Generative AI Masterclass · Bangalore 2026

INTELLIGENCE *AS* INFRASTRUCTURE

How Zentree Labs embedded AI into the decision-making nervous system of a deep-tech engineering firm.

HH

Hemanshu .V

CEO & Founder · Zentree Labs (US, Bangalore, Hyderabad)

200+

CUMULATIVE YEARS of AI

70+

AI/ML ENGINEERS

6

FORTUNE 500 GLOBAL CLIENTS

Acquisitions

AUVIZ → XILINX/AMD
NEURONOIDS -> ZENTREE LABS

Built on First Principles, Not Borrowed Playbooks

Production-grade AI/ML engineering from Bangalore/Hyderabad — semiconductor, Healthcare, edge hardware, and enterprise.

200+	70+	6	3
YEARS	ENGINEERS	CLIENTS	TOOLS

KEY CLIENTS

QUALCOMM	ONEPLUS	STEALTH STARTUP
d-MATRIX	MODULEMD	DEEPVISION

COMPANY EVOLUTION

FOUNDATION

Auviz Systems → Xilinx/AMD, Neuronoids -> Zentree

Hardware acceleration, FPGAs, real-time signal processing. Acquisition gave team Tier-1 semiconductor exposure at global scale.

ZENTREE FOUNDED

Vertical AI Engineering

Computer vision, ADAS, and edge AI. Early clients: Qualcomm, OnePlus, d-Matrix, Nuvia, stealth startups. Applied AI delivered as production systems, not prototypes.

NOW · 2024–2026

Agentic AI Practice

MCP-based agentic pipelines deployed internally first. Architecture proven in production, now offered as external client engagements.

Six Domains — One Integrated Practice

Full AI stack coverage: bare-metal CUDA kernels to high-level agentic orchestration.

PERCEPTION AI

Computer Vision & ADAS

YOLO, RT-DETR, multi-camera tracking at 30+ FPS. NATO-standard vehicle DRI for automotive and drone perception pipelines.

GENERATIVE AI

LLMs, RAG & Semantics

Domain fine-tuned LLMs with structured output generation. Custom RAG pipelines, semantic chunking, hybrid retrieval.

AGENTIC AI

MCP Agent Orchestration

Multi-agent systems on Model Context Protocol. Specialised agents for research, estimation, scoring, and reporting.

EDGE AI

AI Model Optimizer (AMO)

PTQ/QAT quantisation, structured pruning, ONNX export. Targets Qualcomm NPU, d-Matrix PIM, ARM Cortex-M, Jetson.

ML ENGINEERING

PyTorch & CUDA Kernels

ATen operator catalogue, C99 reference kernels for AI accelerators, hardware-portable ML ops for silicon partners.

APPLIED AI

Vertical Domain Solutions

Fraud detection for BFSI, BOM obsolescence for Capgemini, LLM driver generation. Full MLOps to production.

What We Will Cover Today

A structured journey from where the industry is, to what we built, to what you can take away and apply.

We Are Not at the Beginning of AI Adoption

Firms moving from signal to decision in hours — not weeks — consistently outmanoeuvre those that cannot.

> *"The firms that win next decade will be the ones that built AI into how they think and decide."*
>
> — Hemanshu V, Zentree Labs

DECISION SPEED COMPARISON · TRADITIONAL VS. AI-NATIVE

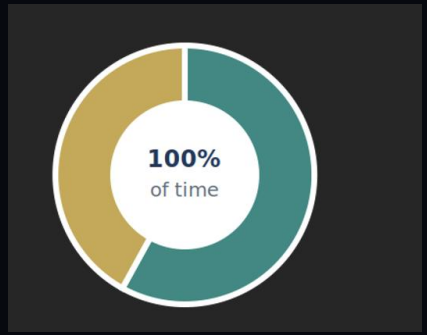
Decision Dimension	Traditional Firm	AI-Native Zentree
Proposal turnaround	5–7 business days	4–6 hours
Market intelligence	Weekly analyst brief	Daily AI digest
Talent allocation	1–3 days of emails	Under 1 hr (RBOT)
RFP qualification	Gut instinct	Scored, data-backed
Content throughput	1 doc/eng/week	3× via AI assistance

The Intelligence We Were Wasting on Process

Over 40% of highly skilled time was spent on tasks requiring intelligence but not deep expertise.

