The Engineering Behind the AI That Gets Smarter Every Day

How Rudy's self-learning architecture creates compounding advantages that are extraordinarily difficult to replicate

Nick Haschka / CEO, OnPoint Generators, Inc.

May 2026

How Rudy's self-learning architecture creates compounding advantages that are extraordinarily difficult to replicate

Executive Summary

Field service industries face an accelerating knowledge crisis. Experienced technicians are retiring faster than they can be replaced. The training pipeline is structurally broken — the people best qualified to teach are the ones organizations can least afford to pull off revenue-generating work. And generic AI tools, however impressive in consumer settings, fail at the point of application: they don't understand trade terminology, can't follow diagnostic chains, and treat a safety-critical wiring procedure the same as a routine FAQ.

Rudy is a production platform purpose-built to solve this problem. What's deployed and working today: three-source hybrid retrieval combining vector search, full-text keyword search, and knowledge graph traversal; an Instructed Retriever that generates multi-step search plans rather than single-query lookups; iterative retrieval with automatic gap detection and query reformulation; query-focused context compression; a self-healing knowledge loop that detects gaps, researches autonomously, and ingests results; 16 self-reinforcing learning loops; a full safety architecture covering 8 hazard categories with per-tenant calibration; and a multi-tenant platform serving 3,000+ documents across 127 manufacturers and 187 equipment models.

What makes this a moat rather than a product: four compounding layers that widen the lead every day the platform operates. Individual compounding — the more a technician uses Rudy, the more Rudy understands their equipment context and expertise level. Community compounding — every correction weighted by authority, every implicit feedback signal, every tribal knowledge tip makes the knowledge base more accurate without requiring explicit effort. Content compounding — user queries surface gaps, the self-healing loop fills them autonomously, and the flywheel accelerates. Cross-trade compounding — diagnostic patterns, safety frameworks, and learning infrastructure transfer structurally to adjacent verticals, so each new trade inherits from every prior one.

The architecture is domain-agnostic. Power generation is the proving ground. Diesel mechanics, marine, heavy equipment, forklifts, pumps, cranes, refrigeration, industrial electrical, elevator, and HVAC are the expansion sequence — ordered by knowledge domain overlap so each trade inherits from the last. A competitor starting today faces not just an engineering gap but a compounding data and trust gap that widens every month.

Part I: The Problem

1. Expertise Is Disappearing

When a 30-year power generation technician retires, they take with them a form of knowledge that no documentation system captures. Pattern recognition developed over thousands of service calls. Equipment-specific quirks that never made it into any manual. Diagnostic shortcuts that turn four-hour jobs into one-hour jobs. Safety insights learned from near-misses that were never written down because the experienced tech knew to avoid the situation in the first place.

The scale of this retirement wave is not a theoretical concern:

- The IEA reports 2.4 energy workers nearing retirement for every new entrant under 25
- Goldman Sachs estimates 750,000+ new power industry workers will be needed by 2030
- The Department of Labor projects nearly half the current power workforce will retire within a decade

Knowledge is fragmented across manufacturer manuals (often outdated), service bulletins (scattered across dealer portals), regulatory codes (paywalled), and individual experience (leaves when the person leaves). No existing system captures the intersection of all four.

2. The Training Pipeline Is Structurally Broken

The expertise crisis would be manageable if training worked. It doesn't — not at scale, not at the margins required.

The fundamental problem is economic. The best trainer in any field service organization is the master technician. That person costs $85 to $120 per hour in loaded labor. Sending them to ride along on a routine call that a junior tech should be able to handle alone destroys the economics of that truck roll. So organizations don't do it. Juniors learn through trial and error, callbacks accumulate, and customer satisfaction erodes.

The problem compounds in several directions:

Resentment. Senior technicians signed up for complex diagnostic work — the problems that require genuine expertise. Babysitting a junior through a 40-point load bank test is not what they trained for. Organizations that assign training duty to their best people often lose them to competitors who don't.

Transfer failure. Classroom training and online modules teach procedures in the abstract. Field competence requires doing, with someone experienced watching and correcting. A junior who can recite the steps for a transfer switch commissioning from a slide deck may freeze in front of an unfamiliar panel. The knowledge doesn't transfer until it's applied under observation.

Availability mismatch. Emergencies don't schedule themselves. A junior tech facing an unexpected fault condition at 2 AM needs the master tech's knowledge, not a call center. The master tech is unavailable, or unavailable for this level of question, or unavailable at this hour.

This is why Rudy matters at a structural level beyond "better search." Rudy is the senior technician riding along on every job, available 24/7, infinitely patient with repeated questions, never resentful of the interruption, and costing nothing incremental per truck roll. It doesn't replace the master technician. It makes the master technician's knowledge available at the moment and location where the junior technician needs it.

3. Why Generic AI Fails

The market's first response to this problem was to point field service organizations at general-purpose AI assistants. These tools are impressive in consumer contexts. They fail in technical field service for reasons that are architectural, not superficial:

Single-pass retrieval. Standard RAG (Retrieval-Augmented Generation) systems embed the query, find similar documents, and synthesize an answer in a single pass. Real diagnostic questions don't work this way. "The QSB7 won't start after a coolant flush — what did we miss?" requires identifying the symptom, correlating with known failure patterns on that engine family, checking the maintenance procedure for common errors, and cross-referencing the coolant system's interaction with the air bleeding procedure. A single retrieval pass can't follow that chain.

Generic embeddings. When a technician queries "DSE 8610 Alarm 1437," the embedding must find an exact match against that controller's fault code table — not a semantically similar passage about alarm conditions in general. Generic models trained on internet text treat trade-specific alphanumeric identifiers as noise.

No expertise weighting. Standard RAG treats a manufacturer service bulletin and a Reddit post as equally authoritative. A field service knowledge system must weight sources by provenance, recency, and the authority of contributors.

No safety awareness. A query about replacing a starter motor on a large genset involves working adjacent to a battery bank capable of delivering thousands of amperes. Generic AI doesn't know this. It answers the mechanical question and omits the electrical safety context entirely. In a safety-critical domain, that omission is not a quality issue. It is a liability.

