Mujahid Ali Quidwai

(224) 322-6446 • quidwaiali@gmail.com • LinkedIn • Google Scholar • Portfolio

PROFESSIONAL SUMMARY

AI Systems Engineer with production deployments in clinical oncology — systems actively used by Mount Sinai oncologists for treatment decisions. Architect of multi-agent genomic curation (OncoCITE: 97.8% precision; Submitted to Nature Cancer) and GPU-accelerated pipelines (92.3% time reduction enabling same-day tumor board). Full-stack AI expertise from fine-tuning 7B-235B models to inference optimization on H100/A100 infrastructure. Research includes a widely cited clinical RAG framework (35+ citations) and first-author work at ACL (BEA Workshop).

PROFESSIONAL EXPERIENCE

Icahn School of Medicine at Mount Sinai New York, NY

Computational Scientist *Dec 2024 – Present*

- **OncoCITE - Multi-Agent Genomic Evidence Extraction (Submitted to Nature Cancer):** Performed systematic EDA on CIViC database (11,312 evidence items, 3,083 publications) quantifying 12 structural bottlenecks: curation latency (median 31d, P90 >21 months), Pareto inefficiency (top 100 sources = 29% coverage), and emerging target gaps (GPRC5D: 0 items despite FDA approval). Architected 6-agent solution (Claude Agent SDK, 22 MCP tools) with state serialization enabling pause-resume for long-running extractions and vision-based PDF extraction (300 DPI) replacing traditional OCR. Developed novel three-way validation framework treating source publications as ground truth—identified 24.2% curation errors in expert-curated databases. Validated on 15-paper corpus: 84% ground truth recovery, 97.8% novel discovery precision, 0% critical errors (n=108). Deployed normalizer to enrich all 11,312 items achieving 83.12% ontology resolution across 20 Tier-2 fields. Estimated 40+ hours of manual curation time saved per publication processed, with potential to scale across 3,000+ unprocessed CIViC sources. Published at ASH 2025.
- **PRIME Model - Predictive Relapse Indicators for Myeloma T-Cell Engagers (Blood 2025):** Co-developed predictive model for patients receiving BCMA- and GPRC5D-targeting T-cell engagers. Integrated clinical, genomic, and treatment response data to identify early relapse indicators. Presented at ASH 2025.
- **MMAP - GPU-Accelerated Genomic Pipeline (56× Speedup):** Engineered production pipeline on Minerva HPC using NVIDIA H100 GPUs with Parabricks (fq2bam, HaplotypeCaller), RAPIDS (cuDF/cuML), and DeepVariant. Orchestrated 57 computational processes across 3 integrated workflows (RNA-seq, WES, Integration) using Nextflow DSL2 with LSF scheduling. Achieved 95.8% processing time reduction (7 days → 3 hours, 56× speedup), enabling 8 patients/day throughput vs. 0.14 previously. Generates 10 clinical outputs including actionable variants, drug recommendations, venetoclax sensitivity predictions, and PSN molecular subtyping. Enabled same-day molecular tumor board readiness, transforming clinical timelines from 14 days to 1-2 days.
- **Multi-Omics Integration (Clinical Lymphoma Myeloma Leukemia 2025):** Applied modified IntegrAO graph neural network methodology to MMRF COMPASS cohort (N=655 samples) for multi-omics patient stratification. Integrated SNV, CNV, TME, and WES data achieving 50% classification granularity improvement (18 vs 12 subgroups) and 258% high-risk detection enhancement (43 vs 12 patients). Identified 18 distinct vulnerability profiles with 94% of patients having actionable targets.

