data from vision models, IoT devices, and sensors
- Potential of the architecture to embed medical knowledge through a rule-based approach
- Need for implementation and expansion on the use of embeddings in model responses
- Opportunities for FL to enable decentralized training with vectorized data from sensors and computer vision models

### A.75 Assessment of a large language model's utility in helping pathology professionals answer general knowledge pathology questions

- Evaluation of the utility and performance of a large language model regarding accuracy, completeness, and its potential as a time-saving tool for pathologists and laboratory directors
- Utilize ChatGPT to answer generalized pathology questions that come to the pathology sector of the hospital
- Pathology, Medical QA, Hospital
- Used ChatGPT 4.0 without fine-tuning or prompt engineering
- Sourced data from an EHR pathology module where hospital care providers send questions to the pathology department
- Accuracy and completeness for the 61 questions was high, with 98
- Utilized a general-purpose LLM for answering the pathology questions, with potential for improvement through few-shot prompt engineering or a fine-tuned model for healthcare
- FL may increase the breadth and privacy of the knowledge of such models, if federated

### A.76 Artificial Intelligence to Automate Health Economic Modelling A Case Study to Evaluate the Potential Application of Large Language Models

- Automated Generation of Economic Models
- Engineering Prompts for Code Creation
- Application of GPT-4 in Health Economics

- Replication of Pharmacoeconomic Models
- Cost-Effectiveness Analysis in Drug Evaluation
- Accuracy and Error Rates in AI-Generated Models
- Few-Shot Prompting for Improved Script Generation
- Necessity of Human Intervention in AI Outputs
- Simplification of Complex Model Components
- Data Security and Privacy in Prompt Design
- Redaction and Handling of Sensitive Information
- Potential of Federated Learning in Model Training

## A.77 Artificial intelligence for health message generation an empirical study using a large language model (LLM) and prompt engineering

- Health Message Generation
- Patient Engagement in Self-Care
- Evaluation of LLMs for Health Communication
- Quality and Clarity of AI-Generated Messages
- Use of Bloom LLM Architecture
- Prompt Engineering for Message Optimization
- Zero-Shot Learning in LLMs
- Exploration of LLM Hyperparameters
- Twitter Data Utilization for Health Studies
- Comparison of AI vs. Human-Generated Content
- Limitations in Sample Size and Diversity
- Potential for Personalized Health Messaging through Federated Learning

## A.78 Analysis of large-language model versus human performance for genetics questions

- Evaluation of LLMs in Medical Genetics
- Comparison of AI and Human Accuracy in QA
- Use of ChatGPT for Genetics Questions
- Medical QA in Genetics and Inheritance
- Performance of LLMs on Memorization vs. Critical Thinking
- Variability in AI Responses
- Social Media as a Source of Medical Questions
- Accuracy of ChatGPT in Medical Contexts
- Impact of Model Fine-Tuning on Performance
- Clinical Applications of AI in Genetics
- Limitations of LLMs in Consistency
- Potential for Federated Learning in Clinical Genetics

## A.79 An AI Dietitian for Type 2 Diabetes Mellitus Management Based on Large Language and Image Recognition Models - Preclinical Concept Validation Study

- AI Dietitian for Type 2 Diabetes Support
- Integration of Multimodal LLM for Diet Recommendations
- Meal Suggestion and Calorie Estimation via AI
- Use of ImageNet and GPT for Health Applications
- Image-to-Text Transformation in Nutrition AI
- Application of WeChat for Dietitian AI
- Evaluation of AI in Registered Dietitian Exams
- Alignment of AI Recommendations with Best Practices
- Challenges in AI Food Recognition
- Professional Dietitian Feedback on AI Suggestions
- Lack of Fine-Tuning in AI Dietitian Models
- Potential for Federated Learning in Dietitian AI Scaling

## A.80 A study of generative large language model for medical research and healthcare

- Synthetic Medical Data Generation
- Training LLMs on Clinical Data
- Evaluation of GPT in Healthcare Settings
- Biomedical NLP Performance
- GatorTronGPT for Medical Text Tasks
- Large-Scale LLMs in Hospital Environments
- Physician Turing Test for AI Validation
- Use of Synthetic Data in Model Training
- Drug-Drug Interaction and Medical Relation Extraction
- Challenges with Large Model Sizes in Healthcare
- Improving AI Performance with Generated Data
- Privacy-Preserving Techniques in Federated Learning

## A.81 A platform for connecting social media data to domain-specific topics using large language models an application to student mental health

