

# QreateAI

AI Products & Engineering

## About QreateAI

QreateAI is led by a founding team of operators who have built, deployed, and scaled production-grade AI systems in high-stakes, real-world environments.

Unlike traditional consultancies that focus on advisory or proof-of-concept work, the QreateAI founding team brings deep hands-on experience across applied AI research, full-stack system development, and live deployments. The team has collectively delivered AI products spanning large language models, speech and audio intelligence, computer vision, probabilistic modeling, and privacy-preserving AI.

The founders have built and deployed AI systems currently being piloted with AIIMS and private hospitals across India, including ambient clinical documentation tools and AI assistants for automated pre-charting and triaging. These systems were designed and shipped under real constraints such as noisy data, latency, clinical accuracy, regulatory considerations, and end-user adoption.

The team's background includes advanced research and applied engineering experience at institutions and organizations such as the **National University of Singapore (NUS)**, **Indian Institute of Science (IISc)**, **National Taiwan University (NTU)**, **Samsung R&D**, and **HyperVerge**. This foundation enables QreateAI to combine research-grade rigor with production-level reliability.

QreateAI's core advantage lies in its operator-led approach. The founders have repeatedly taken AI systems from concept to deployment, aligning technical design with business outcomes. As a result, QreateAI delivers AI solutions that move beyond experimentation to measurable impact, reliability, and scale.

## Founding Team

**Aryaman** — Chief Executive Officer *[PLACEHOLDER: Brief bio, background, key expertise]*

**Subhanshu Arya** — Chief Product Officer *[PLACEHOLDER: Brief bio, background, key expertise]*

**Harsh Pandey** — Chief Technology Officer *[PLACEHOLDER: Brief bio, background, key expertise]*

## How We Work

**Build** — End-to-end product development from concept to production deployment. We take ownership of the full AI product lifecycle, delivering production-ready systems.

**Augment** — Embed senior AI engineers into your team to accelerate development. Our engineers integrate seamlessly with your workflows and tech stack.

**License** — Deploy our production-tested healthcare AI solutions (DoQ, TalQ) within your infrastructure. Proven systems ready for enterprise integration.

## Why QcreateAI

- **Speed** — Operator-led teams move fast. We ship MVPs in weeks, not months.
- **Deep AI Expertise** — Research backgrounds at NUS, IISc, NTU combined with production experience at Samsung R&D and HyperVerge.
- **Proven Delivery** — 8+ projects delivered across healthcare, telecom, government, fintech, and construction tech.
- **Real-World Constraints** — We build for noisy data, latency requirements, regulatory compliance, and end-user adoption.

qcreateai

# Our Products

*Production-tested healthcare AI solutions available for licensing and enterprise deployment. Currently being piloted at major government institutions in India including AIIMS.*

## DoQ AI Assistant

**Industry:** Healthcare & Clinical Services

**Status:** Piloting at AIIMS and private hospitals across India

### Problem

Healthcare providers continue to rely on manual, form-based, or front-desk-driven patient intake processes that are time-consuming, inconsistent, and prone to information loss. Patients often arrive at consultations without having communicated their symptoms in a structured or complete manner, forcing doctors to spend valuable consultation time on basic data collection rather than clinical reasoning.

This leads to shorter effective consultation time, clinician fatigue, and reduced patient throughput, especially in high-volume outpatient settings. Doctors are increasingly burdened with administrative and documentation tasks, diverting attention away from patient interaction and clinical decision-making.

### Solution

DoQ AI is a **pre-medical consultation intelligence assistant** designed to prepare both the patient and the doctor before the clinical visit begins. The system engages patients through voice and chat interfaces to systematically collect all medically relevant information required for an effective consultation.

Using guided, adaptive questioning rooted in clinical reasoning, DoQ AI narrows patient-reported symptoms into probabilistic health concerns, ensuring that only relevant and high-signal information is captured. The system does not replace medical judgment or provide a diagnosis; instead, it functions as a structured clinical intake layer that enhances consultation readiness and efficiency.

