PlantPulse® **AURA Models**

TECHNICAL WHITEPAPER

TECHNICAL WHITEPAPER · REV. 2026.09

Industrial AI models that answer in the language of the plant.

Trained directly by KOPENS, three industrial AI models — diagnosis · search · precise alignment. A mixture-of-experts architecture fits on **one GPU**, solves calculations with formulas, and says "Not measured" for what was not measured. Scoring is done not by judge models but by **formulas · schemas · operational records**. This edition describes the **architecture and validation methods** — performance figures will appear in a revised edition after remeasurement.

3	65,954	0	1
models — 35.9B · 4B · 0.6B	training samples (latest build)	judge models — machine scoring	GPU for serving

September 2026

AURA · on-premises / air-gapped networks · for technical reviewers

kopens®

00 · Summary

Summary — Where General-Purpose Models Become Dangerous in Factories

General-purpose models know much about the world, but **do not know your factory's compressor.** They explain what a bearing is, and **invent vibration values that were never measured.** In a factory the latter is dangerous — if missing data masquerades as "normal," operators believe uninspected equipment is sound. **AURA** is a model family trained to do the opposite, and **KOPENS handles data design, training, evaluation, quantization and serving itself** — this product began when a company with 20 years of industrial data became **able to train models on it.**

3
Language-track models — diagnostics(35.9B MoE) · retrieval(4B) · alignment(0.6B)

5
Benchmarks — all five machine-scored **without judge models**

158
Field diagnostic questions — ground truth from **outside us**(operating records)

1
GPU serving the language model(quantized 21 GB)

Current status — separating "running" from "in development"

Stage	What	Meaning
In production	Language model 35.9B · retrieval(4B) · precision alignment(0.6B)	The in-house inference server **serves the trained model in production** — not the base model, the version we trained
Next stage	Time-series track	A **separate family** for waveforms — about a million parameters, runs **without a GPU.** Product application: next edition(Ch. 15)
Next stage	Vision	**Evaluation criteria built first** — a weld-defect benchmark and industrial images. Next edition

Contents

01 Strengths — structure, before metrics 3
02 Architecture — three language models, one waveform family 4
03 Training and data — where ground truth comes from 5
04 Recipe — mechanisms · fine-tuning · quantization 6
05 What it examines — fault mechanism map 7
06 Telling faults apart — a different sequence, a different fault 8
07 Selecting data — does ground truth come from outside? 9
08 Acquired data — what it covers 10
09 Simulator — physics first, labels derived 11
10 Saying "I don't know" — abstaining is also correct 12
11 Sources of ground truth — calculations · standards · tools 13
12 Training mix — what to include and how much 14
13 Identity and scope — by training, not prompts 15
14 Retrieval — the ceiling on answer quality 16
15 Time-series track — small models for waveforms 17
16 Evaluation — measuring without judge models 18
17 Field diagnostics — false alarms counted separately 19
18 On-premises deployment · next edition · contact 20

Why this edition has no performance metrics. The trained model **has just entered the product path,** and metrics will be published **after they are measured again on that path** — numbers released early keep circulating as evidence even after correction. So this edition fixes **only the architecture and the validation method.** What to measure and how are already defined, so reviewers **can measure it themselves, the same way**(Ch. 16). **Audience** — technical reviewers · data scientists · equipment engineers. Not sales material.

01 · Strengths

What the design delivers — the difference made by "structure" before metrics

The six points below are not benchmark scores but **properties that follow directly from the design**. They therefore remain stable across upgrades, and hold even without performance figures in this version.

DEPLOYMENT

It fits on one GPU

Of the full 35.9B, **only 3B** parameters activate per token in the mixture-of-experts architecture, and the quantized checkpoint is **21 GB**. It requires **one-ninth the compute** of a dense configuration targeting the same quality, so it runs on one GPU. **A model requiring eight GPUs cannot enter most factories regardless of performance.**

OPERATIONS

Upgrades do not require rebuilding the index

We **fixed the retrieval model's embedding dimension (2,560) across versions**. The vector index remains usable after replacing the model — an upgrade that requires re-embedding tens of millions of chunks is **eventually postponed**, and a postponed upgrade is effectively no upgrade.

TRUST

Distinguishes 0 from "not measured"

"0 work orders" is a measurement; "could not read because the connection failed" is not. Reporting the latter as the former **causes equipment that nobody inspected to be trusted as normal.** This distinction is enforced in both training data and evaluation items.

VERIFICATION

Does not use a judge model for scoring

Because answers come from **formulas · schemas · operational records**, scoring is mechanical. The same input produces **the same value regardless of who runs it or when**, and customers can repeat the measurement in their own environments using the same method. When one model scores another, the biases of both are included.

TRAINING

Does not learn fabricated answers

Ground truth comes from **simulator rules · formulas · tool schemas**. Tool trajectories generated by strong models are used **only after passing schema validation** (pass rate 83%) — had the failed 17% been learned, the model would have confidently supplied **arguments that do not exist**.

FIELD

Does not cry wolf

Field diagnostic evaluations count **false alarms separately from missed detections**. Combining both into one "accuracy" score gives **a model that raises alarms for every interval** a high score. Such an assistant is turned off within a week, and a disabled assistant detects nothing.

PROPRIETARY TECHNOLOGY

A company that "builds" models, not one that merely "uses" them

Data design · training · evaluation instrumentation · quantization · serving all run **within one company**. When the model is wrong in the field, we fix it not by covering it with prompts but **through data and training** — the remeasurement plan in Chapter 17 is an example.

INTEGRATED AS ONE SYSTEM

Built by the same team as the platform

The schemas used for tool-call training and evaluation are **the actual tool definitions of the operational PlantPulse platform** (57 types). This integration is difficult to achieve when the model and platform are developed separately — instead of **explaining our tools to someone else's model**, we train our model with our tools.