- LLM Embeddings in Topic Modeling
- Mental Health Theme Identification
- Use of Social Media Data for Mental Health Insights
- BERT-Based Topic Modeling (BERTopic)
- Extraction of Mental Health Triggers
- Containment Sentences in Topic Exploration
- Analysis of Unstructured and Noisy Data
- Qualitative Evaluation of Mental Health Models

- Use of University Subreddits as Data Source
- Challenges in Comparing Qualitative Results
- Customizing Models for Specific Communities
- Federated Learning for Community-Specific Mental Health Models

## A.82 A medical multimodal large language model for future pandemics

- Specialized LLM for Rare Diseases
- Multimodal Approach for Radiograph Interpretation
- Generation of Radiology Reports from Images
- Integration of Vision and Text Models
- Med-MLLM Architecture and Training
- Use of MIMIC-III and MIMIC-CXR Datasets
- Image Understanding and Clinical Phenotype Mapping
- Language Capabilities in Multiple Languages
- Performance Benchmarking of Diagnostic Models
- Reduction in Hallucination and Omission Rates
- Comparison with Existing Models (GPT-4, ChatGPT)
- Federated Learning for Scaling Medical AI Solutions

## A.83 A large language model for electronic health records

- Large-Scale Clinical LLM Training
- GatorTron Architecture and MegaTron Foundation
- Clinical Note-Based Language Model
- Evaluation Metrics: NER, MRE, STS, NLI, MQA
- Model Performance Comparison: GatorTron vs. BioBERT
- Scaling Model Size and Parameter Impact
- Clinical Concept Extraction Accuracy
- Resource Intensity and Hardware Requirements
- Benchmarking against SOTA Models
- Training Data: University of Florida Health Repository
- Modest Performance Gains over Existing Models
- Federated Learning for Privacy and Distributed Training

## A.84 Zero-shot Learning with Minimum Instruction to Extract Social Determinants and Family History from Clinical Notes using GPT Model

- Zero-Shot Approach for Clinical Data Extraction
- Evaluation of GPT-3.5 for Social Determinants
- Prompt Engineering with System and User Roles
- F1 Scores for Demographics, Social Determinants, and Family History
- Comparison with Fine-Tuned and Few-Shot Approaches
- Use of Zero-Shot Prompting in Clinical Notes

- Performance Metrics in Data Extraction Tasks
- Limitations of Non-Fine-Tuned GPT Models
- Challenges in Extracting Specific Clinical Information
- Benefits of Zero-Shot vs. Traditional NER Approaches
- Privacy Preservation through Federated Learning
- Improving Data Extraction with Advanced Techniques

## A.85 Unlocking the Power of EHRs Harnessing Unstructured Data for Machine Learning-based Outcome Predictions

- Incorporating Mental Health in Mortality Prediction
- Utilizing GPT for Extracting Mental Health Information
- Application of GPT Models to Electronic Health Records
- Evaluation of Stress, Anxiety, Depression, and PTSD
- Impact of Mental Health Data on Mortality Prediction Accuracy
- MIMIC-III Dataset for Clinical Data Analysis
- Linear Model for Mortality Prediction
- Potential for LLMs in Extracting Clinical Information
- Improvement in Prediction Models with Mental Health Data
- Shortcomings of Non-LLM Approaches
- Opportunities for Federated Learning in Mental Health Data
- Cross-Silo Data Integration for Mental Health Analysis

## A.86 TaughtNet Learning Multi-Task Biomedical Named Entity Recognition From Single-Task Teachers

- Limitations of Rule-Based NER Systems
- Knowledge Distillation for Entity Recognition
- Fine-Tuning Transformer Models on Clinical Text
- Multi-Task Learning for Entity Classification
- Biomedical Texts and Named Entity Recognition
- Combining Single-Task and Multi-Task Datasets
- Reduction of Model Size via Distillation
- Ensemble of Large Language Models for Knowledge Transfer
- Student-Teacher Architecture in NER Models
- Integration of Diverse Biomedical Datasets
- Evaluation of Model Performance using Cohen Kappa
- Federated Learning for Model Customization in NER

## A.87 Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models

- Synthetic Data Generation for Suicidal Ideation Detection
- Use of Generative Pretrained Transformers in Mental Health
- Integration of Synthetic and Real Data for NLP Models
- Fine-Tuning BERT Models with Mixed Data Sources

