Anomaly-Based Hydraulic Pump Predictive Maintenance with Field Knowledge Accumulation

Data Engineering Project Proposal

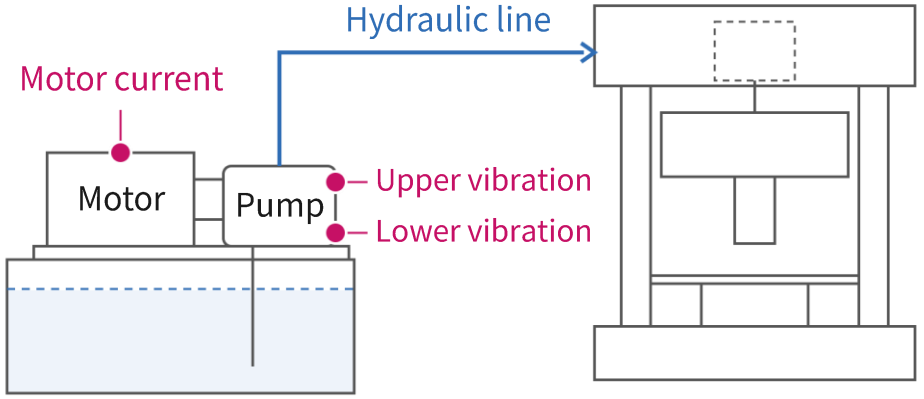
2026.10.22

Team 5
JANG HAIWON, KIM DONGGEUN, KIM MYUNGGUN, YOON SUNGGYOO

Problem Statement

- A **hydraulic pump** failure can stop the line for up to **2–3 days** (repairs, parts supply) [1]
  → Anomalies must be caught **early**
- Sensor-based detectors **already exist**, but they only tell **whether an anomaly has occurred**
  e.g., LSTM-AE trained on normal data [1]
- **Past events** with the same symptom are hard to find, and workers' **tacit knowledge** is not recorded
  "Anomaly signs of the hydraulic pump could be identified by skilled workers or managers from its operating noise." [1]
- Even if detections are given to an LLM, **no history** exists to ground an accurate explanation

→ We need to attach field knowledge to detections and accumulate it over time



Hydraulic unit, press, and sensor locations (● sensor)

Sensor variables (10 Hz)

TimeStamp	**Upper vib.**	**Lower vib.**	**Motor current**	Equip. state

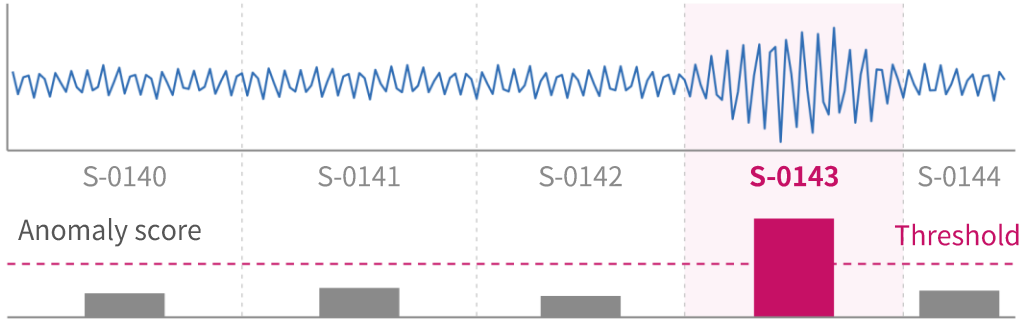
[1] MSS & KOSMO, "Predictive Maintenance AI Dataset for Metal Forming: Guidebook," KAMP, 2022 (in Korean).

Project Goals

Goal 1. Anomalous Segment Selection

Select anomalous segments, saved by **segment ID**

Quality check, segmentation, per-segment anomaly scores, alarm rules

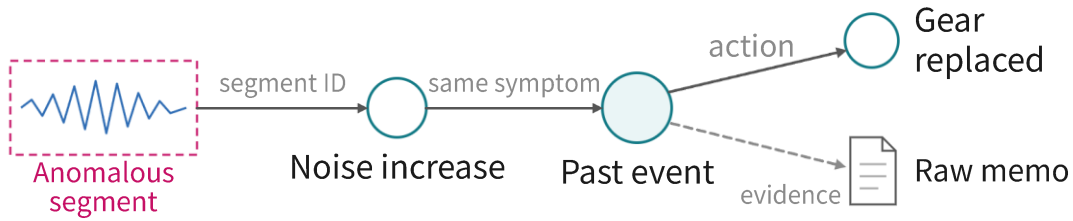


Per-segment anomaly scores and selection (example)