Narrowing the scope is also a strength. It answers questions about equipment · processes · quality · safety, while politics · diplomacy remain out of scope. Korean is the default, and it also answers in English. We spent 5% of training on **deciding what it will "not" answer** — because an assistant that gives plausible answers outside its scope also gives merely plausible answers within its scope.

02 · Architecture

Three Models in the Language Track — Retrieve Broadly, Select Precisely, Answer

Instead of having one large model do everything, **we divided the roles into three.** Retrieval must be inexpensive and broad, ranking must be accurate, and answering must know how to invoke tools. Forcing all three into one model makes one bear the cost of the others.

Model	Size	Format
AURA-35B	**35.9B** **3B** active per token	Mixture of experts (MoE). Diagnosis · calculation · tool calls · reports. Serving context: **32,768** tokens (training sequence: 2,048)
AURA-Embedding-4B	4.0B	**2,560-dimensional** output, up to 1,024 input tokens. Last-token pooling and **asymmetric encoding** that adds instructions only to queries (not to documents)
AURA-Reranker-0.6B	0.6B	A cross-encoder that **reads the query and document together**. Up to 1,024 input tokens, applied only to the top 20~100 candidates

How the Three Work Together



Embedding dimensions do not change between releases. Even when replacing the retrieval model with a new release, **the vector index does not need to be rebuilt.** Re-embedding tens of millions of chunks is a task that becomes daunting after the first attempt, so the dimensions are fixed as a contract.

Why Mixture of Experts (MoE) — Whether It Can "Fit" On Site

Of the full 35.9B, **only 3B** is active per token. Compared with a dense configuration targeting the same quality, **compute is 1/9**, so it runs on **one GPU**. A model requiring eight GPUs cannot enter most factories regardless of performance — this choice determines whether deployment is feasible. For the same reason, **we chose a mixture of experts** — a dense configuration requires 9 times the compute.

How the Retrieval Models Were Trained

Embedding	Contrastive learning (InfoNCE) — other items in the same batch serve as negatives, with separate **hard negatives designed to confuse**. This builds the ability to distinguish documents that appear relevant but do not contain the answer
Reranker	Listwise cross-entropy that selects the answer from a candidate list — it learns **the ranking task itself**

Waveforms Are Treated as a "Different Object" — the Time-Series Track

Language models are strong at **explaining why**, but identifying **what is abnormal** in waveforms arriving every second is a different kind of task. We therefore treat waveforms as **a separate million-parameter model family** — **small enough to run without a GPU**, deployable at the edge, and inexpensive to tune separately for each site. Product integration **will be addressed in the next release** — this product validates value before integration.

One result is already in use — "there are intervals where not invoking a model is better." In some cases, statistics measuring **deviation from the equipment's normal baseline** separated anomalous intervals without invoking a model once. We therefore **do not solve those intervals with a model.** Distinguishing where a model is needed from where statistics suffice comes before "making the large model larger."

03 · Training and Data

One Stage, an Adapter — and Where the Answers Come From

Training consists of **one stage of supervised fine-tuning (SFT)**. No pretraining was performed, and preference training (DPO·RLHF) was **not performed**. The weights are added through an adapter (LoRA), leaving the base weights untouched. Because industrial Q&A does not exist on the internet, it must be **created**, and the key question is **who guarantees the answers in the created Q&A** — learning answers generated by asking another model also teaches that model's hallucinations.

Latest Build Scale

Latest build	**65,954** items — rebuilt for each release
Previous release	49,401 items · 2 epochs · about **75 million** tokens
Sequence length	2,048 (serving context is 32,768)
Language model training time	**1 data-center-grade GPU · 6 hours** (previous release)
Retrieval model training time	Embedding 30 minutes · reranker 1 hour

Composition

Identity · guardrails	5%
Tool calls	5%
Diagnosis · calculation · sensors · standards · procedures	Remaining 90%

Guardrails account for 5% because **"what not to answer" is also a training target** — an assistant that gives plausible answers to out-of-scope questions also gives plausible answers within scope.

Five Data Streams

Stream	Where the answers come from
Physics simulator synthesis	Sensor time series for **18 equipment types × 9 failure modes**. Answers come from **simulator rules**
Formula generation	Industrial calculation questions. Answers are **determined by formulas**, so no judge is needed for grading
Public source texts	Provisions from standards such as the Occupational Safety and Health Act · OSHA, and field terminology
Procedure documents	**Evidence-based** Q&A created from SOPs — answers use only document content
Tool schemas	Tool-call examples generated from OpenAPI · MCP definitions of the live platform

Latest build composition — event chains 40,000 · calculations 8,546 · sensors 5,009, plus tool calls · MCP · platform · standards · terminology. **Calculation capability once collapsed in a release with 0 calculation items**, so we check item counts by stream before every build.

> **Distilled data is used only after "validation."** Some tool trajectories were created as multi-turn examples using a stronger model. However, all generated results are **machine-validated against schemas**, and only those that pass are used for training — **pass rate 83%**. The remaining 17% were plausible calls rejected by the schemas; had the model learned them, it would have confidently supplied **nonexistent arguments**.

Deriving Answers from "Rules"

① SIMULATOR

Failures Are "Created"

Sensor time series are generated by applying 9 failure modes to 18 equipment types. Because the simulator rules define **the sequence in which** vibration·temperature·current change as a bearing wears, the answer to "what is happening in this waveform" is determined.

② FORMULAS

Calculations Are Solved to Generate Questions

Bearing defect frequency · life · NPSH · shell thickness · flow rate · OEE · process capability · heat input. Questions are generated with random values, and **formulas calculate the answers**. The same formulas are used for grading.

③ STANDARDS · PROCEDURES

Use Only What Exists

Statutory provisions and SOPs are used to create **evidence-based** Q&A. The key is enforcing "what is not in the document must not appear in the answer" from the data stage onward.

④ TOOLS

The Schema Is the Judge

Call examples are created from actual tool definitions in the operating platform. Multi-turn trajectories created with a stronger model are retained **only if they pass schema validation** (pass rate 83%).