- Prompt Engineering for Data Generation
- Classification of Suicidal vs. Non-Suicidal Text
- Evaluation of Data Augmentation Techniques in NLP
- Application of GPT, Flan-T5, and Llama2 for Data Synthesis
- Impact of Synthetic Data on BERT Classification Performance
- Ethical Considerations in Generating Mental Health Data
- Opportunities for Federated Learning in Healthcare Data
- Privacy and Security in Synthetic Data for NLP

## A.88 Identification of Ancient Chinese Medical Prescriptions and Case Data Analysis Under Artificial Intelligence GPT Algorithm A Case Study of Song Dynasty Medical Literature

- Mining Medical Data from Archaeological Texts
- Named-Entity Recognition for Traditional Chinese Medicine
- Application of GPT in Historical Medical Records
- Contextual DAG Graphs for Medical Lexical Analysis
- Medical Fine-Tuning of GPT for Archaeological Texts
- Extraction of Ancient Medical Prescriptions
- Engineering Prompts for TCM Detection
- Utilization of Digitalized Chinese Medicine Treatises
- Identification of Prescriptions in Historical Documents
- Evaluation of GPT for Archaeological Medical Data
- Challenges in Describing GPT Utilization
- Federated Learning for Distributed Medical Record Analysis

## A.89 GPTFX - A Novel GPT-3 Based Framework for Mental Health Detection and Explanations

- Explainability of Large Language Models (LLMs)
- Fine-Tuning LLMs for Generating Explanations
- GPT-3 for Explainable AI (xAI) in Mental Health
- Alignment of LLMs with Explainability Goals
- Comparison of GPT and BERT for Mental Health Classification
- Use of Engineered Prompts for Explanation Generation
- Evaluation Metrics for Explainable AI (Rouge, BLEU)
- Interpersonal Risk Factors (IRF) Dataset for Mental Health
- Limitations of Paywalled Models in Research Replicability
- Federated Learning Opportunities for Explainable AI
- Classification of Mental Health Reports
- Embedding Techniques and Model Utilization

## A.90 Focusing on Needs A Chatbot-Based Emotion Regulation Tool for Adolescents

- GPT-2 for Adolescent Emotion Regulation
- Fine-Tuning GPT-2 on Chinese Adolescent Data
- Emotion-Based Dialog Guidance System
- Multi-Output Response Classification for Emotional Support
- Emotion-Weighted Response Selection
- Emotion Regulation Support in Chatbots
- Development and Use of CERD Dataset
- Keyword Extraction and Clustering with KeyBERT
- Evaluation of Chatbot Fluency and Supportiveness
- Challenges in Fine-Tuning and Model Evaluation
- Privacy Preservation in Federated Learning for Chatbots
- Scaling Emotion Regulation Tools in Distributed Settings

## A.91 Empowering Caregivers of Alzheimer's Disease and Related Dementias (ADRD) with a GPT-Powered Voice Assistant Leveraging Peer Insights from Social Media

- Voice Assistant for Alzheimer's Caregivers
- Fine-Tuning GPT for Dementia Support
- Voice-to-Text Interface for Caregiver Assistance
- Prompt Engineering for Dialog Flow in Voice Assistants
- Integration of Reddit Knowledge for Alzheimer's Support
- Customized GPT Model for Empathic Responses
- Use of GPTSimpleVectorIndex for Knowledge Management
- LangChain Library for GPT Optimization
- Caregiver Support Through Social Media Insights
- Evaluation of Answer Relevance and Completeness
- Bias and Curated Knowledge in Alzheimer's Information
- Federated Learning for Privacy-Preserving Model Scaling

## A.92 Early Risk Prediction of Depression Based on Social Media Posts in Arabic

- Arabic Tweet Translation for Mental Health Monitoring
- GPT-3 for Translating Arabic Tweets into English
- NLP Approaches for Depression Detection from Translated Text
- Feature Extraction with BoW and TF-IDF for Depression Classification
- Use of Traditional Classifiers for Analyzing Translated Tweets
- Detection of Depression Signs in Arabic Social Media
- Integration of GPT-3 in Multilingual Mental Health Monitoring
- Evaluation of Classification Performance with F1 Score
- Limitations of LLM Use in Non-Translation Tasks
- Opportunities for Federated Learning in Privacy-Preserving Mental Health Analysis

- Combining LLM Translation with Traditional NLP Methods
- Application of GPT-3 in Multilingual Text Processing