Hallucination in safety-critical domains. Generic AI systems are calibrated to provide confident, helpful answers. In field service, a confidently wrong torque specification or an incorrectly recalled wiring diagram can cause equipment damage, downtime, or injury. The error cost is asymmetric — a "I don't have enough information" answer is recoverable; a confident wrong answer may not be.

Part II: What's Built

Overview

Rudy's production system is not a wrapper around a general-purpose model. It is a purpose-built retrieval and reasoning platform with 16 self-reinforcing learning loops, a self-healing knowledge pipeline, and a safety architecture that treats hazard detection as an architectural constraint rather than a feature.

The following subsections describe each major component. The emphasis throughout is engineering depth and replication difficulty — not just what the system does, but what it would take to reproduce it.

Three-Source Hybrid Retrieval

Rudy retrieves evidence from three complementary sources simultaneously, each capturing a different dimension of relevance that the others cannot:

Vector Search (Qdrant, Gemini Embedding-001, 768 dimensions) — Dense embeddings capture semantic meaning independent of exact terminology. The system understands that "genset won't turn over" relates to "engine fails to crank" without keyword overlap. Embeddings are cached in Redis with a 7-day TTL to eliminate redundant computation. Embeddings are batch-processed with automatic retry logic (3 attempts, exponential backoff).

Keyword Search (PostgreSQL BM25) — Full-text search captures exact terminology matches that semantic similarity misses. When a technician queries a specific part number, fault code, or model designation, lexical matching finds it directly. "DSE 8610 Alarm 1437" must match the fault table entry for that exact alarm on that exact controller — semantic proximity is insufficient.

Knowledge Graph Traversal (PostgreSQL recursive CTEs) — Structural relationships that neither vector nor keyword search can represent: equipment hierarchies (manufacturer → model → system → component → part), diagnostic chains (fault code → symptoms → root causes → resolutions), and cross-document entity relationships. The graph contains 120,000+ relationships with deterministic entity UUIDs and cycle-safe

traversal up to 4 hops deep.

Reciprocal Rank Fusion (RRF) merges results from all three sources into a unified ranking without requiring comparable scores across heterogeneous retrieval systems. Graph results receive a **1.5x boost** during rank fusion, reflecting the high precision of structural traversal for technical queries. All three sources are configured as non-fatal: if any single source fails, retrieval continues with the remaining sources.

The Instructed Retriever

Standard RAG converts a user question into a single embedding lookup. The Instructed Retriever treats retrieval as a multi-step reasoning problem.

When a query arrives, the system generates a **search plan** — a structured strategy that decomposes complex questions into sub-queries, specifies which retrieval sources to emphasize for each sub-query, identifies entity references requiring resolution, and defines what constitutes sufficient evidence for a confident answer.

A query like "What's the oil change procedure for a Cummins QSB7 and what's different about the marine variant?" generates a plan that decomposes into two sub-queries (standard procedure + variant differences), routes each to the appropriate combination of retrieval sources, and defines a merge strategy that presents the standard procedure with marine-specific annotations.

The search plan includes metadata filters — manufacturer, document types, keyword terms — that are wired into both Qdrant vector search and PostgreSQL BM25 keyword search. Reranking is **intent-aware**: a `SPEC_LOOKUP` query emphasizes numerical values; a `TROUBLESHOOT` query emphasizes fault codes and resolution steps. This per-intent scoring transforms reranking from a generic relevance operation into a domain-appropriate judgment.

The planning layer transforms retrieval from statistical similarity into a reasoning process that mirrors how an experienced technician would search for information across a library of sources.

Iterative Retrieval with Gap Detection

When the first retrieval pass doesn't yield sufficient evidence, Rudy automatically detects the gap type and reformulates:

Gap Type	Condition	Response
`vector_empty`	No vector search results	Broaden semantic query, remove filters
`weak_matches`	Low similarity scores	Reformulate with alternative terminology
`graph_empty`	No graph traversal results	Expand search radius, try alternate entry points
`insufficient_evidence`	Low confidence in available evidence	Decompose query, search for supporting sub-topics
`filter_over_constrained`	Too-restrictive metadata filters	Fast path: drop manufacturer filter, retry immediately

The loop runs up to 3 rounds within a 15-second total timeout. The `filter_over_constrained` fast path handles the common case where a technician asks about a specific manufacturer but the relevant cross-reference appears in a multi-manufacturer technical document. Context is preserved across rounds — evidence accumulates rather than restarting.

Gap analysis reads a structured **retrieval manifest** logged per query: source counts, similarity scores, merge events, and boost events. This manifest is the prerequisite for gap detection — the loop is not guessing whether retrieval failed but reading structured diagnostics that describe exactly why.

Context Compression (KARL)

Real-world technical queries often retrieve more evidence than can be synthesized in a single pass. The Knowledge-Aware Retrieval with Lossy compression (KARL) system activates when more than 10 chunks are retrieved, compressing them into a dense, query-focused narrative using Gemini Flash Lite. The maximum synthesis chunk count is 20.

The compression pipeline preserves citation metadata through the compression process. Compressed evidence maintains full traceability to source documents, sections, and page numbers — a regulatory requirement in environments governed by OSHA, NFPA, and EPA, where demonstrating the provenance of every piece of guidance is not optional.

Self-Healing Knowledge Loop

The self-healing loop is one of the more architecturally unusual components of the system. It closes the feedback loop between quality evaluation and knowledge acquisition automatically, without human intervention:

1. The LLM-as-Judge quality pipeline runs against 96 golden questions organized across 4 complexity tiers
2. Answers that receive an F grade are flagged as knowledge gaps
3. The gap triggers an autonomous research job: Gemini with `google_search` grounding researches the topic, generating a synthesis document
4. The synthesis document is ingested into the knowledge base with `synthesized_research` authority tier (0.75)
5. A regression quarantine verifies that the new content doesn't degrade scores on previously passing questions
6. The `research_topics` table prevents duplicate research on the same gap

Since activation, 30+ research documents have been auto-generated and ingested. The self-healing loop runs automatically after every golden test cycle, converting F-grade gaps to research targets without any manual trigger.