Associate Computational Scientist *Oct 2023 – Dec 2024*

- **RAG System for Clinical Decision Support (Clinical Lymphoma Myeloma Leukemia 2024):** Led development of production RAG system using LangGraph orchestration, BAAI/bge-large-en-v1.5 embeddings, and Mistral-7B. Processed 5,000+ documents achieving 88% clinical effectiveness, reducing literature review time by approximately 70% for clinical fellows. Published in medRxiv [36 citations], presented at IMS 2024.
- **CAR-T CRS Prediction (Information MDPI, under review):** Built ML system for early Cytokine Release Syndrome detection in CAR-T patients (ide-cel, cilta-cel) as part of investigator-initiated trial (N=30 enrolled, N=25 analyzed). Engineered time-lagged features from continuous wearable monitoring (Current Health Gen 2, Feverscout) and 92-protein Olink cytokine panel. Evaluated 5 classifiers via StratifiedKFold CV: achieved 84.62% accuracy (ide-cel) and 80.62% (cilta-cel) within 6-hour prediction window. Detected 18/20 CRS episodes with 7-hour median lead time before standard nursing detection. SHAP analysis identified IFN-γ as cross-product predictor; fold-change classifier achieved 90% precision with 40-hour mean lead time. Supports transition to outpatient CAR-T delivery models.
- **Voice ASR Prototype for Clinical Terminology:** Fine-tuned OpenAI Whisper Small for on-device deployment, trained on recorded patient voice dataset to reduce word error rate for multiple myeloma terminology. Prototype demonstrated improved recognition accuracy for domain-specific medical terms that standard ASR models frequently misrecognize.
- **Graduate Student Mentorship:** Mentored 6 Carnegie Mellon University graduate students (Sep–Dec 2024) on CAR-T therapy monitoring capstone project. Guided experiment design, baselining, and reporting for CRS prediction using wearables and cytokines. Coordinated with clinical team on project scope and deliverables.
- **Genomic Foundation Models Integration:** Incorporated Scanpy (single-cell RNA-seq) and Geneformer (gene expression modeling) into precision medicine pipeline. Optimized workflows for processing 10,000+ cell samples and integrated with existing computational infrastructure.
- **Unified Patient Data Management Framework:** Engineered RESTful APIs integrating EHR (EPIC), lab results (REDCap), and genomics data (Nextflow). Built automated pipelines using Git/GitHub and leveraged LLaMA-3 with RAG architecture for structured data extraction.

AYA New York, NY

Entrepreneurial Fellow *Jul 2023 – Oct 2023*

- Conceptualized and built AYA, leveraging reinforcement learning-optimized NLP model to extract and formulate questions from academic video content. Selected for Summer Sprint at NYU's Entrepreneurial Institute (10 selected from 150 startups). Developed end-to-end product combining speech processing, natural language generation, and educational content analysis.

New York University New York, NY
Machine Learning Researcher & Teaching Assistant *Sep 2021 – May 2023*

- **AI-Generated Plagiarism Detection (ACL 2023):** Developed novel multi-faceted NLP approach with Prof. Parijat Dube (IBM Research) achieving 94% accuracy in human-AI text classification. Method employs contrastive loss and LLM-generated paraphrases. DOI: 10.18653/v1/2023.bea-1.58 [31 citations]
- **Teaching Assistant - Data Structures & Algorithms for Bioinformatics:** Supported graduate-level course under Prof. Manpreet S. Katari for 3 semesters (Spring 2022, Fall 2022, Spring 2023), teaching data structures, algorithms, and genomic algorithms to 130+ master's students. Created assignments, held office hours, and bridged computer science fundamentals with bioinformatics applications.

Carcrew New Delhi, India
Technical Lead *Aug 2016 – Sep 2017*

- **Founding Technical Team:** Built MVP for automotive marketplace using Django, Flask, PostgreSQL. Contributed to technical due diligence that secured $2M Series A from TVS Group. Scaled to 10,000+ DAU.