SENIOR ENGINEER TIME AUDIT · 2023 BASELINE

WHERE TIME WAS GOING



5–7 days
Proposal turnaround
Before AI

4–6 hrs
Proposal turnaround
After AI

42% of senior time reclaimed after AI deployment

The Decision Intelligence System

Five-stage AI pipeline. Every major decision enriched with maximum context and delivered at machine speed.

1

INGESTION

RFPs, email, LinkedIn, market reports. PDFs, DOCX, URLs normalised in real time. Structured extraction into clean JSON.

2

RAG ENRICHMENT

Context enriched against 200+ past proposals, delivery records, engineer profiles, pricing history. Hybrid dense+sparse retrieval.

3

AGENTIC ANALYSIS

Concurrent MCP agents for strategic fit, feasibility, technical scoring, and risk flagging. All in parallel, no human queuing.

4

SYNTHESIS

Master agent assembles outputs into executive brief: narrative, action items, confidence level, risks, next steps. Reads in 3 minutes.

5

AUDIT & OUTPUT

Version-controlled outputs with full reasoning traces. Sources cited, confidence scored, uncertainties flagged. Auditable always.

Latency

8–15 min end-to-end for standard RFP qualification.

Quality Floor

Structured evaluation rubric before any human review.

Feedback Loop

Every outcome feeds back. System improves with each engagement.

Governance

Full reasoning audit trails. Never a black box.

The Proposal Intelligence System

Handles qualification, estimation, and drafting end-to-end.

1 Structured Extraction

Parsing agent extracts scope, timeline, budget, and tech stack from any format. Output is clean JSON consumed by all downstream agents.

2 RAG — Learning from Past Work

Retrieves 5–8 most similar past engagements scored on domain match, tech overlap, team fit, and delivery outcome.

3 3-Point Effort Estimation

Optimistic / most-likely / pessimistic estimates from actual historical velocity. Estimation error dropped from ±35% to ±12%.

4 Full Proposal Narrative

LLM generates the complete proposal from structured context. Every claim source-cited. Human polish: 45–60 minutes vs. 2–3 days.

BEFORE VS. AFTER · MEASURED RESULTS

Metric	Before AI	After AI
Total turnaround	5–7 days	4–6 hours
Senior eng. hours	16–24 hrs	1–2 hrs
Bids per month	4–6	14–18
Estimation error	±35%	±12%
Win rate	Baseline	+18%

4 hr

Proposal turnaround. Down from 5–7 days. Full draft: estimates, team composition & terms.

RBOT — Resourcing Intelligence system

Resourcing is one of the most consequential decisions made every week. RBOT makes it well and fast.

1. **Requirements Parsing**
   Intake agent extracts skill requirements from the project brief and builds a weighted skill vector representing the ideal profile.

2. **Live Skill Graph Search**
   Vector matched against 50+ engineers — weighted by expertise, recency, current load, and career growth goals.

3. **Ranked Shortlist with Rationale**
   Not just a name — RBOT explains WHY this person fits, what skill gaps exist, and how to bridge them. Prevents burnout.

4. **Stretch Allocation Flags**
   Identifies 80%-fit allocations where the gap is a skill the engineer wants to develop. Every project becomes a growth opportunity.

PERFORMANCE METRICS

80% — ALLOCATION TIME SAVED
1–3 days reduced to under 1 hour

94% — ACCEPTANCE RATE
Accepted without modification

Live — SKILL GRAPH UPDATES
Auto-refreshed with each project close

RBOT is a live knowledge graph that compounds intelligence with every completed project.

Every Function — AI-Augmented

AI embedded into the operating rhythm of every function — with full transparency about where human judgement remains.

BUSINESS DEVELOPMENT

Lead Scoring & Pipeline Intelligence

Every lead scored against a multi-dimensional ICP rubric before a human hour is invested. Weekly AI competitive digest. Pipeline health dashboards flag at-risk deals 2–3 weeks early.