Goal 2. Field Knowledge & Explanation

Structure **tacit knowledge** and **past actions**

Explain likely causes, grounded in past maintenance events



Example explanation by the AI maintenance agent

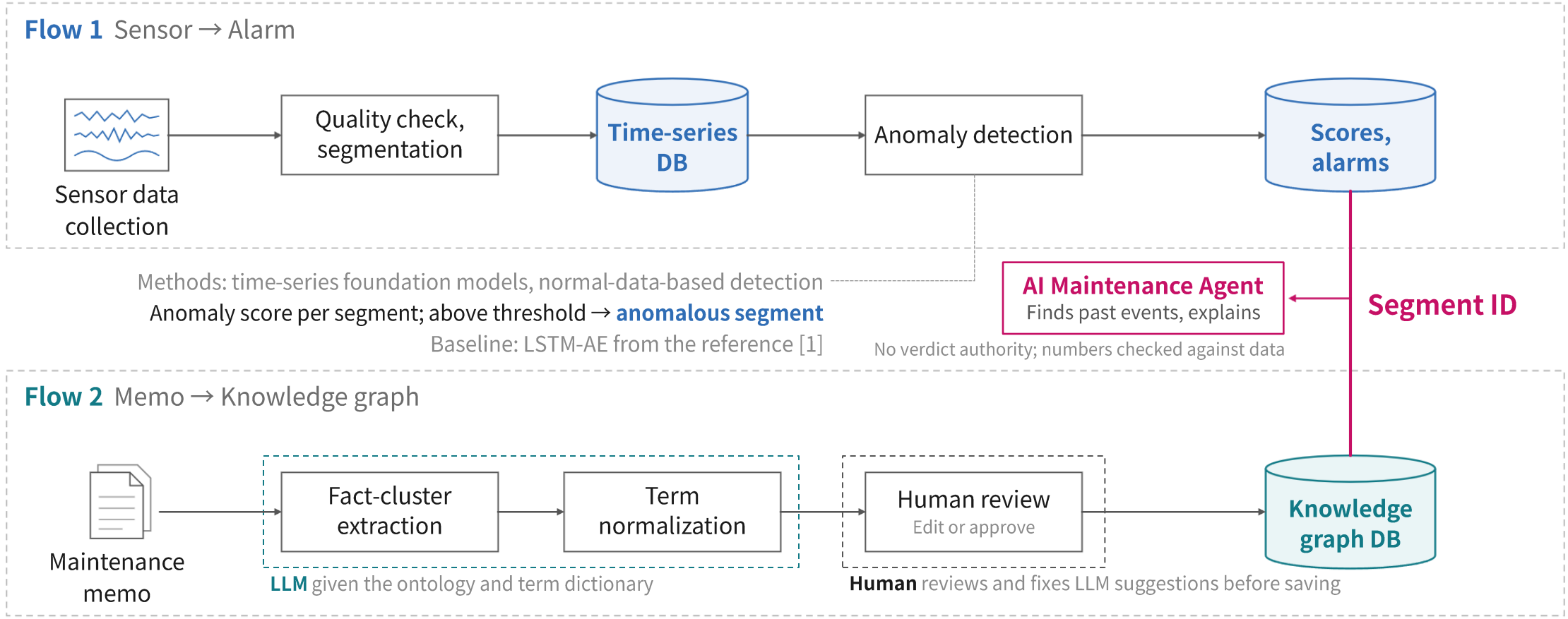
"Last time, a **gear replacement** fixed this noise" (memo M-0007)

Finding past events and explaining with evidence (example)

