From Knowledge Retrieval to Operational Intelligence

AI That Sees, Hears, and Understands Physical Equipment

White Paper v3

Nick Haschka
CEO, OnPoint Generators, Inc.

February 2026

Executive Summary

Field service industries face a compounding crisis. Experienced technicians are retiring faster than they can be replaced, taking decades of hard-won expertise with them. Traditional documentation systems capture procedures but not wisdom. First-generation AI tools retrieve documents but don't understand equipment.

This white paper describes **Rudy**, a platform that began as a self-learning knowledge retrieval system and is evolving into something fundamentally more ambitious: an operational intelligence system that perceives the physical world. Rudy's current production system combines hybrid retrieval (vector search + knowledge graph), community-driven feedback loops, and safety-aware response synthesis to deliver expert-level technical support. Its roadmap extends into multimodal perception—visual diagnostics from photos and video, acoustic signature analysis from equipment audio, real-time telemetry from IoT sensors and industrial controllers—culminating in predictive intelligence that identifies failures before they happen.

> *The core insight driving this evolution: the knowledge technicians need doesn't live only in documents. It lives in the sound a bearing makes before it seizes. In the color of exhaust smoke. In the vibration profile that shifts two weeks before a coupling fails. In the 4–20mA signal that's drifting outside its normal envelope. A system that only reads manuals is solving yesterday's problem. The future of field service AI must perceive the physical world with the same fluency it reads technical documentation.*

While developed for power generation technicians servicing emergency backup systems, the architecture is domain-agnostic. The same principles apply to any industry where complex physical equipment requires expert diagnosis: HVAC, manufacturing, oil and gas, marine systems, aviation maintenance, medical equipment, and telecommunications infrastructure.

Part I: The Foundation — Self-Learning Knowledge Retrieval

The Problem: Knowledge That Disappears

Field technicians working on complex equipment face a daily challenge: the information they need is scattered across sources that don't talk to each other—manufacturer manuals, service bulletins, regulatory codes, and

the tribal knowledge that exists only in experienced technicians' heads.

When a 30-year veteran retires, they take with them:

- Pattern recognition developed over thousands of service calls
- Equipment-specific quirks that never made it into manuals
- Diagnostic shortcuts that turn 4-hour jobs into 1-hour jobs
- Safety insights learned from near-misses

No documentation system captures this. No training program transfers it. It vanishes.

Standard RAG (Retrieval-Augmented Generation) systems—the current industry approach to AI-powered knowledge retrieval—perform a single retrieval pass: embed the query, find similar documents, synthesize an answer. This works for simple lookups but fails for real-world technical support. Single-pass retrieval can't follow diagnostic chains. Generic embeddings don't understand domain terminology. Static knowledge bases don't improve from feedback. And no expertise weighting means all sources are treated as equally authoritative.

Rudy's Current Architecture: What's in Production

Rudy's production system addresses these limitations through three integrated innovations: hybrid retrieval, community-driven learning, and safety-first design.

Hybrid Retrieval: Vector Search + Knowledge Graph

Rudy combines two retrieval mechanisms, dynamically weighted based on query intent:

Vector Search — Dense embeddings via Google's Gemini-001 model (768 dimensions) stored in Qdrant capture semantic meaning. The system understands that "genset won't turn over" relates to "engine fails to crank" even without keyword overlap.

Knowledge Graph Traversal — A Neo4j-backed graph captures structural relationships that vector similarity alone cannot represent: equipment hierarchies (Manufacturer → Model → System → Component → Part), diagnostic chains (Fault Code → Symptoms → Root Causes → Resolutions), document relationships, and regulatory mappings.

An intent-driven orchestrator classifies each query to select the optimal retrieval strategy:

Query Type	Strategy
Specification lookup	Vector-dominant, exact values prioritized
Procedure request	Vector + section navigation
Troubleshooting	Graph traversal + vector enrichment
Code compliance	Graph navigation + vector
Equipment comparison	Multi-query with structured merge

The orchestrator supports multi-step retrieval with up to 5 iterative rounds, clarifying questions when critical information is missing, and conflict detection when sources disagree.

Community-Driven Learning

Rudy implements 16 interconnected feedback loops across six domains: content quality, retrieval optimization, human intelligence, predictive capability, safety, and personalization. Every interaction generates learning

signals that improve future answers—not through periodic retraining, but through continuous feedback integration.