A stale job reaper complements this: jobs stuck in processing for more than 30 minutes are automatically reset; permanent failure is declared after 3 crashes. The ingestion pipeline heals itself.

16 Self-Reinforcing Learning Loops

Most AI systems are static after deployment — they answer the same way on day 1,000 as on day 1. Rudy is designed so that every interaction makes the next interaction better, without retraining and without human intervention. Sixteen learning loops operate continuously across six domains:

Retrieval gets sharper. Every citation a technician clicks is a relevance signal. Chunks with consistently high click-through rates for related queries receive ranking boosts. The system converges toward surfacing the most useful content for each query type — automatically, from usage.

Expert knowledge rises to the top. A 5-tier authority system tracks who contributes reliable corrections and who doesn't. A master tech's correction carries more weight than an apprentice's guess. Approved corrections

boost authority; rejected ones penalize it. The knowledge base converges toward truth, not toward volume.

Safety calibrates itself. Too many warnings cause alert fatigue; too few miss genuine hazards. A proportional controller adjusts per-tenant safety sensitivity based on how technicians respond to warnings. Each organization converges on the right balance for their specific equipment and risk profile.

The system learns your language. Jargon, abbreviations, and equipment aliases are learned from real query patterns. When technicians consistently use shorthand the system doesn't recognize, it learns the mapping. The domain dictionary grows from usage.

Rudy adapts to each technician. Query patterns reveal skill level — the system infers whether to give apprentice-level detail or expert-level brevity. Equipment auto-detection learns which makes and models each technician works on regularly.

Quality improves invisibly. Ingestion quality scoring catches and reprocesses low-quality document chunks. Answer synthesis analysis measures satisfaction by intent type and tunes prompts accordingly. Community pattern detection surfaces emerging knowledge gaps before they accumulate into clusters of bad answers.

Feedback flows without effort. Technicians rarely click thumbs-up/thumbs-down — they're busy. So the system detects behavioral signals from follow-up messages: corrections ("actually it's..."), frustration (repeated questions, all-caps), confusion ("what do you mean"), and satisfaction ("that worked, thanks"). These signals flow continuously at zero added cost, creating a feedback stream that requires no explicit action from the technician.

The critical property of these loops is that they compound. Each loop generates data that other loops consume. Retrieval improvements surface better content, which generates more positive feedback, which strengthens authority signals, which improves the quality of corrections, which makes retrieval better. The system doesn't just learn — it learns faster the more it has already learned.

Safety as Architecture

Safety is not a layer applied after the core retrieval system works. It is an architectural constraint that shapes every component.

Safety analysis runs on every query and every retrieved chunk using pattern-matched detection across 8 hazard categories:

Category	Default Severity
Electrical (high voltage, arc flash, energized equipment, exposed conductor)	Danger
Mechanical (rotating equipment, pinch points, crush hazard, entanglement)	Danger
Fire (flammable materials, hot work, explosive atmosphere, flash point)	Danger
Pressure (pressure vessels, overpressure, rupture, safety valve)	Danger
Lockout/Tagout (LOTO, energy isolation, zero energy, stored energy release)	Danger
Chemical (hazardous materials, toxic, corrosive, inhalation hazard)	Warning
Environmental (spills, hazardous waste, groundwater contamination)	Warning
PPE (personal protective equipment, fall protection, respirator)	Caution

Detection uses pre-compiled regex with case-insensitive word boundary matching for real-time performance. The risk level is calculated from the combination of detected hazards: two or more Danger-level detections produce Critical overall risk; one Danger produces High risk.

Safety warnings are mandatory. A query about "replacing the starter motor" that involves working near a battery bank will include electrical safety warnings even though the question was mechanical. The hazard is injected regardless of whether the technician asked about it.

Per-tenant calibration uses a proportional controller: if technicians are dismissing more than 30% of safety warnings (alert fatigue), sensitivity decreases. If dismissal rate is below 30%, sensitivity increases. The adjustment threshold is 0.05 — changes smaller than 5% are not applied, preventing oscillation. Calibration requires a minimum of 50 interactions before activating.

Every safety-relevant interaction is logged: query content, user identity, tenant scope, detected categories and severities, warning acknowledgment or dismissal, and full citation chains. This audit trail supports OSHA, NFPA 110, EPA, and other regulatory frameworks where demonstrating the accuracy and provenance of safety information is a legal requirement.

LLM-as-Judge Quality Evaluation

Automated quality evaluation runs continuously using an LLM-as-Judge pipeline scoring every answer across six dimensions:

Dimension	What It Measures
Factual Accuracy	Are the facts in the answer correct?
Safety Correctness	Are safety warnings appropriate and accurate?
Hallucination	Does the answer fabricate claims not supported by evidence?
Context Adherence	Does the answer stay grounded in the retrieved context?
Completeness	Does the answer cover all aspects of the question?
Actionability	Can a technician act on this answer in the field?

Hallucination detection is weighted most heavily because in a safety-critical domain, a confidently wrong torque specification or an incorrectly recalled wiring diagram can cause equipment damage or injury. The system is calibrated to strongly prefer "I don't have enough information" over a confident wrong answer.

The golden test suite contains 96 questions across 3 complexity tiers (S1 single-source, S2 two-source, ML multi-layer synthesis), including 15 cross-document questions that require synthesizing information from multiple sources. Multi-layer question generation uses Gemini 2.5 Pro; grading uses Flash for S1/S2 and Pro for ML. The tiered structure measures not just recall of individual document facts but synthesis capability across the full knowledge base.

Parallel General-Knowledge Fallback

A core design principle: Rudy must never give a worse answer than a base model. When RAG retrieval fails to surface the relevant chunk, the system should still answer correctly from base model knowledge rather than returning "not found."

The parallel fallback pipeline runs via `Promise.all()` — both the full RAG retrieval and a general-knowledge generation call launch simultaneously. The fallback adds zero latency because it completes while RAG retrieval is still running.