## A.93 ChatGPT for phenotypes extraction one model to rule them all

- LLMs for Phenotype Extraction from Clinical Reports
- Prompt Engineering for HPO Extraction with GPT-3
- Sentence-Level vs. Report-Level Phenotype Identification
- Evaluation of GPT-3 in Clinical Text Representation
- Accuracy of HPO Label and ID Extraction
- Comparison of GPT-3 with Traditional Phenotype Extraction Tools
- Use of GSC+ and ID-68 Datasets for Evaluation
- Challenges in HPO ID Recall with GPT-3
- PhenoBERT vs. GPT-3 for Clinical Phenotype Extraction
- Impact of Prompt Structure on Model Performance
- Limitations of Baseline GPT in Domain-Specific Knowledge
- Federated Learning for Distributed Phenotype Annotation

## A.94 Boosting Intelligent Diagnostic Process in Internet Hospital A Conversational-AI-Enhanced Framework

- AI Doctor for Initial Patient Triaging in Telemedicine
- Integration of LLMs in Internet Hospitals
- Delegation of Complex Cases to Human Physicians
- LLM Fine-Tuning with Clinical Data for Improved Diagnosis
- Simulation of Patient Load and Service Efficiency
- Reduction of Patient Waiting Times Using AI
- Improvement in Diagnostic Accuracy Over Time
- Challenges with Theoretical Framework and Simulations
- Potential Risks of LLM Hallucinations in Medical Contexts
- Federated Learning for Privacy-Preserving AI Doctors
- Personalization of AI Diagnosis Based on Local Epidemiology
- Theoretical Framework for AI Integration in Telemedicine

## A.95 AI-Integrated Single Platform to Enhance Personal Wellbeing

- AI-Driven Solution for Weight Management
- Integration of Computer Vision for Height and Weight Estimation
- Use of CNN for Food Choice Identification
- ChefTransformer for Meal Recommendations
- Calorie Counting and Exercise Scheduling
- Basal Metabolic Rate Calculation
- Application of MoveNet and Viola-Jones Estimators
- Lack of Fine-Tuning and Prompt Engineering Details

- Potential for Federated Learning in Personalized Health Solutions
- Limited Research on LLM Usage in Meal Recommendations
- Privacy Considerations in Health Data Management

### A.96 A GPT-2 Language Model for Biomedical Texts in Portuguese

- Portuguese Biomedical Text Processing
- Fine-Tuning GPT-2 for Medical Texts in Portuguese
- Training on Biomedical Corpora
- Model Evaluation for Fall Incident Classification
- Comparison with General GPT-2, BioBERT, and Other Embeddings
- Use of AutoTokenizer and Specific Training Parameters
- Performance Metrics: F1 Score of 0.9009
- Dataset Used: WMT16 Biomedical Corpus and Fall Incident Reports
- Shortcomings: Limited Fine-Tuning Data Size
- Opportunities for Federated Learning: Privacy and Data Collection

## Appendix B Federated Learning Paper Codes

### B.1 Contribution-Aware Federated Learning for Smart Healthcare

- Contribution-Aware Federated Learning
- Smart Healthcare Data Distribution Issues
- Fair and Explainable Participant Contribution Evaluation
- FL Model Aggregation Optimization
- Leukemia, Biopsy, and Pneumonia Data Sets
- Contribution Evaluation (CE) Server for Performance Optimization
- Subset Selection for Optimal Model Aggregation
- Potential Processing Overhead for Large Models
- Opportunities for LLMs with CAREFL Strategy
- Training Strategies and non-IID Data Considerations

### B.2 Medical report generation based on multimodal federated learning

- Multimodal Federated Learning for Medical Imaging
- Medical Image Reporting Privacy Concerns
- ResNet-101 and DenseNet-121 for Feature Extraction
- CNN and Transformer-Based Architecture
- FedAvg, FedProx, and FedSW Aggregation
- FedSW Threshold-Based Client Contribution
- IU_Xray Dataset for Training
- Non-IID Data Considerations
- BLEU-1 Metric for Report Quality

- Communication Overhead in Multimodal FL
- Potential Limitations of FedSW
- Scalability in Large Model FL
- Opportunities with LLMs and FedSW
- Medical Imaging and Decision Support Domain

## B.3 FedICU - a federated learning model for reducing the medication prescription errors in intensive care units

- FL for ICU Prescription Error Evaluation
- Patient Privacy in ICU
- Eavesdropping Attack Risks
- Paillier Homomorphic Encryption (PHE)
- ICU, Data Security, and Encryption Domain
- MLP, Logistic Regression, Simple Neural Network Models
- Custom FL Implementation
- FedAvg Aggregation
- MIMIC-IV Dataset
- Key-Value Feature Encoding
- PHE-Enhanced Model Accuracy
- Impact of Encryption on Model Performance
- Differential Privacy with Low Magnitude Parameters
- Lack of Focus on FL Challenges
- Tabular Data Encoding Instead of Textual Data
- Encryption in LLM Training for FL

