Vishesh Goyal

+91 7827253699 Email LinkedIn GitHub Portfolio

AI engineer. Full-stack developer. Published researcher. All simultaneously, until it ships.

4 IEEE-published papers · 5 deployed production systems · 2 real-world freelance clients · End-to-end XAI, RAG, and full-stack delivery.

Education

Manipal Institute of Technology, Bengaluru **Sep 2022 – Jun 2026**

B.Tech in Computer Science Engineering

- Relevant Coursework: Quantum Algorithms and Computing, Deep Learning, Data Structures and Algorithms, Cryptography, Database Management, Computer Networks, Operating Systems, Cloud Computing

The Vivekanand School, New Delhi **Apr 2020 – Jun 2022**

AISSCE – Science (Class XII) *84.2%*

Bal Bharati Public School, Pitampura, New Delhi **Apr 2008 – Mar 2020**

AISSE (Class X) *91.2%*

Technical Skills

- **AI/ML:** XGBoost, scikit-learn, TensorFlow, PyTorch, SHAP, Captum, Sentence Transformers, HuggingFace, IBM WatsonX, OpenAI GPT-4o; model training & evaluation, feature engineering, explainability, RAG pipeline design.
- **NLP & Retrieval:** DeBERTa (NLI), TF-IDF, cosine similarity, semantic chunking, cross-encoder reranking, BGE-large embeddings, ChromaDB, Wikipedia MediaWiki API.
- **Backend:** NodeJS, Express, RESTful APIs, MongoDB, JWT-based authentication, Supabase, bcrypt, Multer, Nodemailer, microservice-style separation.
- **Frontend:** ReactJS, Next.js, TypeScript, Tailwind CSS, Framer Motion, component-based architecture, responsive UI design.
- **Languages:** Python, JavaScript, TypeScript, C.
- **Tools & Platforms:** Git, GitHub, Postman, Linux, VS Code, AWS S3, AWS CloudFront, Vercel, Electron, Qiskit.
- **Domains:** Explainable AI (XAI), Retrieval-Augmented Generation (RAG), Generative AI, Applied ML, Full-Stack Development, Quantum Cryptography (research-level).
- **Certifications:** Applied ML with Generative AI (NUS), Cryptography Specialisation – Cryptography & Information Theory, Symmetric Cryptography, Asymmetric Cryptography & Key Management (Coursera, University of Colorado), Web Development Fundamentals (IBM), Cloud Computing Fundamentals (IBM).

Publications

CrypTon: A Hybrid Quantum-Classical Framework Integrating BB84 QKD with AES for Secure Communication **IEEE NQComp, 2026**

- Implemented a hybrid post-quantum cryptographic system in Qiskit that integrates the BB84 Quantum Key Distribution protocol with AES encryption – replacing classical key exchange with quantum-secured key generation without modifying the cipher.
- Benchmarked across 3 key sizes (128, 192, 256-bit) and 3 file sizes (10KB, 100KB, 1MB): AES throughput was entirely unaffected by quantum key sourcing; QKD key generation added 38–111 seconds depending on key length.
- Achieved **97.3% eavesdropper detection accuracy** across 1,000 trials; interceptor presence spiked QBER to 20–30%, triggering automatic key rejection with zero false positives recorded.
- GitHub | IEEE Link

PRIORIS: Dynamic Adapting Scheduling for HPC – Eliminating Job Failure through Robust Resource Allocation **IEEE ICACIS, Jun 2025**

- Designed a real-time adaptive job scheduling algorithm for High Performance Computing environments that dynamically reorders the execution queue based on live CPU, memory, disk I/O, and network bandwidth availability – replacing failure prediction with failure prevention.
- Introduced a Calculated Priority Metric integrating base priority, estimated runtime, and resource cost; added dependency-driven job promotion, anti-starvation waiting queue cycling every 4th scheduling pass, and limited parallelism for adjacent-priority jobs.
- Achieved **24.7% reduction in makespan** and **31.5% reduction in average job wait time** compared to FCFS baseline, across a simulation of 5,000 synthetic jobs – outperforming even a failure-prediction model that required historical training data.
- GitHub | IEEE Link

Identifying Mango Leaf Diseases with Advanced Deep Learning Approaches and CNNs **IEEE INCIP, Jan 2025**

- Benchmarked three CNN architectures on a dataset of 12,046 mango leaf images spanning 9 classes, with aggressive augmentation to simulate real-world field photography conditions.