Both built as one system

Data Collection · Storage · Retrieval Pipeline

Overall Pipeline

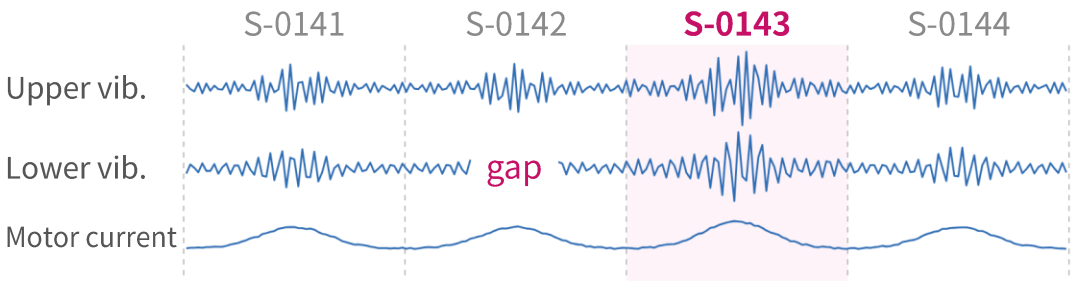


[1] MSS & KOSMO, “Predictive Maintenance AI Dataset for Metal Forming: Guidebook,” KAMP, 2022 (in Korean).

Data Collection & Preprocessing

Sensor data public manufacturing dataset [1]

One hydraulic pump: upper/lower vibration, motor current,
10 Hz, with normal and abnormal runs



Per-stroke segments with segment IDs (example waveform)

Preprocessing plan

1. Check missing values, duplicates, and gaps
2. Segment by press stroke
3. Assign segment IDs (links anomalous segments to memos)
4. Check for operating-condition (load) changes

Maintenance memos synthetic

Without real memos, we write **synthetic memos**
and **ground-truth links** for matching equipment/segments

Synthetic memo record (example)	synthetic = true
memo_id	M-0012 (version 1)
equipment_id	P-01 (assigned in scenario)
segment_id	**S-0143**
raw_text	“P01 pump noise increase. Gear check planned.”

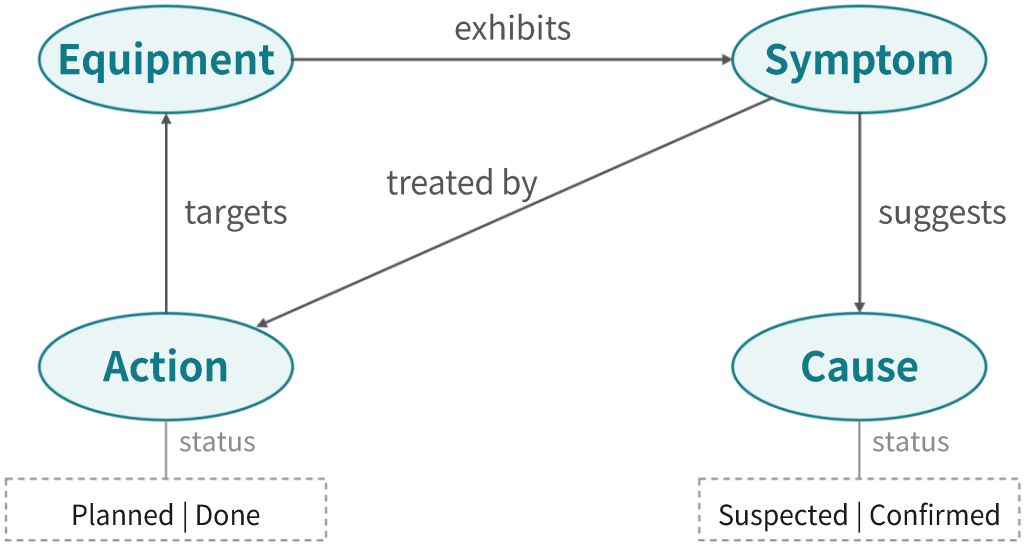
Test cases	What we check
Normal link	Attached to the intended segment?
Duplicate	Is duplicate loading blocked?
Revision	Old version kept and links realigned?
Unknown cause	Suspected vs. confirmed kept apart?

[1] MSS & KOSMO, “Predictive Maintenance AI Dataset for Metal Forming,” built by INTERX, 2022 (in Korean).