## B.4 An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals

- Multi-Hospital Clinical Risk Prediction
- Patient Privacy in FL
- Data Distribution Drift
- Adaptive Federated Learning Framework
- Feature Separation by Clinical Outcome Relationships
- Training Strategies Domain
- Stability-Specific Network
- Shared Stability Network
- Domain-Specific Network
- EHR Feature Vectors
- eICU Dataset
- Stable vs. Domain-Specific Features
- Superiority of Proposed Method
- Focus on Health Parameters Over FL
- Relational Data Limitation
- Fine-Tuning LLMs with Split Feature-Sets

## B.5 Exploring Federated Learning for Speech-based Parkinson's Disease Detection

- Centralized Dementia Risk Models
- Data Privacy Concerns
- Scalability Issues in Centralized Models
- Federated Learning Migration
- Federation of SOTA Parkinson Disease Model
- Dementia Care Domain
- Speech Detection Domain
- Adversarial Auto-Encoder Architecture
- CNN-Based Encoding
- FedAvg, FedAvgM, FedAdam Algorithms
- Small Client Distribution
- MDVR-KCL Dataset
- Voice Recording Data
- Realistic Speech Data Collection
- Limited Sample Size
- Expansion of Clients or Synthetic Data

## B.6 FedTherapist - Mental Health Monitoring with User-Generated Linguistic Expressions on Smartphones via Federated Learning

- Sensitive Nature of Mental Health Data
- Linguistic Expression of Symptoms
- Federated Learning for Mental Health Data
- FedTherapist Model
- Context-Aware Language Learning (CALL) Methodology
- Mental Health Domain
- Psychology Domain
- Fixed-BERT + MLP Architecture
- End-to-End BERT + MLP Architecture
- LLM (LLaMa-7B)
- LoRA for Weight Fine-Tuning
- FedAvg for Weight Aggregation
- Text and Speech Data Collection
- PHQ-9 Test for Depression
- Impact of Text Data on Model Performance
- DistilBERT and Mobile-BERT Performance
- Fixed BERT for Low Overhead
- LLM Resource Constraints
- Dialog Comprehension Fine-Tuning
- Exploration of Smaller LLMs
- Potential of Text-Only Approaches
- CALL Approach for Healthcare Applications

## B.7 Scaling-up medical vision-and-language representation learning with federated learning

- Sensitive Information in Vision and Language Pairs
- Scalable Training Need
- FedMedVLP Model
- Cross-Modal Attention
- Medical Imaging Domain
- Training Strategies
- BERT-BASE Backbone
- Task Evaluation
- ROCO Dataset
- VQA-RAD Dataset
- SLAKE Dataset
- MedVQA-2019 Dataset
- COVID-Fed Dataset
- Superior Performance of FedMedVLP
- Decentralized Approach
- Limited Exploration of FL
- Client Configuration Exploration
- Robustness to Non-IID Data
- Aggregation Approaches Impact

## B.8 Quantitative risk analysis of treatment plans for patients with tumor by mining historical similar patients from electronic health records using federated learning

- Knowledge Asymmetry in Cancer Treatment
- Changing Treatment Parameters
- Patient Heterogeneity
- Limited Access to EHRs
- Federated Learning for EHR Access
- Quantitative Risk Analysis
- Historical Patient Data Mining
- Oncology Domain
- Data Mining in Healthcare
- Feature Extraction with SVM-RFE
- Treatment Selection with DeepLIFT
- Patient Clustering with Cosine Similarity
- Federated Learning Architecture
- Non-IID Data in Medical Treatment
- State-Relevant Treatment Data
- Survival Prediction Accuracy
- Simple FL Validation
- Alternative Aggregation Approaches
- LLMs as Feature Extractors

- Transformer-based Patient Clustering
- Optimization of Feature Extraction

## B.9 Federated learning for secure development of AI models for Parkinson's disease detection using speech from different languages

- Cross-institutional Data Privacy
- Federated Learning Implementation
- Speech Data from Multiple Languages
- Parkinson's Disease Detection
- Real-world Language Corpora
- Wave2Vec Audio Embeddings
- Model Aggregation using FedAvg
- Non-IID Data Across Institutions
- Performance of Federated Models vs. Centralized Models
- Privacy Preservation vs. Model Accuracy
- Challenges in Small Sample Sizes
- Multilingual Speech Processing
- Institutional Data Security