- Selected EfficientNet-B0 as the final model: **98.97% accuracy, 99.10% F1 score**, 0.40% error rate – with only 91 false negatives across 12,046 test images, minimising the dangerous failure mode of missed disease detection.
- GoogLeNet achieved comparable accuracy but was rejected due to 4× longer training time, making it impractical for edge deployment in agricultural settings.
- GitHub | IEEE Link

AI-Powered Personalised Learning Platform: NLP-Driven Article-Centric Chatbot with Sentiment Analysis **IEEE NQComp, 2026**

- Co-authored an NLP-based educational chatbot system that answers questions exclusively from user-uploaded articles using TF-IDF vectorisation and cosine similarity – zero reliance on external LLMs, eliminating hallucination risk entirely for classroom use.
- Achieved **90% question-answering accuracy and 90.84% precision** using lightweight, interpretable classical NLP; an SVM classifier handles parallel sentiment analysis on content tone.
- System is fully auditable and source-bounded – a deliberate architectural choice over transformer-based approaches for educational environments where answer provenance must be verifiable.
- IEEE Link

Experience

Deloitte Capstone Project **Feb 2025 – Jul 2025**

Project Team Leader – Vita AI *Bengaluru, India*

- Led the end-to-end architecture and delivery of **Vita AI**, an AI-powered medical diagnosis platform for doctors that predicts the top-10 most likely diseases from patient symptoms, ranked by clinical severity rather than raw probability – ensuring the worst-case diagnosis is always surfaced first.
- Built a three-stage ML pipeline: XGBoost classifier for disease prediction, SHAP explainer generating per-disease feature attribution breakdowns, and IBM WatsonX translating SHAP outputs into clinically-worded narrative reports for each prediction.
- Deployed a full MERN stack system – ReactJS doctor and patient frontends, NodeJS + Express API gateway, MongoDB for user data, prescriptions, appointment scheduling, and report storage – with JWT-based role separation between doctor and patient access scopes.
- Conducted end-to-end validation with **16 real-world practising doctors**: model outputs were reviewed, challenged, and used to boost and alignment-correct predictions to real clinical judgment before the final system was tested and endorsed by the same cohort.
- Achieved **95.43% test accuracy**; **80% doctor validation accuracy** on real symptom inputs; won the **Deloitte Capstone Project Award**.
- GitHub

HashStudioz Technologies **Jun 2024 – Jul 2024**

Full Stack Intern *Noida, India*

- Developed and optimised RESTful APIs using NodeJS, Express, and MongoDB across multiple product modules; improved backend reliability and reduced redundant query patterns through refactoring and modular route separation.
- Built and iterated on responsive ReactJS dashboards with Tailwind CSS in cross-functional sprints; diagnosed and resolved **10+ user-reported UI/UX issues**, improving interface consistency and stakeholder satisfaction across delivery cycles.

Freelance Work

RGLawz Case Management System **Jan 2025 – Mar 2025**

Solo Builder & Designer – Full-Stack Internal Platform *New Delhi, India*

- Designed and built a complete internal case management platform from scratch for a live New Delhi law firm, replacing physical registers and informal WhatsApp coordination with a structured, searchable, role-based web and desktop application.
- Implemented six core modules: case registration and search, a cause list calendar view showing hearings, full hearing history, client accounts and billing, digital bill generation, and JWT + bcrypt secured RBAC separating admin and staff roles.
- Shipped as both a Vercel-deployed web app and a cross-platform **Electron desktop application** with a **custom AWS S3-based auto-update system** – the app silently checks for newer versions on launch and self-updates without any user intervention, eliminating IT support overhead for a non-technical team.
- System is in **active daily production use**: **238 active cases** registered across multiple Delhi courts, **150+ hearings tracked and updated monthly**, digital billing live, Electron app installed across all firm devices.
- Live Demo (Web)

The Auctores – Business Platform & AI Sales Agent **Mar 2025 – May 2025**

Solo Builder & Designer – Next.js Platform + GPT-4o Sales Pipeline *Remote*

- Built a full-stack business platform for a virtual admin company on Next.js with Sanity CMS for blog management, Supabase for database, and a ClickUp API-integrated client dashboard.
- Designed and engineered a **white-labelled GPT-4o sales agent** – psychologically calibrated through prompt engineering to surface visitor pain points, plant conversion seeds after 5–6 message exchanges, and qualify leads by extracting budget, timeline etc.
- Built a fully automated **CRM pipeline**: a secondary GPT-4o-mini instance reads every completed conversation transcript, extracts structured sales intelligence (lead temperature, budget range, recommended follow-up action), and writes a clean row to a Google Sheets CRM – zero manual data entry, every conversation becomes a sales asset.
- Delivered the CRM pipeline and psychological calibration architecture as additions beyond the original contract scope, because they were the correct solution to the underlying business problem.
- Handoff Version | Live Site

Key Projects

SATORI – RAG-Based AI Study Assistant

Python, ChromaDB, BGE-large Embeddings, IBM WatsonX, Meta Llama 3.3 70B, PyMuPDF, Tesseract OCR, React, TypeScript