A second LLM call evaluates the fallback answer's accuracy, producing a confidence rating (HIGH, MEDIUM, or LOW) with reasoning. The synthesis prompt then merges all available knowledge with strict priority ordering:

1. **Tier 1 (highest):** Customer documentation — manufacturer manuals, spec sheets, bulletins
2. **Tier 2:** Graph knowledge — structured relationships from the knowledge graph
3. **Tier 3:** User-contributed knowledge — tribal knowledge tips, verified corrections
4. **Tier 4 (lowest):** General knowledge — base model answers with confidence indicators

The SSE streaming response includes `knowledgeSources` metadata showing which tiers contributed. When general knowledge is used, its confidence level is displayed so technicians can make informed judgments about verification.

Multi-Tenant Platform

Rudy operates as a fully configurable multi-tenant platform. Every database query, API call, cache key, and vector search is tenant-scoped. Data isolation is enforced at the infrastructure level — there is no code path through which one tenant's documents could appear in another tenant's search results.

Adaptive AI-driven onboarding reduces tenant setup from days to minutes: 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.

Per-tenant configuration includes: credits, authority tiers, safety sensitivity levels, knowledge extraction results, feedback loop parameters, and all retrieval parameters. A tenant operating a fleet of Caterpillar generators gets different defaults, different safety calibration, and different query patterns than one operating mixed-manufacturer industrial equipment.

Embodied Knowledge Extraction

Beyond retrieving from documents, Rudy extracts structured knowledge from its corpus and builds it into a queryable layer:

Glossary Terms — Abbreviations with full forms and definitions, extracted automatically from each ingested document. When a technician uses "LOTO" in a query, the system expands to "Lockout/Tagout" and searches for both forms.

Equipment Entities — Canonical equipment names with manufacturer aliases. "CAT C15," "Caterpillar C15," and "C-15 ACERT" all map to the same engine. At query time, a question about any name variant finds content tagged with all variants.

Extraction runs automatically when document ingestion completes (Pipeline B) and as a cross-document resolution batch (Pipeline C) that merges duplicate entities discovered across different documents.

Additional Production Capabilities

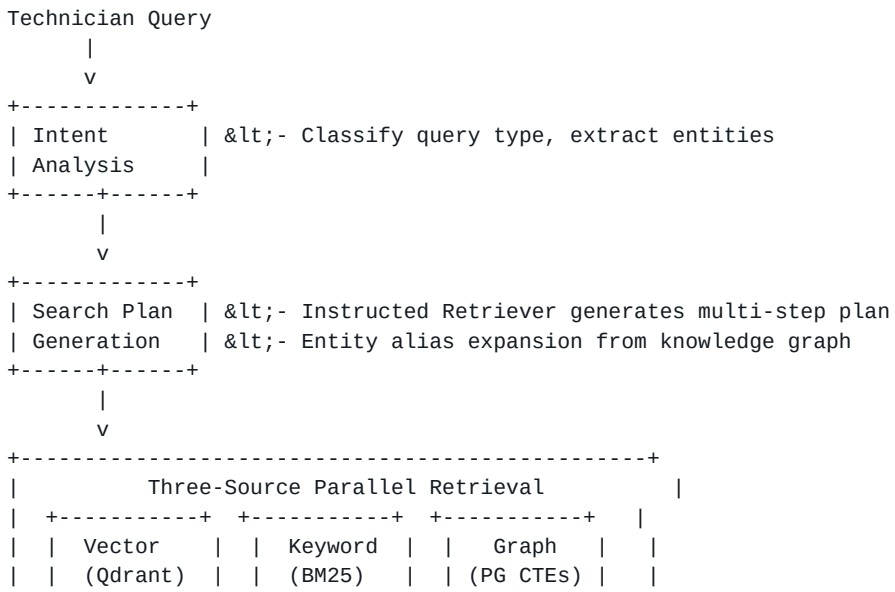
The following capabilities are deployed and working, listed for completeness:

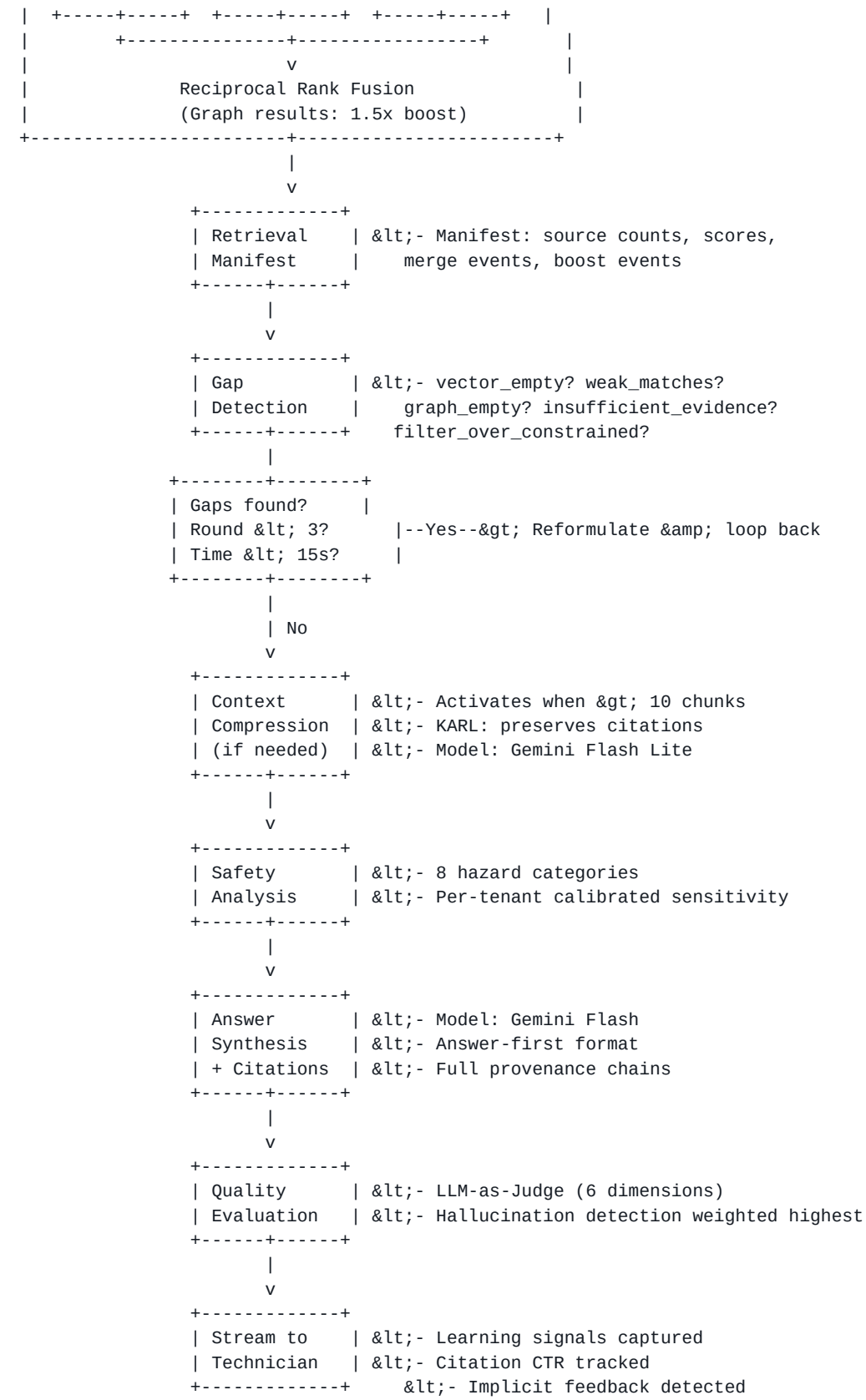
- 55+ admin API endpoints covering stats, corrections, learning, discovery, authority, knowledge, users, query analytics, and synthetic testing
- Customer success analytics with tenant-scoped metrics including active users, session counts, latency percentiles (avg/p50/p95), and satisfaction rates
- Query normalization: 80+ term domain dictionary, 27-cluster synonym expansion for BM25, equipment entity extraction for garbled queries, optional Gemini Flash Lite rewrite with truncated-query completion
- Confidence-gated validation routing uncertain queries through enhanced reasoning with re-retrieval
- Conversational follow-up with Redis conversation history, 20-message window
- Photo upload with client-side compression, GCS signed URL, and Gemini multimodal analysis
- Document acquisition pipeline discovering PDFs, downloading to GCS, and queuing for ingestion
- Query disambiguation: detects ambiguous queries and sends clarification hints while continuing to answer
- Async response groundedness evaluator: post-synthesis background verification that answer claims are grounded in retrieved chunks

Content Scale (May 2026)

Metric	Value
Documents in pipeline	3,000+
Manufacturers covered	127+
Equipment models	187
Curated foundational resources	180
Searchable chunks	58,000+
Knowledge relationships	120,000+

Query Processing Pipeline




Reciprocal Rank Fusion
(Graph results: 1.5x boost)
Retrieval Manifest
Manifest: source counts, scores, merge events, boost events
Gap Detection
vector_empty? weak_matches? graph_empty? insufficient_evidence? filter_over_constrained?
Gaps found? Round < 3? Time < 15s?
Yes
Reformulate & loop back
No
Context Compression (if needed)
Activates when > 10 chunks
KARL: preserves citations
Model: Gemini Flash Lite
Safety Analysis
8 hazard categories
Per-tenant calibrated sensitivity
Answer Synthesis + Citations
Model: Gemini Flash
Answer-first format
Full provenance chains
Quality Evaluation
LLM-as-Judge (6 dimensions)
Hallucination detection weighted highest
Stream to Technician
Learning signals captured
Citation CTR tracked
Implicit feedback detected

Part III: The Compounding Engine

This is the mechanism that converts a good product into a durable business. The engineering described in Part II is reproducible given time and talent. The compounding engine is not — because it requires users, and users generate the data that makes the next user's experience better, which attracts more users.

Four layers operate simultaneously:

Layer 1: Individual Compounding

The more a technician uses Rudy, the more Rudy understands them specifically.

Persona calibration infers expertise level from question phrasing and behavioral patterns. A technician who asks "what does Alarm 1437 mean on a DSE 8610" is at a different competence level than one who asks "the DSE is showing 1437 but the coolant temp looks normal — is there a sensor calibration drift known on QSB7s?" The system infers the difference and calibrates response detail accordingly. Over time, this calibration becomes increasingly accurate.

Equipment auto-detection tracks what a technician works on across sessions. The 24-hour staleness cutoff on auto-loaded equipment context means the system is always working with relevant fleet context, not stale data from a previous engagement. Cross-manufacturer poisoning is guarded against — if a tech switches from a Cummins to a Kohler job, the system detects the topic shift and resets context rather than contaminating the new query with the prior equipment's context.

Implicit feedback signal detection captures behavioral signals without any explicit rating effort. Message patterns indicating frustration (repeating the question, adding "I already tried that"), confusion (asking for clarification, requesting simpler language), or satisfaction (proceeding to follow-up questions rather than retrying) are detected via zero-cost regex and stored as continuous feedback signals.

The retention mechanism follows directly: switching to a competitor means starting from zero. No calibrated persona. No equipment context. No accumulated implicit signals. The accumulated understanding of a specific technician's expertise and fleet is not exportable.

Layer 2: Community Compounding

The 16 learning loops mean the platform improves autonomously from every interaction, not just from explicit feedback.

The contributor authority system ensures that quality converges toward truth rather than toward volume. A master technician's correction carries different weight than an apprentice's — the 5-tier points-based system reflects demonstrated expertise. Rejected corrections penalize authority; approved corrections build it. The knowledge base converges toward what the most experienced contributors validate, not toward what the most frequent contributors submit.