## B.10 Cloud-IIoT-Based Electronic Health Record Privacy-Preserving by CNN and Blockchain-Enabled Federated Learning

- Electronic Health Record (EHR) Privacy
- Securing Distributed Healthcare Data
- Deep Learning for Data Privacy
- Blockchain for Privacy Preservation
- Convolutional Neural Networks (CNN) for Classification
- Cryptography-Based Federated Learning
- Detection of Abnormal Users
- Vector Representation of Healthcare Data
- Integration of Deep Learning and Blockchain
- Data Privacy and Security in Healthcare
- Performance in Malicious Activity Detection
- Scalability and Privacy of Data Handling
- Privacy-Preserving Healthcare Systems
- No Definition for Abnormal Users
- Lack of Detailed Data and Client Information

## B.11 Hyper-Graph Attention Based Federated Learning Methods for Use in Mental Health Detection

- Poor Clinical Performance of Deep Neural Networks
- Limited Dataset Challenges

- Federated Learning for Mental Health
- Structural Hypergraph for Text Representation
- Emotional Lexicon Embeddings
- Directed Graph Representation of Text
- WordNet Embedding for Sentiment Analysis
- Feed Forward Network for Text Classification
- LSTM and Bidirectional LSTM for Sequential Data
- PHQ-9 for Depression Symptom Evaluation
- Amazon MTurk for Data Collection
- Attention-Based Models in Clinical Data
- Performance Comparison of LSTM Approaches
- Challenges with MTurk Data Representativeness
- Opportunities for Advanced LLM Integration

## B.12 Multi-Site Clinical Federated Learning Using Recursive and Attentive Models and NVFlare

- Integration of Federated Learning and NLP
- Data Privacy and Regulatory Compliance
- BERT Training Pipeline for Federated Fine-Tuning
- NVFlare for Secure Communication
- Federated Learning with BERT and LSTM
- Local Training and Fine-Tuning of Models
- IID and Non-IID Aggregation with FedAvg
- Electronic Health Records Data Analysis
- Performance Comparison: BERT vs. LSTM
- Sample Size Effects on Model Performance
- Opportunities for NLP with Federated Learning
- Exploration of LLMs for Larger Sample Sizes

## B.13 Data Quality Aware Hierarchical Federated Reinforcement Learning Framework for Dynamic Treatment Regimes

- Dynamic Treatment Regimes (DTRs) Challenges
- Privacy Concerns in Multi-Hospital Data
- Data Quality and Heterogeneity Issues
- Global Data Quality-Aware DTR
- Hierarchical Federated Reinforcement Learning
- Data Quality Quantification via Immediate Health Status
- Offline Actor-Critic Reinforcement Learning
- Online RL-Based Clustering Scheme
- Markov Decision Process for Treatment Optimization
- LSTM for Patient State Representation
- Aggregation Algorithms: FedAvg, PerFedAvg, IFCA
- eICU and MIMIC-II Datasets Utilization

- Framework Performance and F1 Score Issues
- Mapping DTRs to Data and Feature Extraction
- Opportunities for RL Integration in Federated Learning for LLMs

## B.14 FedCPC An Effective Federated Contrastive Learning Method for Privacy Preserving Early-Stage Alzheimers Speech Detection

- Early-Stage Alzheimer's Disease Detection
- Speech-Based Detection Methods
- Data Privacy Risks in Medical Institutions
- Federated Learning for Privacy Preservation
- Performance Reduction in Federated Learning
- Federated Contrastive Pre-Training (FedCPC)
- Enhanced Representation Learning
- CNN for Speech Data Encoding
- Downstream Models: AST, CNN, CNN-LSTM
- FedAvg Aggregation Method
- MFCCs for Speech Data Representation
- Dataset: NCMMSC AD Recognition Challenge
- Speech Clip Duration and Labeling
- Model Performance Comparison
- Limitations of MFCCs and AST Model
- Training and Fine-tuning as separate, yet federated steps

## B.15 Fair and Privacy-Preserving Alzheimer's Disease Diagnosis Based on Spontaneous Speech Analysis via Federated Learning

- Automatic Speech Analysis for Alzheimer's Diagnosis
- Deep Learning Techniques in Speech Processing
- Centralized Learning Frameworks for Speech Data
- Privacy Concerns with Centralized Data Storage
- Federated Learning-Based Approach for AD Diagnosis
- MFCCs and Linguistic Features Extraction
- Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) Models
- Feature Fusion: Acoustic and Linguistic Embeddings
- ADReSS Challenge Dataset for Evaluation
- Spoken Picture Descriptions Task
- Use of Acoustic and Metalinguistic Features
- Lack of Textual Information Utilized
- Potential for LLM Integration in Speech Analysis
- Flwr Framework for Federated Learning

