DIWAKAR MAHAJAN

New York, NY | E-mail: dm3084@columbia.edu | Mobile: (646) 667-8524 | Google Scholar | Linkedin

WORK EXPERIENCE

IBM Research, Yorktown Heights, NY Feb 2016 – Current

Senior Machine Learning Research Engineer Oct 2020 – Current

LLM–BioFM Integration for Generative Biomedical Reasoning with Agentic Workflows

- Developed BIOVERSE, a modular framework aligning protein and molecular foundation models with LLMs via projection layers and LoRA adapters, enabling zero-shot multimodal reasoning for cell types, molecules, and protein functions.
- Designed contrastive and autoregressive alignment strategies to unify cross-modal embeddings and improve generative biomedical interpretation and developed scalable training pipelines for multimodal LLM finetuning using PyTorch, Hugging Face and ClearML.
- Developed an agentic scientific workflow with RAG components for Cleveland Clinic researchers to screen antibody, nanobody, and TCR candidates using OpenWebUI, Docker-based microservices, and fully local models.

Cross-Modal Foundation Models for Biological Data and Molecular Prediction

- Developed MMELON, a multi-view molecular foundation model combining graph, image, and text embeddings for molecule–target and property prediction.
- Improved performance across 120+ drug discovery benchmarks and applied the model to GPCR target screening for Alzheimer's therapeutics.
- Designed novel topological pretraining objectives and attention-based fusion mechanisms for interpretable multimodal representation learning.

Clinical LLM Benchmarking for EHR Understanding

- Conducted one of the largest evaluations of clinical NLP models, comparing 12 language models (220M–175B params) on tasks requiring reasoning over EHR notes.
- Demonstrated that specialized clinical models outperform large general-purpose LLMs, with limited annotated data or sizes.
- Trained clinical T5 models from scratch on MIMIC-III/IV, showing that in-domain clinical pretraining yields higher accuracy and better parameter efficiency than larger web-trained LLMs.

Clinical Summarization & Question Answering

- Led and scored 2nd place in the external NLP Challenge on clinical abstractive summarization (MEDIQA 2021).
- Developed novel methodologies for improving factual correctness of machine generated clinical summaries.
- Developed and open-sourced a clinically grounded Question-Answering dataset on structured EHR data.

Event Extraction from Clinical Text

- Developed joint learning techniques for extracting medications and their multi-dimensional orthogonal context.
- Utilized the extracted context for event identification, classification and time-line generation.
- Developed Medication Instruction parser to normalize and calculate daily dosage values for prescribed medications.

Advisory Machine Learning Research Engineer Jun 2017 – Sep 2020

Multi-task Clinical Representation Learning for Similarity & Drug Interaction

- Led and won the external NLP Challenge on clinical semantic textual similarity (n2c2 2019).
- Won Drug–Drug Interaction Extraction challenges at TAC 2018 & 2019.
- Developed intermediate multi-task pretraining techniques enabling effective transfer learning for low-resource clinical NLP tasks
- Developed BERT-based models to identify key relations such as Adverse Reactions and Drug–Drug Interactions.
- Strategically leveraged external supervised and unsupervised datasets to address scarcity of annotated clinical data.

Machine Learning Research Engineer Feb 2016 – May 2017

Entity Extraction, Relation Classification & Normalization from Clinical & Biomedical Text

- Won external NLP Challenge on Adverse Drug Event Extraction (TAC 2017).
- Developed hybrid methodologies for extracting disjoint biomedical entities by reducing the label space.
- Employed varied techniques BiLSTM-CRF, BM25 with BERT and Siamese networks, KBE etc.

Research Software Developer, Tata Innovation Labs, India Mar 2011 – Jul 2014

- Applied Topic Modelling on scientific publications and patent database to identify most popular topics of a year.
- Built event extraction models for social media analytics using cascaded CRFs.
- Designed semantic search enhancements for enterprise search systems; trained teams in Apache SOLR integration.