Contributor Authority — Five tiers (Novice through Master) weight corrections by demonstrated expertise. A correction from a master technician with 20 years of experience carries more weight than one from a first-year apprentice. Authority is earned through approved corrections, quality feedback, and peer validation—per-tenant, so expertise in one domain doesn't bleed into another.

Knowledge Gap Detection — The system identifies what it can't answer well by clustering negative feedback by topic, ranking gaps by frequency and severity, and surfacing prioritized gaps to administrators with actionable recommendations.

Automated Quality Monitoring — Continuous metric tracking triggers automated responses when quality degrades: thumbs-down rate spikes trigger prompt review, escalation rate increases trigger retrieval strategy review, correction patterns trigger knowledge base review.

Safety-First Design

Safety warnings are mandatory components of every relevant response, not optional add-ons. The system detects hazards across five categories—electrical, chemical, confined space, hot work, and lockout/tagout—and injects appropriate warnings regardless of whether the technician asked about safety. Every answer includes complete citation chains: document name, section, publication date. Every knowledge modification is logged with full provenance for regulatory compliance.

Multi-Tenant Platform

Rudy operates as a configurable multi-tenant platform serving any industry. New tenants go through an adaptive AI-driven onboarding process: Rudy crawls the tenant's website to extract business context, conducts a targeted AI interview to fill gaps, then auto-generates a complete configuration—system prompt, categories, safety rules, RAG tuning parameters, and starter questions. Credits, authority, knowledge, and retrieval settings are all tenant-scoped.

Part II: The Next Frontier — Multimodal Perception

Why Documents Aren't Enough

The current system excels at retrieving and synthesizing information from text: manuals, bulletins, SOPs, codes, tribal knowledge captured as written corrections. But the physical world communicates through channels that text cannot capture:

- A bearing that's about to fail **sounds** different three days before it seizes
- Exhaust smoke **color** tells an experienced technician what's wrong faster than any diagnostic tree
- A corroded connection is obvious in a **photo** but invisible in a service ticket description
- A vibration profile **shift** of 0.002 inches/second is the difference between normal operation and impending catastrophic failure
- A coolant temperature that's been **trending** up 2°F/week for a month signals a problem no single reading would reveal

An experienced senior technician perceives all of these signals instinctively. If Rudy is to truly replace the expertise that walks out the door when these people retire, it must learn to perceive the physical world—not just read about it.

Horizon 1: Visual and Acoustic Intelligence

Visual Diagnostics

Photo-based fault identification — A technician photographs a controller display showing a fault code, a corroded terminal block, or an unfamiliar component. The system identifies what it's seeing using vision-language models fine-tuned on industrial equipment imagery, then retrieves relevant documentation and community knowledge to deliver a diagnosis with resolution steps.

The technical challenge is not generic image recognition—it's domain-specific visual understanding. A photo of a DSE 8610 controller screen requires understanding the specific layout, alarm codes, and indicator meanings for that controller model. A corroded terminal must be recognized not just as "corrosion" but as a specific failure mode with known causes, remediation procedures, and safety implications.

Annotated image knowledge base — Every photo uploaded builds a searchable visual reference library scoped to the tenant's equipment. Images are tagged with fault codes, equipment models, and resolution outcomes.

Video-based procedure verification — A technician records a maintenance procedure on video. The system verifies each step against the SOP, flags missed steps and safety violations, and captures correct technique as training material for future technicians.

Acoustic Signature Analysis

Baseline learning — The system ingests audio recordings of equipment operating normally, building per-asset acoustic profiles. Each piece of equipment has a characteristic sound signature that varies by load, ambient temperature, operating hours since last service, and other factors.

Deviation detection — When a technician records audio of equipment they suspect is malfunctioning, the system compares the current acoustic signature against the learned baseline. Spectral analysis identifies frequency components associated with specific failure modes: bearing wear produces characteristic broadband noise increases; misalignment creates specific harmonic patterns; cavitation in pumps has a distinctive high-frequency signature.

Community-validated acoustic patterns — When a deviation is detected and subsequently confirmed as a specific failure mode through outcome tracking, that acoustic-to-failure mapping enters the knowledge base. Over time, the system builds a library of "this is what [failure X] sounds like on [equipment Y]"—knowledge that currently exists only in the ears of experienced technicians.