ENGINEERING OPERATIONS

AMO · Feasibility AI · Knowledge Base

AMO cuts model porting time by 60%. Feasibility agent screens requirements against hardware constraints in 30 minutes. Every design decision captured in an AI-queryable knowledge base.

PEOPLE & TALENT

RBOT · JD Generation · Performance AI

RBOT staffs projects in under 1 hour. AI-generated JDs from actual role requirements. Performance review cycles reduced — AI scaffolding lets conversations focus on substance.

FINANCE & REPORTING

Forecasting · Variance Analysis · Reports

Rolling 12-month cash flow models with AI-generated variance explanations. When actuals diverge, structured cause analysis is auto-produced. Weekly reports auto-generated.

The Agentic Backbone — How It Actually Works

Layered MCP architecture running in production today — not aspirational. Real business decisions processed daily.

DATA SOURCES	Client RFPs & SOWs · Engineer Profiles · Project Records · CRM Pipeline · Market Feeds · Financial Models
RAG LAYER	Chroma/Pinecone Vector Store · all-mpnet Embeddings · BM25 Sparse Retrieval · Cross-encoder Re-rank · Semantic Chunking
AGENT LAYER (MCP)	Intake & Parsing Agent · Proposal Gen Agent · Effort Estimation Agent · RBOT Resourcing Agent · BD Scoring Agent · Finance Narrative Agent
ORCHESTRATION	n8n Workflow Automation · Make.com Triggers · FastAPI Gateway · PostgreSQL State Store
OUTPUTS	Proposal Drafts & Decks · Staffing Plans · Go/No-Go Briefs · BD Digests · Finance Reports · Feasibility Analyses

MODELS: Claude Sonnet 4 · GPT-4o · Gemma 3n · Ollama (local) INFRA: LangChain · n8n · Make.com · FastAPI · PostgreSQL · Chroma

Case studies Client Deployments — Not Pilots

Real clients, real transactions and real decisions daily — not demos or pilots.

QUALCOMM Edge AI / SDK

AI Model Optimizer (AMO)
Quantisation and pruning pipeline for Qualcomm NPU targets. Engineers port vision and NLP models without manual

Model porting time down 60%

CAPGEMINI Embedded / Auto

BOM Mgmt & LLM Drivers
LLM-powered component obsolescence detection across Bills of Materials. Separate NLP-to-embedded driver code

Engineer throughput up 4×

d-MATRIX ML Engineering

SRAM-PIM CUDA Kernels
ATen operator implementations and C99 reference kernels for d-Matrix's SRAM-based PIM architecture. Full

Full operator coverage achieved

ModuleMD Industrial AI

Agentic Vision based Skin test AI
CNN Model based in production now.
Patent filed for novel algorithm

Asset uptime improved 12%

The ROI Matrix: How to Prioritize AI Investment

Two axes — Implementation Complexity vs. Business impact— tell you which workflows to pursue, plan for, or defer.

	LOW IMPACT	HIGH IMPACT
HIGH COMPLEXITY	**DEPRIORITISE** Custom fine-tuning Synthetic data gen Complex NLU pipelines Multi-modal	**STRATEGIC BETS** MCP agentic pipelines Decision intelligence RBOT resourcing Predictive maintenance
LOW COMPLEXITY	**PROCEED WITH CARE** Autonomous agents Customer-facing AI Financial approvals	**QUICK WINS** Proposal drafting Meeting summaries JD generation Report templates

ZENTREE'S RECOMMENDED SEQUENCING

1. **Start with RAG, not fine-tuning**
   Ground a general LLM in your proprietary data. 80% of the benefit at 5% of the cost.
2. **Automate one painful workflow fully**
   Don't spread thin. Deploy with rigour, measure against clear baselines.
3. **Build feedback loops from day one**
   Every accepted or rejected output is a training signal. This separates improving systems from plateauing ones.
4. **Expand autonomy as trust accrues**
   Start with AI as drafter. Expand scope only as audit trails and outcomes justify it.
5. **AI strategy is a leadership decision**
   The most impactful AI choices are organisational, not technical. Do not delegate this entirely.