### Impact

- Significant reduction in time spent on repetitive history-taking and irrelevant questioning
- Doctors receive concise, structured pre-consultation summaries for faster clinical context-building
- Increased patient throughput in high-volume settings while preserving consultation quality
- Reduced cognitive load for clinicians and improved time utilization
- Shorter, more purposeful consultations with better patient experience

## Technical Implementation

### Multi-Channel Interfaces

- **Voice (Speech-to-Speech):** Real-time telephony via Twilio SIP, with streaming Speech-to-Text using Deepgram/Twilio ML, LLM-driven response generation, and high-quality Text-to-Speech via ElevenLabs

- **Chat (WhatsApp):** Stateful WhatsApp conversations powered via WABA + Twilio, with session context maintained in Redis for low-latency, multi-turn symptom collection

### **Agentic Clinical Reasoning**

- Multi-stage agentic workflow: Symptom Intake Agent gathers high-level complaints; Doctor Reasoning Agent drills down via adaptive follow-up questions
- Dynamic question generation driven by previous patient inputs, clinical rules, and differential diagnosis reasoning
- Hybrid system combining rule-based medical logic with LLM-based inference for safety and flexibility

### **Medical Safety & Guardrails**

- Medical-grade LLMs including fine-tuned models trained on Indian medical exams (NEET) and domain-specific clinical datasets
- Strong safety controls using constrained prompts, schema-validated JSON outputs, and predefined guardrails to reduce hallucinations

### **LLM Stack & Backend**

- Model-agnostic orchestration leveraging GPT-4/GPT-4o/Claude alongside specialized medical LLMs
- Python FastAPI microservices with custom state-machine-based orchestration layer
- Automated pre-consultation SOAP reports with symptom timelines, risk flags, and confidence scores

qreateai

# TalQ

**Industry:** Healthcare Providers & Hospital Systems

**Status:** Piloting at AIIMS and private hospitals across India

## Problem

Doctors spend a significant portion of consultation time on manual note-taking and post-visit documentation, reducing patient engagement and increasing burnout. In high-volume clinics, documentation can consume 25-40% of a clinician's working time, leading to rushed consultations, delayed records, and reliance on memory-based summaries.

Manual documentation often results in loss of conversational nuance, symptom timelines, and clinical reasoning. Variability in note quality across clinicians leads to inconsistent medical records, affecting continuity of care, audits, compliance, and downstream analytics.

## Solution

TalQ provides a **doctor-in-the-loop conversation intelligence system** that automatically captures and structures real doctor-patient consultations. It records consultations only when explicitly initiated by the doctor, transcribes conversations with medical vocabulary adaptation, and converts them into accurate, transcript-grounded SOAP notes.

The system is designed strictly as a summarization and structuring layer, not a diagnostic tool. Doctors retain full control with the ability to review, edit, and finalize notes before they are stored or shared.

## Impact

- **30-50% reduction in documentation time** per consultation, allowing doctors to focus more on patient interaction
- **Improved consultation quality** as clinicians no longer need to multitask between listening and writing
- **Higher consistency and completeness** of medical records with standardized SOAP notes from full conversation context
- **Faster record finalization**, reducing post-consultation backlog and improving operational throughput
- **Reduced clinician burnout**, contributing to better retention and sustainable operations

## Technical Implementation

### Conversation Capture

- Live, doctor-initiated recording triggered within clinic workflow; doctor explicitly controls start/stop
- Supports clinic room microphones, mobile devices, and web applications via WebSocket streaming
- Continuous audio capture ensures full conversational context without manual segmentation

### Speech Processing Pipeline

- Post-consultation transcription using AWS Transcribe (Healthcare) and Deepgram with automatic fallback

- Medical vocabulary adaptation routing transcripts through fine-tuned medical LLM prior to report generation
- Accurate speaker diarization to separate doctor and patient utterances, preserving conversational roles

### **Clinical Structuring & Safety**

- LLM-based clinical summarization strictly grounded in recorded transcript — no inference beyond spoken content
- Hallucination controls via schema-enforced JSON outputs, evidence grounding, and confidence scoring
- SOAP reports are auto-generated but fully editable by doctor prior to finalization