## B.16 Privacy-preserving Speech-based Depression Diagnosis via Federated Learning

- Diagnosis of Depression Through Speech Analysis
- Deep Learning Models for Mental Health
- Federated Learning (FL) for Privacy Preservation
- Differential Privacy in Federated Learning
- FedAvg, FedMedian, FedTrimmedMean, Krum Aggregation Algorithms
- Cross-Silo and Cross-Device Federated Learning
- Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) Dataset
- Acoustic Features for Depression Detection
- Mel Frequency Cepstral Coefficients (MFCCs)
- Patient Health Questionnaire-8 (PHQ-8) Metrics
- Classification of Depressed vs. Non-Depressed Participants
- Limited Content Evaluation Based on Acoustic Features
- Potential for LLMs in Richer Feature Extraction
- Baselines for LLM Applications in e-Health

## B.17 BVFLEMR an integrated federated learning and blockchain technology for cloud-based medical records recommendation system

- Decentralized Private Storage for Electronic Health Records
- Blockchain-Based Storage Using Hyperledger Fabric
- Continuous Monitoring and Tracking of EHR Updates
- Collaborative Learning on Blockchain-Stored EHR Data
- LightGBM and N-Gram Models for Recommendation
- FedAvg Aggregation Approach
- Patient-Doctor Interaction Reviews and Ratings
- Sentiment and Emotion Analysis for Treatment Recommendations
- Poor Methodology Description
- Undisclosed Database and Data Source Issues
- Preliminary Results and Lack of Detailed Metrics
- Potential for LLMs in Enhanced Vector Encoding
- Framework Description Without Generalizable Conclusions

## B.18 Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting

- Patient Violence Risk in Psychiatric Settings
- Privacy Concerns in Healthcare Data
- Federated Learning for Clinical NLP
- Violence Risk Assessment Using EHR Notes
- Comparison of Federated vs. Local and Centralized Models
- Doc2Vec Vectorization of Clinical Notes
- FedAvg Aggregation in FL

- Limited Data Sample Size
- Opportunity for Training Doc2Vec in FL
- Potential of LLMs for Enhanced Feature Extraction
- Simulated Clients for Federated Learning
- Lack of Significant Contribution from Current FL Methods

## B.19 A Federated Learning Based Privacy-Preserving Smart Healthcare System

- Dementia Detection Privacy Challenges
- Federated Learning for Privacy Preservation
- ADDetector Tool for Alzheimer's Detection
- Acoustic and Linguistic Feature Extraction
- TextCNN for Topic Selection
- Differential Privacy in Federated Learning
- FedAvg Aggregation Strategy
- Small Number of Clients in Experiments
- Potential for Improved Results with Larger Models
- LLMs for Enhanced Linguistic Classification
- Customized Feedback Layer for Patients
- Reducing Communication Overhead with Differential Privacy

## B.20 Evaluating Efficiency and Effectiveness of Federated Learning Approaches in Knowledge Extraction Tasks

- Sensitive Data Extraction and Anonymization
- Federated Learning (FL) for Privacy
- Natural Language Processing (NLP)
- Named Entity Recognition (NER)
- LSTM Architecture
- Categorical Cross-Entropy Loss
- Federated Averaging (FedAvg)
- Mental Health Data
- Sensitive vs. Insensitive Phrases
- Data Privacy and Security
- Accuracy Decay and Latency
- Challenges in Anonymization
- Data Volume Complexity
- Future Research Needs
- Feature Extraction Opportunities

## B.21 Federated learning of predictive models from federated Electronic Health Records

- Computational Efficiency in Healthcare
- Privacy-Aware Solutions

- Decentralized Optimization Framework
- Binary Supervised Classification
- Cluster Primal Dual Splitting (cPDS) Algorithm
- Large-Scale Soft-Margin l1-Regularized SVM
- Decentralized Collaboration
- Electronic Health Records (EHRs)
- Cardiac Event Prediction
- Feature Vectorization (215-Dimension)
- Communication Overhead
- Prediction Accuracy (AUC)
- Feature Importance for Hospitalizations
- Validation Across Different Datasets
- Performance Variability with Data Complexity
- Decentralized Federated Learning Opportunities
- LLMs for Feature Extraction