04 · Recipe

Teach "failure mechanisms," not equipment

Even for the same bearing wear, collecting data separately by manufacturer and model causes **the combinations to grow into the hundreds of thousands.** So we changed the unit of instruction — **component · failure mechanism · operating condition**. Because mechanisms are equipment-independent, **they transfer directly to unseen equipment.**

WHY WE CHANGED THE UNIT

Equipment-centric → mechanism-centric

Equipment-centric collection requires data for every **manufacturer × model × asset**. Mechanism-centric data is reduced to **component + mechanism + operating condition**, so the same amount of data covers **a much broader range of equipment**.

BUILD ONCE, USE IN FOUR PLACES

Reuse the same failure episode

We use **one failure episode** created by a physics- and rule-based generator **across** training · tool training · evaluation · benchmarking. Because all four see the same world, **training and scoring remain aligned.**

Teach the "tracing process," not the "answer"

Answering "Why is the bearing temperature high?" is not the goal. The goal is to **follow this process independently** —

```
alarm → identify the asset → find the tags → read the trend → read the history → list the possible mechanisms
    → weigh the evidence per candidate → ask for more data if it is thin → infer the cause → propose an action → verify recovery
```

Thus, the training data is not **a "symptom → answer" pair**, but **a trajectory that gathers evidence and narrows the possibilities**. Tool calls are included in the data for the same reason — narrowing the cause in the field requires **retrieving more data**.

Fine-tuning recipe

Stage	**One stage of supervised fine-tuning (SFT)** — no pretraining
Weights	Apply as an **adapter (LoRA)** — the base weights remain unchanged
Sequence · epoch	2,048 · 2
Training cost	**1 data center-grade GPU · 6 hours**
Retrieval models	Embedding 30 minutes · reranker 1 hour

Latest build	**65,954** records — rebuilt for every release
Event chains	40,000 — trajectories that narrow the evidence
Industrial calculations	8,546 — answers derived from formulas
Sensors	5,009
Tools · standards · terminology	Tool calls · MCP · platforms · standards provisions · field terminology

For every build, we first check the count in each category. If one category is empty, that capability quietly weakens — calculations became unstable in a release missing calculation data, so **count verification is the first step in the build process**.

Quantization — fit on one GPU without "instability"

① CALIBRATION

Calibrate with domain data

We use calibration samples **drawn from the training set** when reducing to 4-bit. Calibration with general documents from a different distribution **misplaces the quantization points** and degrades quality.

② WHAT REMAINS

Do not quantize the router and output layer

In a mixture of experts, **the logits that determine which expert receives the input** are sensitive. We therefore **exclude the router and output layer from quantization** — quantizing the rest produces 21 GB.

③ DEPLOYMENT FORMAT

Place the adapter over the base

Keep the 21 GB base unchanged and swap **only the adapter (hundreds of MB)**. Because **requantization is unnecessary**, deployment and rollback are simple, with no risk of quality instability from requantization.

④ VERIFICATION

Check that no experts are missing

At 4-bit, the artifact falls to **one-third or less** of the original. If its size remains similar, the expert weights were not quantized, so **size alone provides an initial check**.

05 · What to Observe

Failure Mechanism Map — A Single Signal Does Not Differentiate Them

A temperature increase alone leaves **four** candidates. Therefore, the model learns not pairs such as "temperature rise → bearing," but **which signals move together and which moves first**. The map is shown below.

Rotating Machinery — Eight

Failure	Temperature	Vibration	Noise	Power	Auxiliary Fluid	Indicator → Failure
Bearing wear	+28°C	**4.2×**	+14 dB	+8%	—	125 d
Misalignment	+12°C	3.0×	+8 dB	+15%	—	200 d
Imbalance	+4°C	3.4×	+6 dB	+3%	—	250 d
Insufficient lubrication	+34°C	2.2×	+11 dB	+10%	Lubrication −62%	**83 d**(fastest)
Cooling failure	**+42°C**	1.1×	+3 dB	+6%	Cooling −58%	100 d
Electrical fault	+30°C	1.6×	+5 dB	**+32%**	—	143 d
Cavitation	+6°C	3.6×	**+18 dB**	**−12%**	—	111 d
Fouling	+36°C	1.1×	+2 dB	+22%	Cooling −30%	167 d

Fluid Systems — Three

Failure	Discharge Pressure	Downstream Pressure	Differential Pressure	Duty Cycle	Temperature	Indicator → Failure
Leakage	−18%	**−34%**	—	**+0.45**	+16°C	91 d
Blockage	**+16%**	−29%	**3.6×**	+0.12	+24°C	167 d
Valve sticking	−5%	−7%	1.5×	—	+10°C	125 d

11 Failure Types × 18 Equipment Types

Centrifugal pumps · reciprocating/screw compressors · blowers · chillers · cooling tower fans · gearboxes · induction motors · hydraulic units · injection molding machines · presses · CNC spindles · conveyor drives · agitators · centrifuges · heat exchanger circulation pumps · boiler feedwater pumps · vacuum pumps. **Applicable failures differ by equipment** — the six rotating failures apply to all 18 types, the three fluid failures to 11 types, and cavitation and fouling to 4 types each.

"Valve Sticking" Is Not Detected by the Mean

The pressure mean is **8.29 bar**(normal 8.49), nearly the same, but the **standard deviation is 0.309**(normal 0.104). Because the symptom is not the value but **the fluctuation itself**, monitoring only the mean misses it — this is **why dispersion is monitored as well**.

06 · How to distinguish

Even with the same two signals, a different "order" indicates a different fault

The ability to narrow down the cause in the field comes not from the magnitude of values but from **what moved first**. Therefore, the training data is not a snapshot but a **trajectory with a time axis**.