The safety calibration loop is a specific example of community compounding that has no analog in a static system. Different organizations have different risk tolerances, different equipment types, and different technician experience levels. A system with a single global safety sensitivity serves no organization well. The per-tenant proportional controller learns from actual dismissal patterns, converging on a calibration that reduces alert fatigue without sacrificing genuine hazard detection. That calibration is a learned artifact that took interactions to produce.

Citation click-through tracking creates a feedback signal that requires no explicit action from the technician. Every time someone clicks through to view a source document, that signal is recorded. Chunks with consistently high CTR for related queries are boosted in future rankings. The ranking system improves from usage without asking anyone to rate anything.

A competitor can build retrieval. They cannot purchase 10,000 expert-weighted corrections, accumulated calibration signals, and citation engagement histories. That data exists only as a byproduct of operating the platform at scale.

Layer 3: Content Compounding

The self-healing knowledge loop creates a flywheel between usage and content quality.

User queries are the detection mechanism. When a query against the golden test suite produces an F-grade answer, that is evidence of a knowledge gap. The gap triggers autonomous research — Gemini with `google_search` grounding produces a synthesis document. The document is ingested, regression-quarantined, and becomes part of the knowledge base. The next time a technician asks a similar question, the gap is filled.

But the flywheel is broader than the golden test suite. More users generate more query patterns. More query patterns surface more gap types. More gaps trigger more research. More research documents improve answer quality. Better answer quality attracts more users.

Document collection agents (in progress) extend this further: a Search Agent autonomously searches the web for relevant technical documents; a Browse-Along Agent collects materials from authenticated sites as technicians navigate them. Users with journeyman or higher authority can upload manuals directly. Every inbound document expands coverage, which improves answers for the entire user base.

The content flywheel has no natural ceiling. There is always more technical documentation to acquire, more failure patterns to document, more tribal knowledge to capture. The larger the knowledge base, the more comprehensive the gap detection, and the more precisely new research can target what's actually missing.

Layer 4: Cross-Trade Compounding

The architecture is domain-agnostic. The retrieval pipeline, safety framework, authority system, learning loops, quality evaluation, and multi-tenant platform all transfer structurally to adjacent trades.

The cross-trade compounding is not metaphorical — it is mechanical. A Cummins ISX in an emergency generator is the same engine as a Cummins ISX in a Peterbilt truck. Hydraulics failure modes on an excavator share root-cause patterns with hydraulics on a forklift and a crane. HVAC refrigerant circuit diagnostics follow the same logical structure as generator cooling system diagnostics. Safety frameworks for electrical hazards in power generation transfer directly to electrical hazards in HVAC and industrial equipment.

The expansion economics compound with each vertical:

- **First vertical (power generation):** Build everything — retrieval infrastructure, safety framework, authority system, learning loops, quality evaluation, multi-tenant platform, content for the specific trade
- **Second vertical (HVAC, diesel mechanics):** Reuse approximately 60% of infrastructure; seed with domain-specific content and terminology; calibrate safety patterns for the new domain
- **Third and fourth verticals:** Increasingly, the work is domain seeding rather than infrastructure construction

- **By the fifth vertical:** Mostly assembling from existing building blocks, with a narrowing slice of new work for domain-specific terminology and equipment hierarchies

Each vertical also contributes back to prior verticals. Diesel mechanics knowledge informs generator diagnostics. Industrial hydraulics knowledge informs heavy equipment maintenance. The cross-trade knowledge graph grows denser with each new domain, creating retrieval benefits that compound across the entire platform.

A competitor entering power generation today must build all of this infrastructure. A competitor entering power generation after Rudy has already expanded to three additional trades is competing against a system with four verticals' worth of compounded learning.

Part IV: The Moat

What would a well-funded competitor need to replicate Rudy's position? The honest answer, system by system:

Engineering Complexity

The individual components of Rudy's retrieval system are publicly documented. Vector search, BM25 keyword search, knowledge graph traversal, RRF rank fusion — these are known techniques with available tooling. Any competent ML engineering team can build each of them.

The difficulty is not building any one component. The difficulty is building 11+ interacting systems that work correctly together under production load, calibrated for safety-critical output, with learning loops that improve all components simultaneously from usage.

The safety calibration loop depends on having a safety detection system to calibrate. The safety detection system depends on a per-tenant sensitivity model. The per-tenant sensitivity model depends on interaction history. The interaction history depends on users trusting the safety warnings enough to dismiss or acknowledge them. That trust depends on the safety warnings being accurate. The accuracy depends on the calibration loop.

Circular dependencies of this kind don't have clean "build it and it will work" development paths. They require iterative production operation to close the loops. Each loop closure reveals new edge cases. The edge cases require system-wide adjustments. The adjustments require revalidating the interactions between loops. This is years of production work, not months of engineering.

Accumulated Knowledge

The platform currently holds:

- 3,000+ documents across 127 manufacturers
- 187 equipment models with normalized metadata
- 120,000+ knowledge relationships
- 30+ autonomously generated research documents
- Glossary terms and equipment entity aliases extracted from every document
- Per-tenant safety calibration data

- Citation engagement histories across all queries

None of this was scraped from a public source. The calibration data, engagement histories, implicit feedback signals, and authority scores are proprietary artifacts of platform operation. They cannot be purchased, synthesized from public data, or transferred from another domain. A competitor building the same architecture today would begin with all of these tables empty.

The gap widens daily. Every query generates engagement data. Every correction updates authority scores. Every safety interaction refines calibration. Every self-healing cycle adds research documents. The accumulation is continuous and cannot be paused or replicated retroactively.

Trust Network

The 5-tier authority system represents accumulated trust that took verified contributions to build. Master-tier contributors have demonstrated domain expertise through approved corrections over time. That demonstrated authority is the basis for weighting their future contributions differently from a new account.

A new entrant's platform starts with every user at Novice. The value of the authority network — the signal that separates an expert correction from a guess — only exists after years of operation. A competitor cannot purchase it or bootstrap it artificially without defeating its purpose.

Trade-Specific Calibration

Generic configuration produces a functional system. Field-tested calibration produces a useful one.