PUBLICATIONS

- Quidwai MA, et al. "OncoCITE: AI-Driven Genomic Evidence Curation." (Submitted to Nature Cancer, 2025)
- Quidwai MA, et al. "Oncodif: An Auditable AI Framework for Automated Genomic Curation." Blood 146, 2646 (ASH 2025)
- Quidwai MA, Laganà A. "A RAG Chatbot for Precision Medicine of Multiple Myeloma." medRxiv 2024. DOI: 10.1101/2024.03.14.24304293 [36 citations]
- Quidwai MA, et al. "2P-145 Innovative AI-Driven Decision Support Tool for Multiple Myeloma Using RAG." Clin Lymph Myel Leuk 24, S123-S124 (IMS 2024) [1 citation]
- Quidwai MA, Li C, Dube P. "Beyond Black Box AI-Generated Plagiarism Detection." ACL BEA 2023. DOI: 10.18653/v1/2023.bea-1.58 [31 citations] — IBM Research
- Rajeeve S, ..., Quidwai M, et al. "Early Detection of CRS Using Wearable Devices." Information (MDPI), under review; SSRN 5217949
- Mouhieddine T, ..., Quidwai M, et al. "PRIME: Predictive Relapse Indicators for Myeloma T-Cell Engagers." Blood 146 (Suppl 1), 3996-3996 (ASH 2025)
- Hamidi H, ..., Quidwai M, et al. "Integrating Microenvironment with Tumor Multi-Omic." Clin Lymph Myel Leuk 25, S181-S182 (IMS 2025)

EDUCATION

Master of Science in Computer Engineering (Honors) 2021 – 2023
New York University, Tandon School of Engineering | New York, NY
Advisors: Prof. Dennis Shasha, Prof. Manpreet S. Katari, Prof. Parijat Dube (IBM Research)
Coursework: Deep Learning, Machine Learning, High Performance ML, NLP, ML for Cyber-Security, Algorithms & Data Structures for Bioinformatics
Scholarship: NYU Tandon Graduate Scholarship — Merit-based award of $8,000/year for academic excellence

Bachelor of Technology in Computer Science 2012 – 2016
Dr. A.P.J. Abdul Kalam Technical University | Ghaziabad, India

TECHNICAL SKILLS

LLM & Agents: Fine-tuning (7B-235B, Qwen/Mistral/MedGemma, LoRA/QLoRA, DPO, Unsloth), Multi-Agent Orchestration (Claude/OpenAI Agent SDK, LangGraph, Supervisor-Worker, MCP 22 tools), RAG (Vector DBs, Tool-calling), State Serialization, Deterministic Replay, Chain-of-Thought, Guardrails

Serving & MLOps: vLLM, SGLang, TensorRT-LLM, Ollama, KV Cache Optimization, <100ms TTFT, Streaming Inference, Quantization (FP8/INT4), OpenAI Evals, DeepEval, LangSmith, Logfire, MLflow, CI/CD, A/B Testing, Three-way Validation

Multimodal AI: Vision (Claude 3.5, Qwen-VL, DeepSeek-OCR, PDF-to-Image 300 DPI), Voice (Whisper Fine-tuning, Silero VAD, Piper TTS), Domain-specific ASR

Genomics & Bio: Nextflow, Parabricks (fq2bam, markdup, haplotypecaller), DeepVariant, NGS Pipelines (STAR, BWA-MEM, fastp, GATK, FeatureCounts), Variant Calling (Mutect2, Lancet, HaplotypeCaller, Manta), CNV/Fusion (FACETS, BEDTools, Arriba), Annotation (VEP, SnpEff), Biomarkers (MSI, TMB, HRD, GEP70), Single-Cell (Scanpy, Geneformer), PhyloWGS

Healthcare Data: Precision Medicine, Clinical NLP, EHR Integration (EPIC/REDCap), HIPAA, CAP/CLIA Validation, Ontologies (CIViC, ClinVar, RxNorm, HPO, NCIt), PubMed E-utilities

HPC & Cloud: NVIDIA H100/A100 Clusters (196-GPU Minerva), Multi-GPU (CUDA, NVLink), LSF/SLURM, Docker/Singularity, AWS (Bedrock, SageMaker, OpenSearch), Azure OpenAI, RAPIDS (cuDF/cuML)

Languages & Core: Python (PyTorch, Transformers, Pandas, scikit-learn, XGBoost), SQL (PostgreSQL, MongoDB), Bash, Git