Fault	Leading signal	Following signal
Bearing wear	**Vibration** rises	Temperature follows · Defect-frequency harmonics
Insufficient lubrication	**Lubricant flow** drops	Temperature surges from frictional heat
Cooling failure	**Coolant flow** drops	Only temperature rises (vibration remains normal)
Misalignment	2× component dominant · **Axial** vibration	Power increases
Unbalance	1× component dominant · **Radial**	Temperature change is minimal
Electrical fault	**Current imbalance**	Power consumption · Temperature rise together
Cavitation	**Broadband noise** surges	Discharge drops → **Power decreases**
Fouling	Heat-transfer surface contamination → **Differential pressure** rises	Accompanied by temperature rise

When temperature rises — rules for distinguishing four modes

If vibration rises "together"

It indicates bearing wear · misalignment · unbalance. Next, distinguish by whether it is 1× or 2×, radial or axial.

If only temperature rises

It indicates cooling failure · fouling. Check whether coolant flow dropped first or differential pressure rose.

If lubricant flow dropped "first"

It is insufficient lubrication. The temperature surge is a result, not the cause.

The only mode in which power "decreases"

It is **cavitation**. Power rises in every other mode — this alone confirms it.

How to distinguish the three fluid faults

Leakage	Discharge pressure and downstream pressure drop **together**, but downstream pressure drops more, widening the gap. And **utilization stays near 1.0** — it cannot stop because it must replace the leakage. Differential pressure remains unchanged
Blockage	Discharge pressure instead **rises** (it accumulates upstream). **A 3.6-fold differential pressure is the primary signal** — it resembles fouling, but the temperature rise is smaller and the differential pressure rises more steeply
Stuck valve	The mean remains within the normal range. **Dispersion** is the signal

The evidence is expressed in sentences. The training-data answer does not end with a one-line determination; it also provides sentences supporting that determination — "Temperature rose first at D+11, and vibration followed at D+19 (**8-day lag**) — thermal cause preceded", "Lubricant flow fell from 78.3% → 49.8% at **0.86%/day** — continuous consumption or leakage", "Vibration trend over the last 14 days: **+0.309 mm/s·day** — accelerating". Values and periods must be included so that **people can verify the result.**

07 · How to Select Data

First ask, "Does the ground truth come from outside us?"

Before collecting more data, establish **selection criteria**. If we create problems using our own rules and score them by those same rules, it is **circular** — the most valuable data breaks that loop. We therefore ask three questions.

CRITERION ①

Does the ground truth come from outside us?

Does it carry **labels we did not create**, such as operational records and maintenance history? Only such data lets us ask, "Were our assumptions correct?"

CRITERION ②

Is it a laboratory failure or a field failure?

Artificial defects and accelerated tests differ from **actual failures during operation**. Field data includes missing values, maintenance events, and changing operating conditions — those too are reality.

CRITERION ③

Does the form involve fewer assumptions?

Raw high-frequency vibration waveforms must be converted into features, and **that conversion embeds our assumptions.** Signals such as pressure, temperature, and current are **already in engineering units** and involve fewer such assumptions.

EXCLUSION CRITERION

Is there "causality" between the sensor and label?

Some data contains physically impossible values, or has a "failure imminent" label despite normal sensor readings. **We do not use data without causality, regardless of its size** — training on it instills a habit of making unsupported conclusions.

We do not accept public-dataset labels at face value

We download, inspect, and **verify them ourselves.** In one case, **72** segments classified as "normal" were statistically indistinguishable from failure segments and were excluded from evaluation. We also found that, among 33 normal-difficulty items, **11 were windows whose values remained frozen for two full hours** yet were labeled "normal." **Interpreting a stuck sensor as equipment failure, or vice versa,** is a major source of field errors, so we handle such windows separately in evaluation.

Train with synthetic data; verify with measured data. Evaluating a simulator-generated model using simulator-generated data applies the same assumptions twice. We therefore use synthetic data **for training**, while **evaluation of "whether it was correct" uses data whose ground truth comes from outside**.

08 · Acquired Data

Ground Truth from Outside — What Does It Cover?

These are public datasets that passed the criteria above and were **actually downloaded and opened**. Each was selected based on **which coverage gap it fills** — the criterion is not size but **the area covered**.

Data	Type	What It Covers
Subway Air Compressor MetroPT-3 · CC BY 4.0	**Measured**	Air compressor on an operating train, **6 months · 1.51 million rows**(10-second intervals). 7 analog + 8 digital. **4 failures are documented in operational records** — a representative case where ground truth comes from outside
Hydraulic Test Rig UCI · ZeMA	**Measured**	**2,205** 60-second load cycles, 17 sensors. Cooler · valve · internal pump leakage · accumulator were **degraded in stages to establish labels**. It **aligns almost directly with our fluid systems**
Chemical Process Tennessee Eastman	Simulation	**21 process faults · 52 variables**. Fills the entire process-system coverage gap
Circulation Pump Test Rig SKAB	**Measured**	Records of actual valve operation · leakage. It is valuable because it **measures vibration and pressure together**
Semiconductor Process SECOM	Measured	Pass/fail from **590 sensors** — high-dimensional process monitoring
Naval Propulsion System Naval Propulsion	Simulation	**11,934** gas-turbine propulsion-system instances. The degradation coefficient is **a continuous-value label in 0.001 increments** — a form absent from the normal/abnormal binary classification
Power-Generation Gas Turbine UCI · 5 years	**Measured**	11 sensors + CO · NOx emissions. **The dimension of reading combustion state from exhaust gases** was absent from temperature · vibration data
Sensor Drift	Measured	**Cases where the instrument, not the equipment, fails** — a source of frequent field misdiagnosis
Steel-Plate Surface Defects	Measured	**7 defect types** — material systems

By Industry

Rail · pneumatics · hydraulics · chemical processes · pump systems · semiconductors · shipbuilding/defense propulsion · power generation(gas turbines) · instrumentation · metals. **Each industry differs in "what speaks first"**, and the goal is to secure each dimension one by one.

Next to Add

Wind power(gearbox · generator SCADA + failure history) · gearbox multiclass · bearing **full-life trajectory**. The last is used to **calibrate the degradation rate with measurements** — fix the data before fixing the model.