The 27-cluster synonym expansion for BM25 queries was built from observed query patterns — terms that field technicians use interchangeably but that don't share lemmas. The 80+ term domain dictionary reflects abbreviations that appear in technician queries and must be expanded for accurate retrieval. The per-tenant safety calibration reflects actual dismissal patterns on actual equipment. The query reformulation strategies for each gap type were refined through production iteration.

This calibration is not documented anywhere. It exists in the system configuration, in the learned parameters of the feedback loops, and in the data accumulated from production operation. It cannot be reconstructed from first principles without operating the system at scale.

Time

Compounding effects mean that the lead grows with time, not proportionally to effort. A competitor starting 12 months from now faces:

- 12 more months of citation engagement data in Rudy's ranking system
- 12 more months of authority scores accumulating across the contributor base
- 12 more months of safety calibration per tenant
- 12 more months of implicit feedback signals informing persona calibration
- 12 more months of self-healing research documents filling knowledge gaps
- 12 more months of cross-trade learning if Rudy has expanded to additional verticals

The gap is not linear. A knowledge base that has been operating for two years is not twice as valuable as one that has been operating for one year — it is materially more valuable, because the second year's learning builds on the first year's foundation, the community's authority scores are more differentiated, and the calibration has had more cycles to converge.

Starting later doesn't mean starting with a smaller lead to close. In compounding systems, starting later means starting further behind and falling further behind each additional month.

Part V: The Vision

The compounding engine described in Part III points toward two near-term directions that are not speculative — they are extensions of mechanisms already operating in production.

Personal Adaptation

The current platform learns about technicians in aggregate: which query patterns indicate expertise, which equipment this user works on, which implicit signals indicate frustration or satisfaction. The next layer is applying this learning persistently across sessions and proactively.

Equipment fleet memory — A technician's equipment context should persist across sessions, not reset at logout. The system already tracks equipment auto-detection within a session; the extension is building a durable fleet profile that surfaces relevant information proactively. When a service bulletin is ingested for equipment that appears in a technician's fleet history, the system can surface it without waiting for a query.

Proactive surfacing — The retrieval loop today is reactive: a technician asks, the system answers. With persistent equipment context and continuous knowledge ingestion, the system can alert technicians to relevant new information — manufacturer TSBs for equipment they service regularly, emerging failure patterns detected across the community, regulatory changes affecting their work.

Response calibration to demonstrated expertise — Persona calibration already infers expertise from query phrasing. The refinement is calibrating not just response detail but response structure: a master technician asking about a complex fault may want a differential diagnosis tree; a junior technician asking the same question may need a step-by-step procedure with safety callouts at each step. The system already has the inference mechanism; the extension is applying it more granularly.

Career Development Engine

The authority system was designed for content quality control. It also happens to be a competency record.

A technician who has been using Rudy for two years has a history of approved corrections, query patterns across equipment types, and demonstrated expertise across domains. That history is a competency profile that neither the technician nor their employer could easily construct otherwise.

Skill gap identification — If a technician regularly queries about procedures they should be performing independently, that is a signal about a training gap. If their query patterns on a specific equipment type are shifting from "how to" to "why does" questions, that is evidence of growing competency. The implicit feedback system already detects behavioral signals; the extension is interpreting those signals as competency indicators.

Growth trajectory — The authority tier system already tracks reputation. Making that trajectory visible to the technician creates a career development dimension that no other tool in their workflow provides.

Employer visibility (opt-in) — With technician consent, competency profiles can be shared with employers for workforce planning and training prioritization. The recruiting pipeline application — identifying technicians with demonstrated expertise in specific equipment types — is a direct extension of the authority system's existing

data.

Trade Expansion

The 11-vertical expansion sequence is ordered by knowledge domain overlap — each trade inherits a substantial foundation from prior verticals, so each launch is faster and cheaper than the last:

#	Trade	Overlap	Why This Order
1	Generator / Power Systems	BASE	Beachhead — Cummins, Cat, MTU, Kohler, Generac
2	Diesel Truck Mechanic	~60%	Same engines (Cummins ISX, Cat C15, Detroit DD15), fuel, cooling, DPF/SCR
3	Marine Diesel	~55%	Same engine blocks with marine ratings. Tiny niche (~30-50K), dominate fast
4	Heavy Equipment	~50%	Cat + JD engines known. Hydraulics is the big new domain
5	Forklifts	~55%	Inherits hydraulics from #4, diesel from base. Small niche, few OEMs
6	Pump Technicians	~50%	Inherits hydraulics + industrial electrical. Bridges into plant operations
7	Crane & Rigging	~45%	Inherits hydraulics from #4. Highly specialized, high willingness to pay
8	Refrigeration / Cold Chain	~25%	Seeds the refrigeration domain before full HVAC. Commercial focus
9	Industrial Electrical	~40%	Electrical foundation from generators. Big market, enter after playbook proven
10	Elevator & Escalator	~30%	Builds on industrial electrical. Heavily regulated = sticky
11	HVAC & Refrigeration	~20%	Largest market (~400K). Refrigeration domain from #8. Pipeline battle-tested

Each vertical is not a separate product — it is a tenant configuration on the same platform, seeded with domain-specific content and safety calibration. The retrieval infrastructure, learning loops, quality evaluation, authority system, and multi-tenant architecture transfer without modification.

The compounding math: vertical #2 needs ~40% new content (60% reuse from generators). By vertical #5-6, the platform is mostly assembling from existing knowledge domains. By vertical #9-11, the document collection pipeline, self-healing knowledge loop, and user contributions are doing most of the work. The marginal cost of each new vertical declines as the platform matures.

Part VI: Future Ancillaries

The capabilities described in this section are natural extensions of the existing architecture, not separate product lines. They are presented as such because they require either user volume (to activate remaining learning loops) or hardware integration (to ingest physical-world data) that the current platform does not yet have.

Visual Diagnostics

Photo upload with Gemini multimodal analysis is already deployed in production. The extension is making this bidirectional: not just technicians sending photos to get answers, but an accumulating annotated image library that becomes a searchable visual reference.