Numbers We're Willing to Stand Behind

Internal benchmarking across 18 months. Not projections — operational changes with documented baselines and verified outcomes.

4 hr

PROPOSAL TURNAROUND

Down from 3–7 days. Full draft with estimates, team composition & terms.

70%

BD RESEARCH SAVED

Market scanning, competitor positioning, ICP enrichment — fully automated.

80%

FASTER RESOURCING

RBOT staffing in under 60 mins. Previously 1–3 days of back-and-forth.

3×

CONTENT THROUGHPUT

Docs, proposals, presentations per engineer per sprint. Quality unchanged.

+18%

PROPOSAL WIN RATE

More bids with consistent high quality — better selection, higher volume.

60%

MODEL PORTING TIME

AMO cuts the time to port a model from research framework to target hardware.

94%

RBOT ACCEPTANCE RATE

Staffing recommendations accepted without modification. A strong trust signal.

42%

SENIOR TIME RECLAIMED

Returned to engineering, R&D, architecture, and high-value client work.

What Founders often Get Wrong

We have made most of these mistakes ourselves and watched clients make them too.

Also consider build vs. buy

1 Mistaking a demo for a deployment

AI on curated demo data is not AI that works in production. Build for failure modes first. The distance between prototype and reliable system is where most projects stall.

2 Reaching for fine-tuning before RAG

Fine-tuning costs time, money, and expertise. RAG achieves 80% of the benefit at 5% of the cost. Fine-tune only when RAG has clearly plateaued and you can prove it.

3 10 tools instead of 1 deployed well

Most companies have 12 AI subscriptions and deeply use 2. Pick the highest-leverage workflow, deploy with rigour, then expand. The AI market is engineered to create FOMO.

4 Autonomy before trust has been earned

The fastest way to make a team reject an AI system is too much autonomy too soon. Start as drafter. Expand scope only as confidence and audit trails justify it.

5 Not closing the feedback loop

Every output your team acts on — or rejects — is a training signal. Systems that improve vs. plateau differ only in whether outcomes are systematically captured.

6 Delegating AI strategy to engineering

Which workflows to automate and which to keep human-in-loop are leadership decisions. The most impactful AI choices are organisational, not technical.

What We Are Building Next

The internal AI systems are version 1.0. A more ambitious programme is underway — moving from AI-augmented to AI-native operations.

In Progress · H1 2026

Persistent Multi-Agent Memory

Cross-session episodic and semantic memory. RBOT compounds learning from every project; allocation proposal improves with every RFP.

In Progress · H1 2026

Edge LLMs for Real-Time Perception

Quantised Gemma 3n + Phi-3 alongside vision models on edge hardware for multi-modal awareness. New automotive HMI use cases.

Planned · H2 2026

Engineering Intelligence Platform

Packaging internal systems — proposal engine, RBOT, feasibility agent, knowledge base — as a deployable product for engineering firms.

Planned · 2027

Agentic AI Consulting Practice

Helping enterprise clients build decision intelligence pipelines using the same architecture proven internally. Deployed systems, not workshops.

NORTH STAR METRIC · END 2026

No decision made by Zentree leadership is made without AI-synthesized context.

Not replacing judgement — ensuring every judgement is exercised with the best available information at machine speed.

THESIS FOR FOUNDERS IN THIS ROOM

The firms that define the next decade are not building the largest models.

They are integrating intelligence most deeply into how they operate. The technology is no longer the barrier — organisational willingness to change how decisions are made is the real differentiator in 2026 and beyond.

The Best AI Is the One Your Organisation *Actually Uses*

Start with one painful workflow. Build a small, reliable system.

Measure it against a clear outcome. Then earn the right to expand.

That is how Zentree built it — and it is still how we build.

zentreelabs.com

USA, Bangalore, Hyderabad

info@zentreelabs.com

OPEN FOR

- AI Consulting
- Agentic System Builds
- Edge AI Delivery
- PoC to Production
- Team Augmentation
- Engineering Intelligence Platform