09 · Simulator

Define the physics first, then "derive" the labels

Industrial Q&A is not available on the internet, so it must be created. The key is the **order of creation** — assigning labels first and adding signals leaves **no causality between sensors and ground truth.** So we reverse the order.

① **Simulate degradation**	Advance the latent state `d` (0~1) over time. Each failure mode has a different rate, and **accelerates as it progresses**
② **Sensors are observations**	Sensor values are **noisy observations** of that state. Which sensor moves first follows the **actual failure physics**
③ **Derive labels**	**Derive** the ground truth from `d`. This makes sensor → ground-truth inference **possible in principle**
④ **The generation formula guarantees bounds**	Values are not clipped afterward — the generation formula itself produces values only within **physically possible bounds**
⑤ **Stop at failure**	Stop the trajectory at `d=1.0`. **When equipment fails, it is serviced rather than kept running**

Verify that causality actually exists — measured over 20,000 rows

Mode	Temperature	Vibration	Power	Auxiliary flow	Interpretation
Bearing wear	1.33×	**3.40×**	1.05×	—	Vibration dominates
Insufficient lubrication	1.47×	2.06×	1.13×	**0.39×**	Flow drops first
Cooling failure	**1.53×**	1.13×	1.11×	0.52×	Only temperature changes, without vibration
Cavitation	1.29×	2.67×	**0.63×**	—	Power "decreases"
Fouling	1.36×	0.91×	**1.51×**	0.85×	Power rises without vibration

Vibration alone cannot distinguish bearing wear from cooling failure. Temperature · power · flow must be considered together — we proceed only after **numerically verifying** that the distinction holds in the generated data. There are **0** physical-bound violations.

Physics determines "correctness"; the model determines "expression"

Simulator-generated sentences share the same structure, so without intervention the model **memorizes the template.** We therefore vary only the expression — **87** question forms · **50** verdict forms. **Physics determines what is true**; the model determines only how to say it.

No dependence on other models

215,021 generated items(93%) are created **without another model** — they come from physics · formulas · original specifications · tool schemas. Therefore, **the performance ceiling is not limited by another model.**

10 · Saying "I Don't Know"

Without evidence, "Defer Judgment" is the correct answer

The most dangerous failure in industrial AI is not being wrong, but **making assertions without evidence**. Therefore, from the training stage, **"I don't know yet" is taught as the correct answer.**

The simulator knows the answer but does not reveal it

The generator knows the failure mode of the trajectory. However, **if the sensors have not moved yet, it does not assign a label.** Early degradation cannot be diagnosed from sensor data, and assigning an answer anyway teaches the **habit of "being certain without evidence."**

The decision criteria are fixed numerically

A definitive diagnosis is issued only with **at least two pieces of evidence**, or **one piece of evidence + degradation of at least 0.45**. Below that is **defer judgment**, which is also a correct answer.

Therefore, a 3rd of the data answers "do not confirm"

Sensor snapshot	Breakdown
Definitive diagnosis	**895**
Defer judgment	**662** — "The current observations do not yet distinguish them"
Normal	500

In event chains, **definitive diagnoses account for 64.4%**, while the rest are deferred or normal. Deferral phrases were also diversified into **30 variants**, preventing the model from memorizing "I don't know" as **a single stock phrase**.

> **"0" and "not measured" present the same issue.** "0 work orders this week" is a measurement, while "the connection was lost, so it could not be read" is not. If the latter is reported as the former, **operators will believe that equipment no one inspected is normal.** Contexts with missing values were deliberately included in the training data, and in those cases, **answers stating "not measured"** were set as correct.

Answers are provided in a form that people can verify

Whether definitive or deferred, each answer states **which signal · from what value to what value · over how many days** it changed. "Check the temperature" is not evidence; **"The temperature rose from 62→89°C" is evidence** — the evaluation also mechanically counts **whether a numeric value follows the signal name** (Chapter 16).

11 · Sources of Correct Answers

Calculations · Standards · Procedures · Tools — Where Each Correct Answer Comes From

Each data category has **a different authority that guarantees the correct answer.** Only data with a clear authority is used for training.

Category	Scale	Who guarantees the correct answer
Industrial calculations	**8,546**	**Formulas**. Questions are generated with random numbers, and formulas produce the answers — no other model is needed, and **there is no upper limit**
Standards · regulations	**1,326**	**Original provisions**. Answers are derived **only from the text** of 184 provisions of the Occupational Safety and Health Act and 479 provisions of overseas occupational safety regulations
Procedures(SOP)	**1,087**	**Predefined procedures**. We create 106 SOP documents, **fix them first**, and derive answers only from them
Tool calls	**1,059**	**Schema**. Required arguments are declared, guaranteeing the correct answer
Field terminology	140	**Manually curated dictionaries** in 28 categories — English-Korean pairs and spelling variants (트립 / Trip / tripping) are included together

Calculations — 30 families, seven excluded from training

Heat·fluids(LMTD · heat transfer · Reynolds number · pipe friction · pump shaft power · orifice · NPSH · affinity laws), rotating machinery(bearing fault frequency · life · belt speed · reducer torque · shaft torsion), welding·mechanical(heat input · bolt torque · tolerance · fillet welding · key shear), pressure vessels·piping(pipe thickness · cylindrical shell · head · maximum allowable working pressure · safety valve area), quality·electrical(process capability · OEE · motor current · thermal expansion).

The holdout is "within the same family"

The **circumferential** equation for a cylindrical shell is taught, while the **longitudinal** equation is excluded. This single gap distinguishes memorizing a formula from **selecting and setting up the right equation for the situation**.

The question "values" are also realistic

A question is itself unrealistic if it yields values such as a bolt load exceeding yield or a bearing life of 256 hours. Allowable stress ratios and load ratios are generated after **calibration to field ranges**.