Ontology & Memo Processing

- **Ontology**: a predefined schema of equipment, symptoms, causes, actions, and their allowed relations
  Defines what to extract (scope), how to align terms (normalization), and how to store them (graph structure)

Example



Classes (ellipses), allowed relations, status attributes (dashed)

Memo processing example

“ P01 pump noise increase . Inspection planned .”
Equipment Symptom Action Status

1. LLM fact-cluster extraction equipment, symptom, action, status
2. Term normalization P01 → P-01 (alias list)
3. Human review edit or approve, then save

Equipment	Symptom	Action	Status
P-01	Noise increase	Inspection	**Planned**

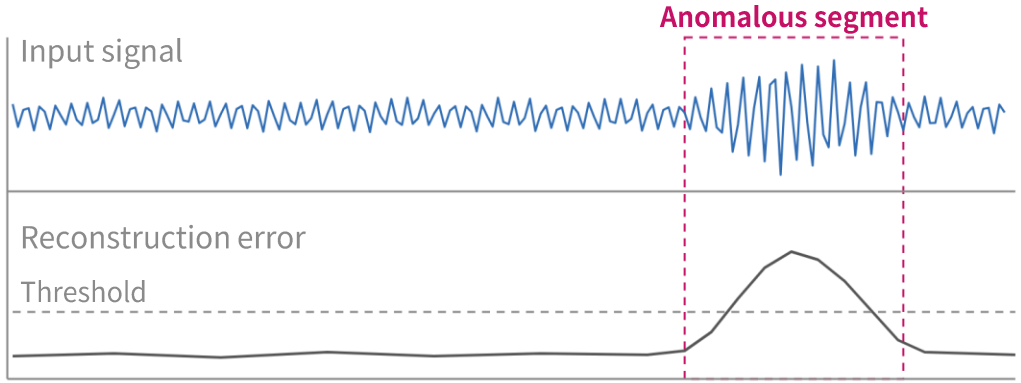
LLM output is only a suggestion;
it is saved only after a Human reviews and corrects it

Comparison: synonym dictionary vs. LLM, on the same ground truth

Related Work

MOMENT time-series foundation model [1]

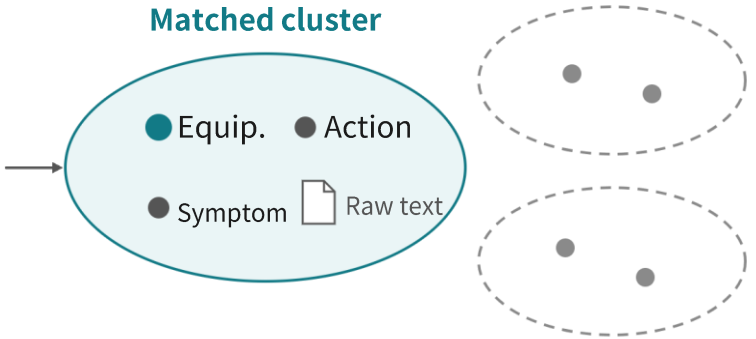
General model pretrained on the public Time-series Pile
Detects anomalies by reconstruction error (zero-shot)



(a) Finding anomalies by reconstruction error (concept)

OG-RAG ontology-grounded retrieval [2]

Stores and retrieves ontology-grounded fact clusters (hyperedges), so evidence is easy to trace to the raw text



(b) Retrieval by fact cluster, traced to the raw text (concept)

In this project

Flow 1: a candidate detection method, compared with normal-data-based methods

In this project

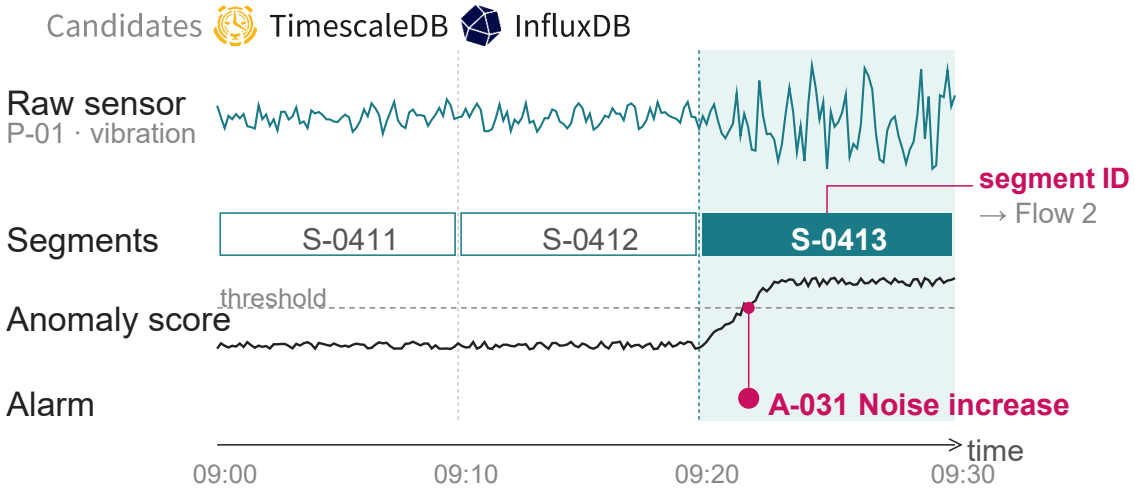
Flow 2: reference for storage and retrieval (hyperedges in our DB design)

[1] M. Goswami et al., “MOMENT,” ICML 2024. [2] K. Sharma et al., “OG-RAG,” EMNLP 2025.