Horizon 2: Physical World Data Ingestion

The Data That Already Exists

Industrial equipment is already generating enormous volumes of operational data. SCADA systems, building management systems, genset controllers, PLCs, and IoT sensor networks produce continuous streams of temperature readings, pressure curves, vibration profiles, fuel levels, run-hour counters, and hundreds of other parameters. The vast majority of this data goes unread by humans—it's logged, maybe archived, and rarely analyzed until something has already failed.

IoT and Sensor Integration

Protocol support — MQTT, OPC-UA, Modbus TCP/RTU, BACnet, and REST-based APIs cover the major industrial communication protocols. Rudy ingests data from SCADA systems, BMS platforms, and IoT sensor networks without requiring replacement of existing infrastructure.

Continuous ingestion and correlation — Temperature, pressure, flow, vibration, fuel level, and other telemetry streams are ingested in real time, time-series indexed, and correlated with equipment identity from the knowledge graph. A temperature reading isn't just a number—it's a reading from a specific sensor on a specific component of a specific unit with a specific maintenance history and a specific set of OEM-defined operating limits.

Controller and PLC Data

Genset controllers — Direct integration with DSE, ComAp, DEIF, and other genset controllers provides real-time access to fault codes, operating parameters, and event logs. Fault codes are interpreted immediately against the knowledge base.

Industrial PLCs — Data ingestion from Allen-Bradley, Siemens, and Modbus-connected PLCs extends coverage to manufacturing, process control, and facility automation equipment.

Analog Signal Interpretation

Legacy instrumentation—4–20mA current loops, thermocouple outputs, RTDs, pressure transducers—represents the majority of installed sensing infrastructure in many facilities. Rudy learns the normal operating envelope for each sensor on each unit. A 4–20mA signal representing 0–100 PSI on a specific oil pressure transducer has meaning only in context: what equipment it's on, what the OEM specification says, what the actual operating range has been historically, and whether it's been trending.

Digital Twin Construction

The convergence of documentation, sensor data, maintenance history, and community knowledge enables construction of a living digital representation of each asset—a digital twin that reflects not just the manufacturer's design specifications but the actual, current operational state of the equipment.

A digital twin in this context is not a 3D visualization. It's a knowledge object that answers: "What do we know about this specific unit right now?" The answer combines:

- OEM specifications and procedures (from ingested documentation)
- Current operating parameters (from sensor data)
- Maintenance history (from work orders and technician interactions)
- Known quirks and field fixes (from community knowledge)
- Active alerts and predicted issues (from analytical models)

Horizon 3: Predictive Intelligence

Horizons 1 and 2 give Rudy eyes, ears, and a nervous system. Horizon 3 is the brain that makes sense of it all—using accumulated operational data and maintenance history to shift from reactive support ("what's wrong?") to predictive intelligence ("what's about to go wrong?").

Failure Pattern Recognition

Signal fusion — Cross-reference sensor trends, maintenance logs, query patterns, and community-reported issues across thousands of similar assets. When coolant temperature trending up 2°F/week coincides with a vibration shift on a specific bearing and the last three technicians who worked on similar units asked about water pump symptoms—those signals, individually unremarkable, collectively indicate an impending water pump failure.

Pattern library — Confirmed failure-precursor patterns enter a pattern library, scoped by equipment type and operating context. Privacy is maintained through rigorous anonymization—patterns are only surfaced when

detected across 5+ distinct equipment instances from 3+ different users.

Proactive alerting — When a unit's current data matches a known failure-precursor pattern, the system alerts the operator with the specific failure risk, confidence level, recommended inspection or preventive action, and supporting evidence from similar past events.

Condition-Based Maintenance Scheduling

The most direct economic impact of predictive intelligence is the shift from time-based to condition-based maintenance. Current practice: change the oil every 500 hours, replace the air filter every 1000 hours, overhaul the turbo every 10,000 hours—regardless of actual condition.

Condition-based scheduling uses operating data to determine when maintenance is actually needed. The economic case: unnecessary maintenance is expensive (parts, labor, downtime), but missed maintenance is catastrophic (unplanned failure, extended outage, safety risk). Condition-based scheduling optimizes the tradeoff using data instead of conservative generic intervals.

Fleet-Wide Anomaly Detection

When a new failure mode emerges on one unit, the system checks whether any other units in the fleet show early signs of the same pattern. This is particularly valuable for systematic issues—a bad batch of fuel filters, a firmware bug in a controller update, a design flaw that manifests only under specific operating conditions. The first failure is expensive; the second through fiftieth are preventable.