Every photograph of a fault display, corroded connection, or failure mode — tagged with equipment, fault condition, and resolution outcome — builds a visual knowledge base that cannot be constructed from manufacturer documentation. Manufacturers document what components look like new. The visual knowledge base documents what they look like at every stage of failure.

Video-based procedure verification extends this further: a technician recording a maintenance procedure can have each step verified against the SOP, with missed steps and safety violations flagged in real time. Correct technique captured on video becomes training material directly from the field.

Acoustic Signature Analysis

Experienced technicians diagnose by ear. A bearing that is about to seize sounds different before it fails. Cavitation in a pump has a characteristic high-frequency signature. Misalignment creates specific harmonic patterns at rotation frequency and multiples. This diagnostic knowledge exists in experienced technicians' ears and nowhere else.

The approach is empirical: establish per-asset acoustic baselines during normal operation, detect deviations using spectral analysis, and validate failure-mode mappings through outcome tracking when deviations are subsequently confirmed as specific failure modes. Community validation means that each confirmed acoustic-to-failure mapping improves detection for every user on similar equipment.

IoT and Sensor Integration

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 that go largely unanalyzed. MQTT, OPC-UA, Modbus TCP/RTU, BACnet, and REST-based APIs cover the major industrial communication protocols.

The value is context fusion: a temperature reading is not just a number — it is a reading from a specific sensor on a specific component of a specific unit with a specific maintenance history and OEM-defined operating limits. Correlating that reading with the maintenance history, the community's pattern database, and the equipment's known failure modes transforms an isolated data point into an actionable signal.

Predictive Intelligence

The convergence of sensor telemetry, maintenance history, and community failure patterns creates the conditions for shifting from reactive support to predictive alerting. Individual sensor readings don't predict failures. Combinations of weak signals — temperature trending up 2°F per week, vibration shift on a specific bearing, query patterns from technicians who recently worked on similar units — collectively indicate an impending failure that no single signal would surface.

Condition-based maintenance scheduling is the direct economic case: replace time-based maintenance intervals with data-driven determinations of when maintenance is actually needed. The economic asymmetry is sharp — unnecessary maintenance is expensive; missed maintenance is catastrophic. Data closes that gap.

Frontier capabilities (longer-term): edge-deployed AI for sites without connectivity; natural language equipment control with human-in-the-loop confirmation and complete audit logging.

Part VII: Domain Expansion

The architecture's domain-agnosticism is not a marketing claim — it is a consequence of how the platform was built. The retrieval pipeline does not assume knowledge about any specific trade. The safety framework is a configurable pattern-matching system that accepts any hazard taxonomy. The authority system is a points-based tier mechanism with no domain-specific logic. The learning loops track signals that exist in any field service context: query satisfaction, citation engagement, correction authority, safety calibration.

Expanding to a new domain requires: domain knowledge seeding (equipment hierarchies, terminology glossaries, initial document corpus), safety pattern configuration for the domain's specific hazard categories, and authority hierarchy calibration appropriate to the trade's expertise structure. The retrieval infrastructure, feedback systems, quality evaluation, and admin tooling transfer without modification.

Domain	Knowledge Crisis	Multimodal Value	Predictive Value
HVAC	High tribal knowledge, fragmented manuals, same retiring-workforce dynamic	Acoustic: compressor health. Visual: refrigerant leaks, coil condition	Refrigerant pressure trends, energy consumption anomalies
Medical Equipment	Very high regulatory density, moderate tribal knowledge	Visual: error displays, wear indicators, calibration drift indicators	Sensor drift detection, calibration degradation prediction
Manufacturing	High tribal knowledge, many OEMs with proprietary diagnostics	Acoustic: motor and pump health. Visual: wear patterns, alignment	Vibration trends, production quality correlation with equipment state
Oil and Gas	High tribal knowledge, high regulatory, serious safety stakes	Acoustic: valve and pump cavitation. Visual: corrosion, joint integrity	Pressure and flow trends, wellhead performance degradation
Aviation MRO	Very high regulatory density, very high tribal knowledge	Visual: structural inspection, fastener condition. Acoustic: engine health	Flight data trends, component lifecycle analysis
Marine	Very high tribal knowledge, harsh environment failure modes	Acoustic: engine and hull. Visual: corrosion, bearing wear	Engine performance curves, hull fouling rate prediction

The key message for each vertical is the same as for power generation: the knowledge that keeps equipment running safely is fragmented, at risk of retirement, and impossible to capture with documentation alone. The architecture that solves this in power generation solves it structurally in every other trade. The cross-trade compounding from Part III means each additional vertical benefits from everything learned in all prior verticals.

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 operational experience in power generation service, Nick has seen firsthand what the knowledge crisis costs: experienced technicians taking decades of diagnostic intuition

into retirement, juniors learning through expensive callbacks, and customers bearing the cost of fragmented expertise.

Rudy began as an internal tool to make OnPoint's own technicians more effective. It became clear that the architecture being built was not solving a company-specific problem but an industry-wide one — and that the multi-tenant platform that emerged could serve any field service organization facing the same structural challenge. The same platform that helps an OnPoint technician in the field is equally available to any competitor willing to build their technicians' knowledge base on it, because the network effects that compound from a larger user base benefit everyone using the system.

The expansion thesis — from power generation to every trade where complex equipment requires expert diagnosis — follows directly from the architecture. The platform doesn't know it's about generators. It knows about knowledge retrieval, safety-critical information, expert authority, and continuous learning. Those properties transfer.

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. Craswell, N., et al. (2020). Overview of the TREC 2019 Deep Learning Track. *TREC 2019*.
6. Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. *arXiv preprint*.
7. Mobley, R. K. (2002). An Introduction to Predictive Maintenance. *Butterworth-Heinemann*.
8. IEA World Energy Employment Report. International Energy Agency.
9. Goldman Sachs Global Investment Research. Power Generation Workforce Outlook.
10. U.S. Department of Labor, Bureau of Labor Statistics. Power Plant Operators, Distributors, and Dispatchers — Occupational Outlook Handbook.

OnPoint Generators, Inc. May 2026