DB Structure

Flow 1 storage time-series DB

Raw sensor data, segments, anomaly scores, alarms

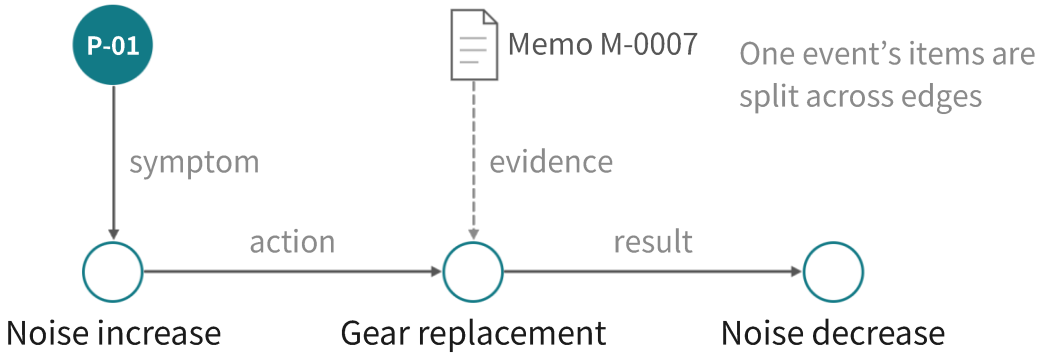


(a) Time-series rows: raw data, segments, scores, alarms on one time axis

Flow 2 storage graph DB

Equipment, events, symptoms, actions, memos
(linked by **segment ID**, not raw sensor data)

Candidates Neo4j Apache AGE



(b) Knowledge graph: equipment, symptom, action links

Save checks
Both stores

1 IDs, required fields

2 Block duplicates

3 Keep revisions as versions

4 Trace evidence by raw memo and version ID

DB Candidates

Type	Candidate	Overview [1]	Query language
Time-series DB	TimescaleDB	PostgreSQL extension adding time-series features	SQL
	InfluxDB 3 Core	Purpose-built time-series DB, light and fast	SQL, InfluxQL
Graph DB	Neo4j	The most widely used graph DB	Cypher
	Apache AGE	PostgreSQL extension adding graph features	openCypher
	TypeDB	Expresses multi-entity relations directly	TypeQL

Selection criteria

After the proposal, same queries

Ingest speed	Segment query	Link query	Update effort	Ops burden

[1] Official docs of TimescaleDB, InfluxDB 3 Core, Neo4j, Apache AGE, and TypeDB.

Demo & Evaluation Plan

Demo (in preparation)

1 **Prepare test cases with synthetic memos**
Normal link, duplicate, revision, unknown cause

2 **Save to DB after Human review**
Block duplicates, keep revisions as versions

3 **Agent retrieves and explains**
Anomalous segment → past events → grounded explanation

Retrieval questions

Q1 What was done last time on the same pump?
Anomalous segment → event → past same-symptom events → action

Q2 Which memo did this link come from? Revisions?
Link → raw memo and version → revision history

Metrics

Flow 1	**Detection performance** (incl. misses, false alarms), ingest volume, latency, gap handling
Flow 2	Extraction accuracy, planned/done status errors, wrong-merge rate
Retrieval	Retrieval accuracy and time, **explanation grounding** (cited events match ground-truth links?)

Timeline

	Midterm (Proposal) Current stage	After the proposal	Final
Tasks	Finalize data setup Survey DB candidates Design the pipeline	Select the DBs Small demo with synthetic memos	Connect DB Build and validate retrieval
Outputs	Proposal talk	DB rationale, demo results	Final presentation
Scope		**Small demo · synthetic memos** → **real-world application**	

Thank you

Thank you for listening to the presentation

Pohang University of Science and Technology