### **Backend & Compliance**

- Python FastAPI microservices with asynchronous pipelines for long consultations (30-40 minutes)
- Outputs in structured JSON, clinically formatted PDFs, and EMR-friendly schemas
- Audio recordings encrypted at rest/transit, automatically deleted; transcripts undergo PII masking

qreateai

# Client Work

*Selected projects delivered for enterprise clients across telecom, government, fintech, and construction technology.*

## ObZ — Enterprise AI Chatbot

**Client:** Major telecom operator (Indonesia)

**Industry:** Customer Experience & Contact Center Automation

### Problem

Telecom operators receive massive volumes of service requests related to billing, service status, plan details, applications, grievances, and policy information. A significant portion of these interactions is repetitive and informational, overwhelming call centers and help desks. This results in long wait times, inconsistent responses, and poor service experience for customers.

Traditional support systems are constrained by working hours, agent capacity, and manual workflows. During peak demand periods such as service outages or billing cycles, support teams struggle to scale. This leads to service delays, increased operational costs, and reduced trust in service delivery.

### Solution

ObZ is an **intelligent, conversational automation platform** designed to serve telecom customers across digital channels. It uses natural language understanding to handle high-volume service queries related to accounts, services, applications, and policies in real time.

Built as a scalable service automation layer, the chatbot integrates with telecom systems and CRMs to deliver accurate, context-aware responses. It supports multilingual interactions, seamless handoff to human agents for complex cases, and operates continuously to ensure reliable, round-the-clock service delivery.

### Impact

- **30-60% reduction in call center and helpdesk load** by automating routine service requests
- **Faster service resolution** with significantly reduced wait times for customers
- **24/7 service availability** improving accessibility across geographies and time zones
- **Lower operational costs** by minimizing dependency on manual support staff
- **Improved service transparency and trust** through consistent, accurate, and timely responses

### Technical Implementation

- **Omnichannel Support:** Web, mobile apps, messaging platforms, IVR-to-text, and government portals
- **Multilingual:** Automatic language detection with regional language and transliteration support (Hinglish, Spanish, Japanese, etc.)
- **NLP Pipeline:** Intent detection, entity extraction (account IDs, services, policies), multi-turn context management

- **Knowledge Base:** Structured knowledge linking services, workflows, policies with human-in-the-loop learning
- **Integration:** API integration with telecom infrastructure, CRMs, and service databases; seamless agent handoff
- **Architecture:** Modular API-first microservices with high availability, auto-scaling, PII masking, and encryption

qreateai



# Floorplan Annotation Platform

**Client:** Stanford & MIT-backed construction tech company (San Francisco)

**Industry:** Architecture, Engineering & Construction Technology

## Problem

Architecture, engineering, and construction teams rely heavily on manual annotation of PDF floor plans — a process that is slow, repetitive, and prone to inconsistencies. Large projects often involve hundreds of plans, where manual labeling can take hours per floor plan, delaying downstream design, analysis, and execution workflows.

Manual labeling introduces variability in room names, boundaries, and spatial definitions across teams and projects. This inconsistency reduces data reliability for BIM workflows, construction planning, and spatial analytics, leading to rework, misinterpretation, and quality control challenges.

## Solution

The Floorplan Annotation Platform provides a **browser-based, instant-annotation system that lets teams annotate uploaded PDF floorplans without** preprocessing or backend setup. Users can create, edit, and manage room-level annotations in real time, supported by assisted room label suggestions that improve speed and consistency.

Designed as a **data preparation and annotation layer**, the platform converts static floor plans into structured, analysis-ready datasets. Its frontend-only architecture enables easy integration into existing AEC workflows, allowing teams to annotate at scale while maintaining human oversight and accuracy.