- Built a full-stack production-grade Retrieval-Augmented Generation system that accepts up to 20 PDFs and builds a session-isolated personal knowledge bank using BGE-large-en-v1.5 embeddings stored in ChromaDB.
- Implemented a **dual-mode answer architecture**: Strict Mode returns answers built entirely from retrieved PDF chunks with source citations and page numbers (zero hallucination risk); LLM Tutor Mode grounds IBM WatsonX Llama 3.3 70B responses in the same retrieved chunks, giving users depth without losing document accuracy.
- Built a hybrid three-signal retrieval pipeline: dense vector similarity (BGE-large), cross-encoder rerankings for joint query-document scoring, and keyword score boosting for exact terminology matching ensuring retrieval that understands both meaning and phrasing simultaneously.
- Engineered a PDF extraction pipeline handling scanned pages via Tesseract OCR, topic-aware chunking across page boundaries, and inline equation rendering. Equation image regions are cropped, stored separately, and surfaced alongside text answers so users see the actual rendered formula.
- Implemented three-turn conversational memory with active follow-up detection.
- GitHub

RealityCheck – LLM Hallucination Correction Pipeline

Python, NLP, DeBERTa (NLI), Sentence Transformers, IBM WatsonX, Wikipedia API

- Developed a modular, model-agnostic six-phase hallucination correction pipeline that intercepts LLM responses before they reach the user: claim extraction, semantic decomposition, Wikipedia-grounded evidence retrieval, four-signal NLI verification, reasoning override, and deterministic answer synthesis – no secondary LLM call in the correction loop.
- Evidence retrieval uses a cache-first architecture. Retrieved chunks are embedded (multi-qa-MiniLM-L6-dot-v1) and ranked by semantic similarity before verification.
- Claim verification fuses four independent signals: semantic similarity gate, DeBERTa-based NLI scoring (entailment / contradiction / neutral), rule-based overrides targeting known NLI failure modes, and Noisy-OR aggregation across evidence chunks.
- Contradicted claims are repaired using domain-aware correction templates derived from the best contradicting evidence; insufficient-evidence claims are epistemically hedged rather than deleted preserving informational completeness unlike WikiChat, which omits unverifiable content.
- Evaluated on **TruthfulQA** across IBM WatsonX Granite, Meta Llama, and MistralAI Mistral: improved answer-level accuracy by **up to 30 percentage points**, hallucination recall from **37–45% to 78–83%**, overcorrection rate held below **7%**.
- GitHub

AgriVerse – Explainable AI Crop Recommendation System

Python, XGBoost, SHAP, ReactJS, NodeJS, Express, Open-Meteo API

- Built a deployed, full-stack explainable AI system for crop recommendation: accepts soil parameters (N, P, K, pH), fetches **live weather data** via Open-Meteo API for the user's location, and predicts top-3 crops with probability scores across 22 crop types using XGBoost trained on 2,200 instances.
- Achieved **99.80% overall accuracy**; 18 of 22 crop classes returned perfect F1 scores of 1.00; remaining 4 classes scored 0.98–0.99. Used novel dropout simulation during training (randomly masking input features) to build robustness to incomplete real-world soil data – critical for Indian smallholder agriculture.
- Integrated two-tier SHAP explainability: per-crop local waterfall charts showing which soil and weather variables pushed each recommendation, and a precomputed global feature importance chart for systemic pattern analysis by researchers and extension workers.
- Introduced a **Trust Score** (derived from probability margin between top crop and competitors, dynamically penalised for missing inputs or API failures).
- Built a **Counterfactual Engine**: computes Euclidean distance between the user's N/P/K/pH profile and centroid profiles of all other crop categories; excludes weather variables (non-amendable) and returns concrete soil adjustment guidance – e.g., "Rice can be grown: decrease Phosphorus from 50 to 47.58."
- GitHub

SkillAI – AI Career Recommendation Engine

Python, XGBoost, PyTorch, Deep Neural Networks, Captum, IBM WatsonX, ReactJS, MongoDB

- Developed a career guidance platform using a **two-layer hierarchical ML architecture** trained on the 2025 U.S. Department of Labor O*NET dataset (1,000+ occupations): XGBoost first clusters the occupation space to identify which broad career domain the user's skill profile belongs to, then a specialised DNN trained only on that cluster finds the top-10 best-fit careers within it.
- The hierarchical design deliberately prevents flat-classifier noise: the XGBoost handles coarse domain separation; the cluster-specific DNN handles fine-grained discrimination between adjacent roles (e.g., Data Scientist vs. ML Engineer) without ever seeing occupations from other clusters.
- Applied Captum to generate per-career feature attributions: for each of the top-10 recommendations, users see exactly which skills drove the match and which worked against it, with magnitude – translated into plain-language career reports via IBM WatsonX.
- GitHub

Extracurricular

- **Leadership:** Outreach Partnerships Head, E-Cell MIT-B (Dec 2024 – Aug 2025); drove external partnerships, coordinated industry collaborations, and managed stakeholder engagement for entrepreneurship events across the institution.
- **Market & Financial Systems:** Actively track Indian equity markets and derivatives; practice paper trading with a focus on understanding risk-reward dynamics, market microstructure, and data-driven decision-making under uncertainty – skills that directly translate to quantitative reasoning in applied ML contexts.
- **Philosophy & Behavioural Science:** Independent study of moral philosophy, behavioural economics, and cognitive science to build systems-level thinking about human decision-making – informing how I approach AI system design, explainability, and the human-in-the-loop problems that matter most in applied AI.