Project Concept Note

Title: AI-Assisted Health Care Database: Text-to-Text LLM-Based Screening
Principal Investigator (PI): Dr. Nikku Yadav
Co-Investigators: Dr. Durga Toshniwal, Mr. Abhinav Bahuguna, Ms. Akanksha Uniyal
Proposed Funding Agency: Biotechnology Industry Research Assistance Council (BIRAC)
Domain Area: Artificial Intelligence in Healthcare, Data Analytics, Clinical Informatics

1. Background and Rationale

India faces a significant challenge in integrating large volumes of unstructured clinical data—ranging from electronic health records (EHRs), diagnostic reports, physician notes, and patient feedback—into actionable healthcare insights. Current systems rely heavily on structured data and lack the ability to understand contextual, text-based medical information.

The advent of Large Language Models (LLMs) such as GPT and domain-specific medical transformers presents an opportunity to revolutionize healthcare data interpretation. By applying text-to-text AI models, we can extract, summarize, and classify health information to assist in disease screening, triaging, and decision support.

This project proposes the development of an AI-assisted healthcare database that integrates LLM-based text-to-text screening to automate the extraction and interpretation of clinical narratives, improving early detection, documentation accuracy, and patient management.

2. Objectives

1. To develop an integrated healthcare database capable of processing and storing multimodal health data (EHRs, text reports, prescriptions).
2. To design and train a Text-to-Text LLM module for disease screening, risk prediction, and clinical insight generation.
3. To validate the AI system using real-world health data (anonymized) from hospital partners.
4. To develop a user-friendly dashboard for clinicians and researchers for interactive queries and report generation.

3. Methodology and Work Plan

Phase 1: Requirement Analysis and Data Collection (Months 1–3)

- Identify data sources: EHRs, clinical notes, lab reports, and discharge summaries.
- Define data architecture and privacy framework following ICMR and HIPAA guidelines.
- Annotate sample datasets for training and validation.

Phase 2: LLM Integration and Model Development (Months 4–10)

- Fine-tune an open-source LLM (e.g., LLaMA, Mistral, or BioGPT) on medical text datasets.

- Develop a text-to-text interface capable of screening inputs like "patient complaint + test results" to generate diagnostic probabilities and suggestions.
- Implement prompt engineering for differential diagnosis, red-flag identification, and medication checks.

Phase 3: Database and Interface Development (Months 11–14)

- Design a secure healthcare database integrated with the LLM output.
- Build APIs for hospital EMR systems and create a web-based clinician dashboard.
- Enable searchable, query-based summaries for doctors and researchers.

Phase 4: Validation and Testing (Months 15–18)

- Evaluate the model on predefined clinical conditions (e.g., diabetes, hypertension, respiratory infections).
- Assess accuracy, recall, and interpretability.
- Conduct usability testing with healthcare professionals.

. Expected Outcomes

- A functional prototype of an AI-assisted healthcare database integrated with an LLM screening engine.
- Improved efficiency of medical data interpretation and screening accuracy for common diseases.
- Scalable architecture for integration with hospital information systems.
- A foundation for further research in AI-driven clinical decision support and precision health.

5 Novelty and Innovation

- First-of-its-kind Text-to-Text LLM-based screening framework for Indian healthcare settings.
- Combines natural language understanding with medical reasoning.
- Provides context-aware, explainable screening outputs rather than simple classification.
- Enables real-time data augmentation and multilingual processing (English, Hindi, regional languages).

6. Societal and Commercial Impact

- Societal: Enables faster and more equitable healthcare screening, particularly in semi-urban and rural health setups.
- Clinical: Reduces clinician workload through automated report summaries and diagnostic support.
- Commercial: Potential to scale as an AI HealthTech platform for hospitals, diagnostics centers, and telemedicine services.

7. Team Expertise

- **Dr. Nikku Yadav (PI):** Expertise in biomedical data analytics and clinical informatics.
- **Dr. Durga Toshniwal:** Specialization in AI, data mining, and machine learning applications in healthcare.
- **Mr. Abhinav Bahuguna:** Experience in health data systems, software development, and AI deployment.
- **Ms. Akanksha Uniyal:** Skilled in NLP model development, data preprocessing, and model validation.

8. Estimated Budget (Indicative)

Category	Estimated Cost (INR)
Personnel & Research Assistants	12,00,000
Data Procurement & Annotation	6,00,000
Computing & Cloud Resources	10,00,000
Software Development & Integration	8,00,000
Dissemination & Publication	2,00,000
Total (Approx.)	**38,00,000**

9. Deliverables

1. Functional AI-based healthcare database and screening prototype.
2. Technical documentation and validation reports.
3. Publication(s) in peer-reviewed journals.