## Impact

- **50-70% reduction in floor plan annotation time**, accelerating project onboarding and design review cycles
- **Improved labeling consistency and accuracy**, reducing downstream rework in BIM, CAD, and construction planning
- **Faster turnaround for large projects**, with support for bulk annotation of 100+ floorplans simultaneously
- **Lower operational overhead**, eliminating the need for specialized tools, preprocessing pipelines, or backend infrastructure
- **Higher data quality for downstream systems**, enabling reliable spatial analytics and digital twins

## Technical Implementation

- **Frontend-Only Architecture:** Pure React.js application, all processing client-side for low latency and simplified deployment
- **PDF Annotation Engine:** Direct support for multi-page PDF floorplans with real-time bounding box creation and editing
- **Assisted Labeling:** Automatic room classification suggesting room names based on spatial and layout cues
- **Human-in-the-Loop:** Annotators can accept, override, or correct all suggestions
- **Scalability:** Optimized for high-volume annotation with fast rendering and consistent labeling behavior

qreateai

# Intelligent Traffic Management System

**Client:** State Government in partnership with educational institution

**Industry:** Transportation Planning, Smart City Infrastructure

**Recognition:** Bengaluru Mobility Challenge 2024 ([dataforpublicgood.org.in](https://dataforpublicgood.org.in))

## Problem

Urban traffic management remains largely reactive, dependent on manual monitoring and static signal plans derived from historical averages. Such systems are unable to respond dynamically to real-time fluctuations caused by peak demand, incidents, or localized disruptions. This results in delayed interventions, inefficient signal utilization, and persistent congestion across critical junctions.

Most cities lack granular, continuous traffic data that captures vehicle movements, turning patterns, and inter-junction flows. Despite extensive deployment of CCTV and Safe City camera networks, their use is largely confined to passive surveillance and post-incident analysis, missing opportunities to improve traffic flow and optimize signal control.

## Solution

The Intelligent Traffic Management System introduces an **AI-driven, data-centric approach** to urban traffic control by transforming existing CCTV infrastructure into a real-time traffic intelligence layer. The system continuously analyzes live video feeds to detect, track, and re-identify vehicles across multiple junctions, enabling a dynamic understanding of traffic flow.

By combining short-term traffic forecasting with turning movement prediction, the solution enables proactive signal planning and congestion mitigation instead of reactive enforcement. The platform integrates with downstream traffic optimization systems, planning dashboards, and decision-support tools.

## Impact

- **Shift from reactive to proactive traffic management** — anticipate congestion before it escalates through forecasting
- **City-wide traffic intelligence at operational scale** — near real-time visibility across multiple junctions simultaneously
- **Data-driven infrastructure planning** — origin-destination flow intelligence for road redesigns, flyovers, and policy interventions
- **Reduced operational dependence on manual processes** — AI-driven analytics improve consistency and scalability
- **Validated real-world relevance** — demonstrated through city-scale pilot deployment and national-level recognition

## Technical Implementation

- **Data Sources:** Live CCTV feeds from Safe City cameras across North Bengaluru, junction-mounted sensors, historical traffic patterns
- **Computer Vision:** Deep learning pipelines for vehicle detection, classification, and multi-object tracking across diverse Indian traffic conditions
- **Vehicle Re-ID:** Robust feature extraction to track vehicles across non-overlapping camera views under varying lighting and occlusions

- **Prediction Models:** LSTMs and time-series models for short-horizon traffic volume prediction and junction-level dynamics
- **Architecture:** Modular, research-to-production-ready design for cloud or edge-assisted deployment; 15-minute update cycles
- **Scalability:** Designed for city-scale deployment (aligned with Bengaluru's 9,000+ CCTV cameras)

qreateai

# Lobo AI — Personalized Market Intelligence

**Client:** San Francisco-based investment platform

**Industry:** Financial Services, FinTech

## Problem

Stock market investors are exposed to an overwhelming volume of news, reports, and updates spread across multiple platforms. Filtering this information to identify what is relevant to their specific portfolio is time-consuming and cognitively taxing. Critical stock-specific updates are often missed while investors spend excessive time consuming irrelevant or redundant content.

Most financial insights are delivered in text-heavy formats that require focused screen time, limiting accessibility during commutes or multitasking periods. The lack of personalized, audio-first financial content makes it difficult for users to stay informed in a convenient, habit-forming manner aligned with daily routines.

## Solution

Lobo AI is an **AI-powered personalized audio intelligence platform** for stock market investors. Users select up to 20 stocks and automatically receive daily, stock-specific audio briefings that include relevant news updates, summaries, and contextual insights.