HONORS & AWARDS

IBM Outstanding Technical Achievement Award, 2020.
2nd Place Award – Radiology Report Summarization Challenge – **MediQA at NAACL 2021**.
1st Place Award – Clinical Semantic Textual Similarity – **National NLP Clinical Challenge 2019**.
1st Place Award – Drug-Drug Interaction Extraction Challenge – **Text Analytics Conference 2019**.
2nd Place Award – Drug-Drug Interaction Extraction Challenge – **Text Analytics Conference 2018**.
2nd Place Award – Clinical Adverse Drug Event Extraction – **Text Analytics Conference 2017**.
2nd Place Award – NTT DoCoMo Challenge-Multimedia Grand Challenge – **ACM Multimedia Conference 2012.**

EDUCATION

Columbia University, New York, USA 2016
Master of Science, Computer Science (Natural Language Processing)
Guru Gobind Singh Indraprastha University, India 2010
Bachelor of Technology, Information Technology

SKILLS

Programming Languages:	Python, Java, C++, R, MATLAB
NLP/ML/LLM:	Deep learning, LLMs, foundation model pretraining, LLM finetuning, multimodal modeling, cross-modal alignment, contrastive learning, NLP pipelines, Retrieval-Augmented Generation, Eval, prompt engineering, LLM Agentic Workflows
Frameworks/Libraries:	PyTorch, Tensorflow, Lightning, Transformers, Hugging Face libraries, SpaCy, Scikit, SciPy, NumPy, Mallet, Workflow orchestration (OpenWebUI, Docker microservices)

SELECTED PUBLICATIONS & PATENTS (See the Complete list here)

- BioVERSE: A Modular Framework for Integrating Biomedical Modalities with Language Models in Precision Medicine. *Conference on Intelligent Systems for Molecular Biology 2025*
- Capturing Individual-level Social Determinants from Clinical Text. *AMIA Annual Symposium Proceedings 2024*
- Multimodal Molecular Representation Learning for Small Molecule Drug Discovery-Pretraining and Early Fusion Architectures. *American Chemical Society (ACS) Fall Meeting 2024*
- Clinical natural language processing for secondary uses. *Journal of biomedical informatics. 2024*
- Artificial intelligence-assisted non-pharmaceutical intervention data curation. 2024 (US Patent 12,062,454)
- MISMATCH: Fine-grained Evaluation of Machine-generated Text with Mismatch Error Types. *Findings of the Association for Computational Linguistics: ACL 2023*
- Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics 2023*
- Extracting medication changes in clinical narratives using pre-trained language models. *Journal of biomedical informatics 2023*
- Do We Still Need Clinical Language Models? *Proceedings of Machine Learning Research 2023*
- Auto-generating ground truth on clinical text by leveraging structured electronic health record data. 2023 (US Patent 11,782,942)
- Towards generalizable methods for automating risk score calculation. *Proceedings of the 21st Workshop on Biomedical Language Processing ACL 2022*
- Reducing physicians' cognitive load during chart review: a problem-oriented summary of the patient electronic record. *AMIA Annual Symposium Proceedings 2022*
- Evaluating Social Determinants of Health in Clinical Communications Data. *American Medical Informatics Association 2021*
- AI-assisted tracking of worldwide non-pharmaceutical interventions for COVID-19. *Nature Scientific data 2021*
- An Exploration of Reasons Behind Drug De-escalation and Discontinuation Events in Clinical Notes. *AMIA 2021*
- SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents. *ACL 2021*
- emrKBQA: A Clinical Knowledge-Base Question Answering Dataset. *Proceedings of the 20th Workshop on Biomedical Language Processing. ACL 2021*
- IBMResearch at MEDIQA 2021: toward improving factual correctness of radiology report abstractive summarization. *Proceedings of the 20th Workshop on Biomedical Language Processing. ACL 2021*
- Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *Machine Learning for Health (ML4H) at NeurIPS 2020*
- Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning. *JMIR Medical Informatics 2020*