Autonomous Diagnostic Agents

Complex failures require multi-step diagnostic reasoning. Rudy's diagnostic agents conduct structured troubleshooting sessions:

1. Analyze initial symptoms from the technician's description, photos, and available sensor data
2. Formulate and rank hypotheses using the knowledge graph's diagnostic chains
3. Request specific measurements or observations to discriminate between hypotheses
4. Narrow to root cause through iterative evidence gathering
5. Deliver a diagnosis with resolution procedure, cited sources, and confidence level

Horizon 4: Frontier Applications

Multimodal Fusion Reasoning

Horizons 1 through 3 create separate perception channels: text, vision, audio, telemetry. The frontier capability is fusing these channels into unified situational awareness.

> *"The unit sounds different AND the oil pressure is trending down AND the last three technicians asked about this bearing AND the acoustic signature matches a pattern we've seen precede bearing failure on similar units AND the OEM bulletin from last month mentioned a bearing supplier quality issue."*

No single signal is definitive. The fusion of five weak signals produces a strong, actionable conclusion. This multimodal reasoning—combining what Rudy reads, sees, hears, and senses into a single analytical framework—is where the system transcends any individual capability.

Reinforcement Learning from Operations

By tracking outcomes through the existing feedback and outcome-tracking infrastructure, the system continuously optimizes its recommendations based on what works in the real world—not just what the manual says. Which repair approach results in longer MTBF? Which troubleshooting sequence resolves fastest? Which

parts substitution works as well as OEM at lower cost?

Edge-Deployed AI

Field service often happens where connectivity doesn't: remote wellhead sites, offshore platforms, underground utility vaults, rural substations. Edge deployment means running inference on ruggedized edge devices—distilled models optimized for local execution that sync knowledge updates when connectivity is available. The same diagnostic agents, the same knowledge base (cached locally), the same safety warnings—available with no internet.

Natural Language Equipment Control

The ultimate extension: from "read" to "write." Today, Rudy reads data from equipment and provides guidance. The frontier is interpreting a technician's intent, generating the correct control sequence, requiring explicit human confirmation, and executing—safely adjusting setpoints, initiating test sequences, or performing controlled shutdowns. This capability requires the most rigorous safety engineering of any feature on the roadmap: human-in-the-loop confirmation for every action, equipment-specific safety limits that cannot be overridden, and complete audit logging.

Technical Architecture

Current Production Stack

Layer	Technology	Purpose
Frontend	Next.js 14 + React	Admin dashboards, technician interface
API	tRPC + Node.js	Type-safe API with streaming support
Database	PostgreSQL (Drizzle ORM)	Users, queries, documents, feedback, corrections
Vector Store	Qdrant	Semantic search across document chunks
Knowledge Graph	Neo4j	Equipment hierarchies, diagnostic chains
Cache	Redis	Query result caching, session management
Ingestion	Python FastAPI + Docling	PDF parsing, semantic chunking, classification
LLM	OpenAI + Anthropic	Answer synthesis with provider failover
Embeddings	Google Gemini-001	768-dimensional dense vectors
Infrastructure	Docker Compose	All services orchestrated with health checks

Architecture Extensions for Physical World AI

Multimodal Ingestion Pipeline — Extending the current document ingestion pipeline to process images (equipment photos, annotated diagrams), audio (acoustic recordings with spectral analysis), and video (procedure recordings with frame-level analysis). Each modality gets its own embedding space with cross-modal alignment.

Time-Series Data Infrastructure — Sensor telemetry requires purpose-built time-series storage (TimescaleDB or InfluxDB), stream processing for real-time anomaly detection, and integration with the existing knowledge graph to provide semantic context for numerical data.

Edge Inference Runtime — Distilled models packaged for execution on edge hardware (NVIDIA Jetson, Intel NUC, or equivalent ruggedized platforms). Knowledge base subsets cached locally and synchronized when connectivity permits.

Safety-Critical Design Across All Horizons

As Rudy extends into perceiving and acting on the physical world, safety engineering becomes exponentially more important. The principles established in the text-based system—transparent provenance, mandatory safety warnings, full audit trails—extend to every new modality:

Visual safety — When photo analysis identifies a safety-relevant condition (exposed conductors, missing guards, chemical spills), safety warnings are elevated to the highest priority regardless of what the technician asked about.

Sensor safety — Operating parameter thresholds are defined per-asset from OEM specifications and enforced automatically. When a sensor reading exceeds safe limits, the system generates an immediate alert.