## B.22 Fold-stratified cross-validation for unbiased and privacy-preserving federated learning

- Data Leakage in Federated Learning
- Centralized Validation Set Risks
- Privacy-Preserving Validation Techniques
- Fold-Stratified Cross-Validation (FSCV)
- Data Duplication and Deduplication
- XGBoost Model
- Synthetic Dataset for Validation
- MIMIC-III Dataset
- ICU Patient Records
- Interfold Deduplication
- Covariate Identification for Privacy
- Low Computational Overhead
- Privacy Preservation Focus
- Client-Based Validation Strategies

## B.23 Federated Learning on Clinical Benchmark Data -Performance Assessment

- Reliability and Performance of Federated Learning (FL)
- Benchmarking FL
- Clinical Benchmark Dataset
- MNIST Dataset
- MIMIC-III Dataset
- ECG Dataset
- Client Setup (3 Clients for MIMIC and ECG; 10 Clients for MNIST)
- Epochs and Batch Sizes
- Basic and Imbalanced Data Experiments

- Model Architectures (ANN for MNIST, LSTM for MIMIC-III, CNN for ECG)
- Area Under the Receiver Operating Characteristic Curve (AUROC)
- F1-Score Performance
- Handling Imbalanced and Skewed Data
- Comparative Performance Across Datasets
- Limitations of FedAvg Aggregation
- Non-Textual Data Focus
- Multimodal LLM Opportunities

## B.24 Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning

- Adverse Drug Reaction (ADR) Prediction
- Challenges with Electronic Health Data
- Federated Learning Framework for ADR
- Local Model Aggregation Methods
- Drug Safety Domain
- Training Strategies in Federated Learning
- Parameter Metadata for Class Ratios
- FedAvg and FedADR Algorithms
- Limited IBM MarketScan Explorys Claims-EMR Data Set (LCED)
- Patient-Level Features (Demographics, Diagnostic Codes, etc.)
- Centralized vs. Federated Model Privacy
- Non-Textual Data Focus
- Potential for Federated Learning with Text Data

## B.25 Federated Learning-Based Secure Electronic Health Record Sharing Scheme in Medical Informatics

- Vulnerability of EHRs to Cyberattacks
- Federated Learning-based EHR Sharing Scheme
- Decentralized Learning and Data Privacy
- Convolutional Neural Network (CNN) for EHRs
- InterPlanetary File System (IPFS) for Data Storage
- Blockchain for Data Integrity
- Smart Contracts for Data Access Control
- Private vs. Public IPFS Storage
- FedAvg Aggregation Algorithm
- X-ray Images for COVID Detection
- Training and Model Storage on IPFS
- Convergence and Performance Compared to Centralized Models
- Privacy and Data Integrity Using Blockchain
- Non-Textual Data Focus
- Applications of IPFS and Smart Contracts in LLMs

## B.26 Federated Learning for Privacy Preservation of Healthcare Data From Smartphone-Based Side-Channel Attacks

- Side-Channel Attacks on Smartphones
- Vibration Exploitation for Sensitive Information
- Federated Learning for Privacy Protection
- Neural Network for Text Inference
- Distributed Training Across 10 Clients
- Magnetometer and Gyroscopic Data
- User Postures During Data Collection
- Dataset Characteristics (Sensors, CSV Format)
- Prediction of Keystrokes from Metadata
- Challenges with Small Client and Sample Size
- Unclear Relevance to Health Data
- Limited Application to eHealth and LLMs

## B.27 Federated Learning-Inspired User Personality Prediction Using Sentiment Analysis and Topic Preference

- Identifying Personality Traits from Topic Engagement
- Sentiment Analysis for Personality Prediction
- Psychology and Classification Domain
- Emotion-Categorized Chinese Weibo Comments
- Use of LSTMs, GRUs, and BERT for Classification
- Training Data from ChineseNlpCorpus and Weibo
- Challenges in Confirming Use of Federated Learning
- Potential for FL in Analyzing Patient Traits

## B.28 A Resource-Constrained and Privacy-Preserving Edge-Computing-Enabled Clinical Decision System - A Federated Reinforcement Learning Approach

- Security in Federated Deep Networks
- Integration of MEC and Software-Defined Networking
- Use of Double Deep Q-Network (DDQN) and Fully Decentralized Federated Framework (FDFF)
- Additively Homomorphic Encryption for Privacy
- Medical IoT Application
- Personal Health Information (PHI) and Historical EMRs
- Performance in Clinical Decision Support Systems (CDSSs)
- Challenges in Practical Implementation and Computational Overhead
- Opportunities for Federated Learning with LLMs and Reinforcement Learning