Tool calls are taught in four forms

Tool selection	**Which tool to call** among several tools
Required argument completion	The most frequent field errors are **missing required arguments and type errors**
Failure handling	**Do not fabricate a value** when the tool returns **reading rejected**
Tool description	Explain in plain language what the tool does

Instrument reading is a representative case. If repeated readings from a blurry image differ, the tool returns **"unreadable"**. The model is trained **not to guess and report a number** in that case — fabricated instrument values in the field lead directly to **incorrect action**.

12 · Training Mix

What You Include and How Much Defines the Shape of Capability

Having a large dataset is not the same as **mixing it well**. If one category dominates, the other capabilities are **quietly diluted** — so each category is mixed with a **cap**.

Mix of the Latest Build

Event chains (diagnostic traces)	**57%**
Narrative domains	22%
Industrial calculations	12%
Sensor snapshots	7%
Standards · terms · procedures	4%
Identity · guardrails	**2%**

For event chains, **20 sets of ten thousand are generated, but only 4 sets are used**. Without a cap, that category accounts for 87% of the total and **overwhelms everything else.**

What Actually Happened

When we heavily weighted narrative data, calculation performance became unstable. Where the bearing fault frequency should have been calculated, **the model fell back to the statement "check the harmonics."** This was because the training data contained no calculation process. **We would not have found this without the calculation bench**, so the first assembly step is now **checking counts by category**.

Quality Filters Run "Immediately After Generation"

Answers that are too short	Discard
Answers with insufficient Korean	Discard
Other-language leakage	Discard — Chinese-character annotations in Korean (온도(溫度)) are not flagged as false positives

Filtering **immediately after generation is cheaper** than filtering just before training, and the reason for each removal is clearer then.

Prompts — Long Ones Suppress Tool Calls

Using the same model · same questions, we measured tool calls while **changing only the system prompt**. The results ranked as follows —

Prompt	Result
2 lines (145 characters)	**Best**
Condensed version (1,189 characters)	Slight difference
2 lines + list of 57 tool names	Clearly worse
Long body (5,488 characters)	Lowest

The improvement did not come from stating "fewer" rules — the long body buries the tool-call signal. Listing tool "names" in the body is also counterproductive (the model knows them from the schema alone). We therefore have one operating rule — **measure with the bench before adding another rule.** If adding it lowers the score, that rule **costs more than it provides.**

13 · Identity and Scope

Identity is embedded through "training," not prompts

Prompts can be **forgotten, truncated, or overridden by users.** If identity collapses then, it cannot be explained to customers. Therefore, we chose not to rely on prompts.

We train under three conditions

We train the same question-answer pairs under ① **our prompt** ② **no prompt** ③ **another company's prompt**(including another assistant's identity statement). This preserves identity **even without a prompt or when overridden by another company's prompt.**

At serving time, prompts are "reinforcement"

Identity should remain intact without a prompt; **if it collapses, the training is wrong** — we make this a **validation item** and measure it under **both conditions, with and without a prompt.**

The default system prompt has four clauses and 209 characters

Clause	Why it is there
① **Identity**	It is an industrial AI assistant created by KOPENS — if underlying information leaks, it cannot be explained to customers
② **Evidence and "I don't know"**	It specializes in equipment · processes · quality · safety, **answers with evidence, and does not speculate when uncertain**
③ **Language**	Korean is the default, and it also answers in English. Multilingual models **tend to follow the input language**, so this is specified
④ **Scope**	Politics · diplomacy and similar topics are excluded — this is **scope definition, not censorship.** Equipment diagnostics do not require political commentary

The order is also intentional. Identity and guardrail data account for only **2%** of the total — too much degrades general performance.
Design also includes deciding how much training to use to define scope.

There is "one" source of truth

One system prompt string is the sole source, and **data generation · evaluation · serving all reference it.** Copies inevitably diverge — in fact, a copy once appeared in a build script and caused inconsistent wording.

Evaluate with the prompt "used in deployment"

Because prompts affect results, a score measured with another prompt **is not the deployment environment's score.** Therefore, evaluation and serving use the same source of truth.

14 · Retrieval

For industrial RAG, "retrieval quality" sets the upper bound on answer quality

Even the best model produces **a wrong answer when given wrong evidence**. General-purpose embeddings view "temperature rise due to bearing wear" and "temperature rise due to insufficient coolant" as **nearly identical sentences** — even though they require completely different actions in the field. We therefore trained a separate retrieval model.

EMBEDDING 4B

We add "deliberately confusing wrong answers" separately

In contrastive learning that uses other items in the same batch as negative examples, we deliberately mix in **documents that look relevant but are not the answer**. This is where the ability to distinguish them comes from.

RERANKER 0.6B

It learns "ranking" itself

It is trained to select the correct answer from a candidate list. As a cross-encoder that **reads the query and document together**, it detects **subtle differences** missed by embeddings.

Asymmetric encoding — add instructions only to queries

We add an instruction telling the query "what to find," but **not to the document.** Documents are long-lived assets once indexed; adding instructions to them requires **a full reindex whenever the instruction changes**.

We fixed the dimension as a "contract"

Upgrades do not require rebuilding the vector index. We fixed the retrieval model's output dimension(2,560) across versions. An upgrade that requires re-embedding tens of millions of chunks **is ultimately postponed**, and a postponed upgrade is equivalent to no upgrade — so we chose the dimension for **operability** rather than performance headroom.

Prevent "leakage" before evaluation

Group-level split	Items that **share a question or answer document** are treated as one group and assigned together to training or evaluation. Random row-level splitting places **the same document on both sides** and only inflates the score
Equipment-level split	Fault diagnosis is split **by equipment** — evaluation uses **equipment absent from training**
Breakdown by source	Fault diagnosis · procedures · specifications are evaluated **separately**. The overall average obscures the less-represented sources
Wording robustness	We measure again after **rewriting the same question in field language** — if the score collapses, the model memorized the format

15 · Time-Series Track

We use "small models" for waveforms — one per site

Language models excel at **explaining why**. For waveforms arriving every second, **finding what is abnormal** is a different task and requires a different scale.