Predictive safety — Failure predictions carry safety classifications. A predicted bearing failure on a non-critical fan is a maintenance planning item. A predicted failure on a fuel system component is a safety-critical alert requiring immediate action.

Control safety — Any future equipment control capability implements defense-in-depth: human-in-the-loop confirmation for every action, equipment-specific safety limits that cannot be overridden, automatic rollback on unexpected responses, and complete audit logging.

Expansion to Other Domains

While developed for power generation, Rudy's architecture is domain-agnostic. The same principles—knowledge retrieval, community learning, multimodal perception, predictive intelligence—apply wherever physical equipment requires expert diagnosis, knowledge is fragmented, regulatory compliance intersects with practical procedures, and operational data exists but is underutilized.

Domain	Knowledge Crisis	Multimodal Value	Predictive Value
HVAC	High tribal knowledge	Acoustic: compressor health. Visual: refrigerant leaks	Refrigerant pressure trends, energy anomalies
Medical Equipment	Very high regulatory	Visual: error displays, wear indicators	Sensor drift, calibration degradation
Manufacturing	High tribal, many OEMs	Acoustic: motor/pump health. Visual: wear patterns	Vibration trends, quality correlation
Oil & Gas	High tribal, high regulatory	Acoustic: valve/pump cavitation. Visual: corrosion	Pressure/flow trends, wellhead performance

Domain	Knowledge Crisis	Multimodal Value	Predictive Value
Aviation	Very high regulatory	Visual: structural inspection. Acoustic: engine health	Flight data trends, component lifecycle
Marine	Very high tribal knowledge	Acoustic: engine/hull. Visual: corrosion, wear	Engine performance curves, hull fouling rates

Expanding to a new domain requires: domain knowledge seeding (terminology, equipment hierarchies), initial document corpus, safety pattern configuration, and authority hierarchy calibration. The core retrieval pipeline, feedback system, multimodal processing, and admin tooling transfer without modification.

Conclusion

The field service industry's knowledge crisis is real, accelerating, and not solvable by better documentation or simple chatbots. The expertise that keeps critical infrastructure running—the pattern recognition, the equipment quirks, the diagnostic intuition built over decades—is walking out the door as experienced technicians retire.

Rudy's approach is to build a system that captures this expertise across every channel it exists in. Text, certainly—manuals, SOPs, tribal knowledge captured as corrections. But also the visual patterns that experienced eyes recognize instantly, the sounds that trained ears distinguish effortlessly, the sensor trends that seasoned operators notice intuitively. And then to extend beyond human perception into domains where machines have inherent advantages: continuous monitoring of hundreds of parameters simultaneously, pattern recognition across thousands of similar assets, predictive modeling that identifies failures days or weeks before they manifest.

The current production system—self-learning knowledge retrieval with community-driven authority, safety-aware synthesis, and multi-tenant configurability—is the foundation. The roadmap—multimodal perception, physical world data ingestion, predictive intelligence, and frontier AI applications—is where the platform becomes something genuinely new: not a better manual, not a better search engine, but an operational intelligence system that perceives, learns, predicts, and acts.

> *The future of field service support is not better answers to technician questions. It's a system that sees what your equipment is telling you, hears what your technicians hear, knows what your best people know, and tells you what's coming next.*

We invite collaboration with organizations interested in exploring how multimodal AI applied to physical equipment can transform their operations.

About the Author

Nick Haschka is CEO of OnPoint Generators, Inc., a field service organization specializing in emergency backup power systems. With over two decades of experience in power generation service, Nick has witnessed firsthand the knowledge fragmentation and expertise loss that motivated the development of Rudy—and the operational reality that inspired its evolution from a knowledge retrieval system to a platform for physical-world AI.

Contact

Nick Haschka
CEO, OnPoint Generators, Inc.
Email: nick@onpointgen.com
Phone: 408-581-0885

References

1. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS 2020*.
2. NFPA 110: Standard for Emergency and Standby Power Systems. National Fire Protection Association.
3. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP 2019*.
4. Es, S., et al. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint*.
5. Peng, Z., et al. (2023). Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint*.
6. Girdhar, R., et al. (2023). ImageBind: One Embedding Space to Bind Them All. *CVPR 2023*.
7. Mobley, R. K. (2002). An Introduction to Predictive Maintenance. *Butterworth-Heinemann*.
8. Grieves, M. (2014). Digital Twin: Manufacturing Excellence through Virtual Factory Replication. *White Paper, Florida Institute of Technology*.