Built as a personalization and content delivery layer, Lobo AI converts large volumes of financial text data into customized, voice-based briefings tailored to each user's portfolio. All curated news and insights are delivered before the stock market opens, ensuring users are fully informed and prepared to make timely investment decisions.

## Impact

- **Pre-market readiness** — consolidated view of all relevant stock-specific news before trading begins
- **Significant reduction in information overload** — filtered market news covering only selected stocks
- **Improved decision-making efficiency** — investors start the trading day informed and confident
- **Higher engagement and retention** — audio-first delivery integrates into morning routines
- **Faster reaction to market developments** — timely, summarized updates delivered ahead of market hours

## Technical Implementation

- **Data Sources:** User-selected stocks (up to 20), real-time financial news feeds, exchange announcements, company press releases
- **News Filtering:** Entity recognition maps incoming news to user-selected stocks; relevance ranking and deduplication
- **AI Summarization:** NLP models for news classification, extractive and abstractive summarization, contextual framing
- **Audio Generation:** Text-to-Speech models optimized for financial content clarity and pacing; scheduled before market open
- **System Design:** Event-driven ingestion pipeline, batch processing for pre-market compilation, web-based UI for playback

- **Scalability:** Supports thousands of concurrent users with stateless backend services and scalable job queues

qreateai

# Prediction Market Platforms

**Clients:** Two enterprise clients (under NDA)

**Industry:** Prediction Markets & Gamified Forecasting

*We have built and deployed two distinct prediction market platforms for enterprise clients, enabling users to stake tokens on event outcomes across categories including sports, crypto, finance, weather, and entertainment.*

## Problem

Decision-makers, researchers, and analysts often lack scalable methods to quantify what a diverse group of informed participants believes will happen in the future. Traditional forecasting relies on expert reports, polls, or heuristics that are slow, biased, and costly. There is no efficient mechanism to aggregate dispersed beliefs into real-time probability estimates.

Many token-based platforms struggle to maintain active participation beyond the initial onboarding phase. Users earn tokens but lack meaningful ways to utilize them. Without interactive use cases, token utility diminishes and user retention drops, leaving ecosystems stagnant.

## Solution

Both platforms provide structured, gamified environments where users stake tokens on predicted outcomes. **Platform A (Because)** enables users to stake on price ranges for assets (crypto, stocks, forex, commodities), with collective stake distribution surfacing crowd-sourced confidence levels.

**Platform B (Circles)** focuses on binary Yes/No predictions across sports, weather, entertainment, and crypto. It features AI-powered multi-language support, automated payout pipelines, and third-party token integration for global reach.

## Impact

- **Emergent probability signals** — crowd-sourced confidence levels providing insight into collective expectations
- **Increased token utility and engagement** — users have compelling reasons to earn, hold, and stake tokens
- **Global accessibility** — multi-language support (English, Japanese, Mongolian) expands participant base
- **Operational efficiency** — automated event scheduling, outcome resolution, and payout processing
- **Trust through transparency** — proportional payout logic, refund guarantees, and audit logging
- **Gamification that retains** — streak tracking, leaderboards, and win rate statistics

## Technical Implementation

### Shared Architecture

- **Backend:** Node.js/Express with TypeScript, MongoDB (Mongoose ODM), BullMQ + Redis for queue processing
- **Admin Dashboard:** React/Next.js with TanStack Query, Tailwind CSS, ApexCharts/Recharts for analytics

- **Validation & Safety:** Zod schema validation, JWT authentication, MongoDB transactions for atomic operations

#### **Platform-Specific Features**

- **Platform A:** React Native (Expo) mobile app, price-range pool logic, Firebase Cloud Messaging
- **Platform B:** Next.js 15 web app with Turbopack, binary Yes/No markets, OpenAI-powered translations, third-party token integration

qreateai



## Let's Build Together

Whether you're looking to build a new AI product from scratch, augment your team with senior AI engineers, or license our proven healthcare solutions, we'd love to hear from you.

**Email:** [team@qareai.com](mailto:team@qareai.com)

**Website:** [qcreate.qareai.in](https://qcreate.qareai.in)

**LinkedIn:** [QcreateAI](#)

qcreateai