A million-parameter scale is an "advantage"

It runs without a GPU, so it **can be deployed on an edge box**, and tailoring it to each asset costs little. **10,000 models for 10,000 assets** is a practical scale — an operating model that is not viable with large models.

We teach mechanisms, not equipment names

With bearing wear, vibration rises first and temperature follows, **whether in a pump or gearbox**. This allows us to start with public measurements and physics simulators, **even without customer data**.

Four ways industrial signals differ from general time series

Operation and shutdown are mixed	**A 0 during shutdown is not an anomaly.** Counting it as one floods the system with alerts
Normal changes with load	4 A at no load and 7 A under load are **both normal** — a single absolute threshold cannot handle this
Measurement gaps are common	**A sensor outage differs from a value of 0.** We therefore **mark missing data separately** for the model
Failures are rare	4 incidents in 6 months — about 2% of the total. **Rare events require a separate measurement method**

Three use cases

① Continuous monitoring

It assigns normal/anomaly scores by window and alerts when the threshold is exceeded.

② Prioritizing inspections

For "the temperature rose; what should be checked first," it provides an **order**.

③ Suppressing false alarms

Current and pressure rise under load and can look like a failure. **Reading that as normal** is half of this track.

④ Preserve an anomaly "fingerprint"

Storing features extracted from waveforms allows us to answer **"which past failure does this anomaly resemble?"** — covered on the next page.

Both tracks examine the same site. The time-series side identifies **what** is abnormal, while the language model explains **why**. And **we do not use models where statistics suffice** — some segments can be distinguished solely by deviation from the normal baseline, and adding a model there only increases cost.

16 · Evaluation

No judge model — the same value regardless of who measures it

Industrial AI has no leaderboard. We therefore built the measurement system ourselves, with one principle — **we do not let another model score it.** Because ground truth comes from formulas, schemas, and operational records, machines do the scoring, producing **the same score for the same input, regardless of who runs it or when**.

Axis	Items	What is measured and how
Industrial calculations	**600**	Engineering formulas in **30 families** (bearing fault frequency · life · NPSH · ASME shell thickness · orifice flow · OEE · process capability · welding heat input …). Items are generated randomly, and formulas calculate the ground truth. Correct if **relative error is within 1%**
Tool calls	22	Provides the schemas of **57 actual tools** on the operating platform as context. Tool names are checked against the set, arguments against each tool's declared JSON Schema (required fields · regex · enum), and multistep calls for **order**
Grounding fidelity	8	Provides actual analysis prompts as context and checks whether the response uses **only the given facts**. In particular, whether it states when the context says **"no data"**
Field diagnosis	**158**	6 months of actual compressor measurements — **the only one of the five whose ground truth comes from outside us** (Chapter 17)
Retrieval	277	Embeddings search the full corpus (Recall@1/5/10 · MRR@10); the reranker selects the correct answer from 8 candidates

7 families were excluded entirely from training

Of the 30 families, **seven** were not included in the training data. Testing included families would measure **memorization**. Testing excluded ones shows whether the model "can formulate an unseen equation."

Retrieval evaluation verified "0 leakage"

Items sharing question and answer documents were split into training and evaluation **as groups**. If the same document appears on both sides, the score rises while capability remains unchanged.

Evaluation also distinguishes "0" from "unknown." We deliberately included **contexts with missing values** in the training data and set **answers stating "not measured"** as correct in those cases; the grounding fidelity axis measures this directly.

What it means for scoring to be "mechanized"

Axis	Rule
Industrial calculations	Use **the last occurrence of the "Answer: …" line**; if absent, use the first number parseable when scanning backward from the end of the body (because the model may show several intermediate calculations, **the last is the conclusion**). Comparison uses $\lvert answer - truth \rvert / \lvert truth \rvert \le 0.01$; if the ground truth is 0, compare absolute values. **Units are not parsed** — attempting to distinguish "Hz" from "hz" or "mm/s" from "mm·s^{-1}" turns notation differences into scoring errors. The item specifies the unit, and only the number is checked
Tool calls	The tool name is checked against the declared set, and arguments are counted in **three separate categories** — missing required fields · undeclared arguments · format violations. Regexes are actually matched: if a date is constrained by `^\d{4}-\d{2}-\d{2}$` , `2026-09-19` passes and `2026/09/19` violates it. **Because the operating platform enforces the same format, calls that fail this validation actually receive 400**
Evidence citation	We encoded the rule that "listing names alone is not evidence." It checks whether **a numeric value appears within 80 characters after the signal name** — "Check the temperature" scores 0, while "The temperature rose from 62→89°C" scores 1. Counting is by **signal**, not sentence or claim

How reviewers can verify it "themselves" — without trusting us

What	How
Is the trained version actually being served?	Expose the base model and trained version simultaneously **under different names**, then send the same question to both. **If the responses are effectively identical, the training was not applied**
Did training leak into evaluation?	Compare evaluation questions and answer documents against the training set and count exact matches. Items sharing questions or documents must be split **as groups**
Did it memorize the format?	Rewrite the same question **in field language** and measure again. If the score collapses, it memorized the format rather than the content

More items do not provide more information. Field-diagnosis failures comprise **four independent events**. Overlapping windows produce 240 items and make the error bounds look "narrower," but they merely reslice the same 83 hours — **looking more precise is not the same as becoming more accurate.** We therefore use **158** nonoverlapping segments. Scores are **being remeasured on the current product path** and, once complete, will be entered **in the same table**.

17 · Field Diagnosis

To Avoid Creating a Boy Who Cried Wolf — Count False Alarms Separately

This is the only one of the five axes where **the ground truth comes from outside**. We use **6 months of measurements** from an operating compressor and **faults recorded in the operating logs** for that period as ground truth. The scorecard is not a rule we created but **what actually happened**.

Composition of the 158 Questions — "Normal" Is Divided into Three Types

Interval	Questions	Why It Is Divided This Way
Fault	**44**	Intervals recorded as faults in the operating logs. Missing one here is a **missed detection**
High-load normal (hard normal)	**40**	🔴 Intervals that **look abnormal but are normal**. Temperature and vibration are elevated, but the equipment is sound. Raising an alarm here is a **false alarm**
Ordinary normal	60	Quiet intervals. There is no excuse for raising an alarm here
Pre-fault	14	**Not scored.** We only observe "how far ahead the model detects" — scoring an interval with no definable ground truth makes the score false

Why False Alarms Are Separated from Missed Detections

If both are combined into one "accuracy" metric, **a model that raises an alarm for every interval** receives a good score because it catches every fault. Such an assistant is **turned off within a week in the field.** Once turned off, it catches nothing.

Scoring in One Line

We extract **one normal/abnormal classification line** from the model's answer and compare it automatically with the records. A person does not read the narrative and assign a score — because the score would change with the reader.

🔴 Reselecting the "Hard Normals" — When the Scorer Was Wrong

Initially, we selected the **intervals with the highest motor current** as hard normals. But the 40 selected this way **were statistically indistinguishable from the fault intervals** — current · intake valve duty · pressure were effectively identical. They were **abnormal intervals absent from the operating logs**. Had we left them unchanged, a model that read the signs "correctly" would miss 40 questions — the **scorer, not the model, would be wrong**.

We therefore changed the criterion. A compressor in normal operation **cycles on and off.** We excluded windows with continuous operation in which a specific pressure signal disappeared entirely(72), and the remaining "hard normals" have **current 2.3 times that of ordinary normals**, but the compressor cycles normally and pressure remains present. They can be solved only by **distinguishing "the value is high" from "it is a fault."** This has an implication for reviewers — **labels in public data must not be accepted at face value.**

How We Selected the Data

The ground truth comes from outside us(the operator's fault records are the labels). These are actual faults **during operation, not in a laboratory** — many public datasets induce faults artificially or through accelerated testing. Because the signals use **engineering units**(pressure · temperature · current), our assumptions do not enter at the feature-extraction stage.

How We Segmented the Intervals

We split the data into 2-hour windows **without overlap**(overlap increases only the number of questions, not the information), and summarize each window as 8 rows at 15-minute intervals. We discard windows with fewer than 8 rows due to collection gaps, windows **within 6 hours before or after maintenance**, and windows that merely overlap a fault interval.

We Defined the Answer Format

```
판정: 정상 | 이상      (Korean for "Verdict: normal | abnormal" — the literal string the scorer matches)
```

The scorer extracts this line with a regular expression(full-width colons are also accepted). If the format is absent, the answer is accepted only when exactly one of the two terms appears within the first 200 characters; if both or neither appear, it is **counted separately as a format violation** — that is not a judgment issue but an **instruction-following** issue, and mixing it with wrong answers obscures the cause.

We first verified that the task was valid. Before evaluating our model, we tested it three times with a strong external general-purpose model and revised the questions — without signal descriptions, the task became an encoding-guessing exercise(the valve signal's meaning was interpreted in reverse), and even with descriptions, continuous operation could not be identified as abnormal without knowing **"this equipment's usual pattern."** The task became valid only after providing the mean from a period well before the fault as a reference value. Providing a reference value does not leak the answer; it is the **actual diagnostic process**.

We will add measurement-based data in the next edition. The next edition will add **measurement-based data** — field data best demonstrates sensor drift · missing values · changes in operating conditions. We will **evaluate again with the same 158 questions** to compare editions(the questions must remain fixed for the comparison to be valid).

18 · Deployment and Next Edition

On-Premises — Models and Data Stay In-House

Nothing leaves the premises. The models run **on the customer's hardware, using the customer's data**. With no network requirements, they work as-is in closed networks.

Delivery Format

Format	Quantized checkpoints as a compressed archive (`tar.zst`). It contains one model directory and is **served as-is**
Container Image	By separate agreement
Interface	**OpenAI-compatible endpoint** — existing clients work as-is
Network Requirement	**None.** Both models and data stay in-house

GPU

Recommended	A current-generation GPU with **native 4-bit floating-point (NVFP4) acceleration**
Previous Generation	It works — **but provides no acceleration benefit**
Language Model	Served on **1 GPU** (21 GB weights + KV cache)
Two Retrieval Models	Both can be loaded on one GPU

Size

Model	Quantized
AURA-35B	Approx. **21 GB**
AURA-Embedding-4B	Approx. 6 GB
AURA-Reranker-0.6B	Approx. 1 GB

Included in the Next Edition

Confirmed

- **Performance metrics** — once remeasurement on the production path is complete, the table on page 16 will be filled in
- **Measurement-based diagnostic data** — field data will be added to synthetic data (pages 3 · 17)
- **Multi-turn tool use** — this edition focuses on single calls
- **Preference learning** — it is on the roadmap
- **Vision** — a weld defect assessment benchmark and industrial images are ready. **We establish the criteria first, then train**

Separate Track — Waveform Analysis Series

This **million-parameter-class** series detects anomalies in waveforms and is **separate** from the language model. It is **small enough to run without a GPU** and can be deployed at the edge. This edition covers **the architecture and the criteria**; **product integration is covered in the next edition.** Structurally, **identifying what is anomalous in a waveform and explaining why** are separate tasks (page 15).

License

Adoption terms are **negotiated per contract** with KOPENS. Open-source component licenses and notices are delivered **with the model artifacts**.

See Also

AURA Model Overview	docs.plantpulse.io /ai/overview/aura-models
AI Platform Documentation	docs.plantpulse.io /ai
Data Lake Technical White Paper	docs.plantpulse.io /platform/whitepaper

Contact

Phone · Email	1588-5268 · webmaster@kopens.com
Inquiries · Demo	kopens.com/contact
Products · Documentation	plantpulse.io · docs.plantpulse.io

㈜한국오픈솔루션 · KOPENS, Inc. © 2026 KOPENS. PlantPulse® is a registered trademark of ㈜한국오픈솔루션. This document is Rev. 2026.09 — the **architecture and validation methods edition**.