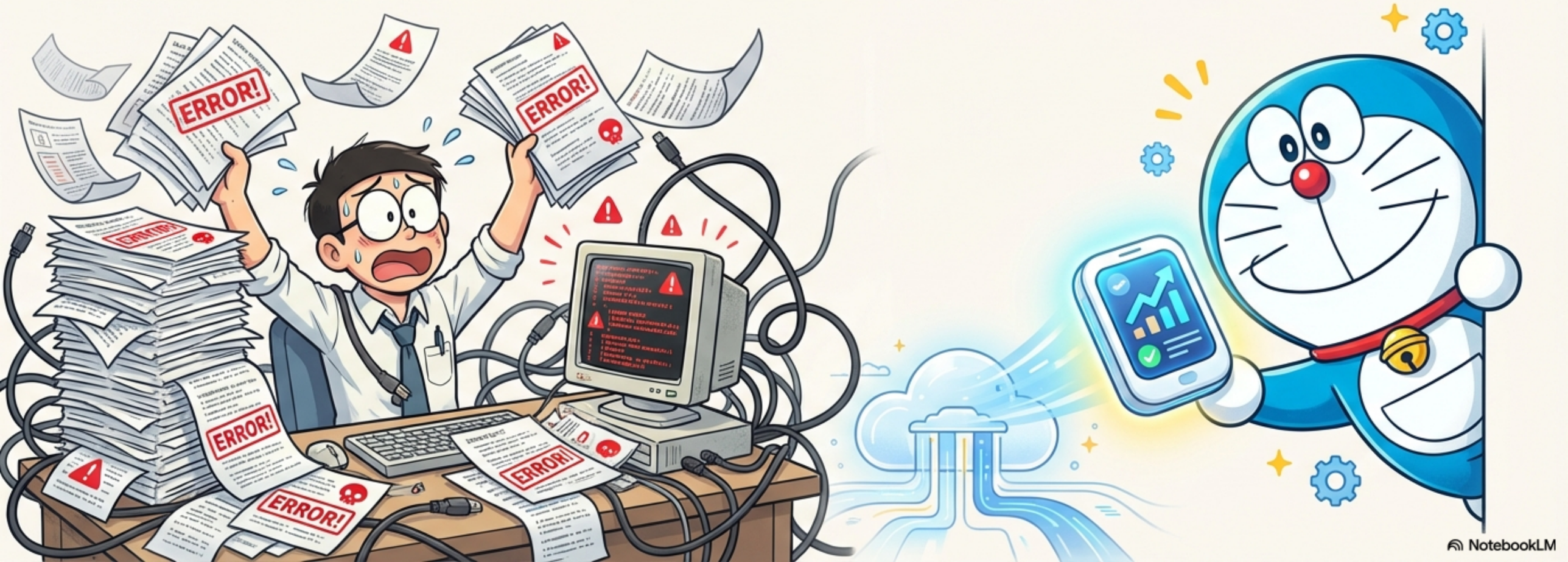
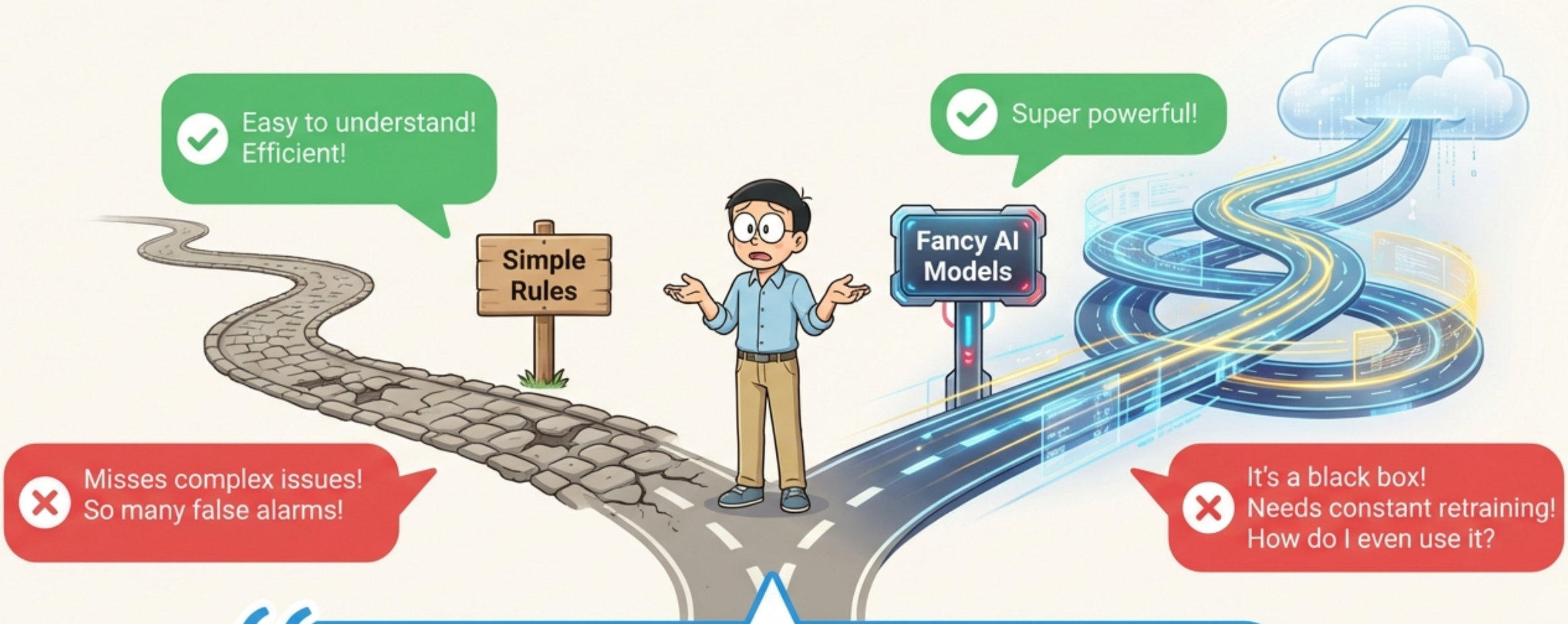
# MonitorAssistant: A Gadget from the Future for Cloud Monitoring!

## Simplifying Cloud Service Monitoring via Large Language Models



# "Doraemon, help! Monitoring our cloud is so hard!"



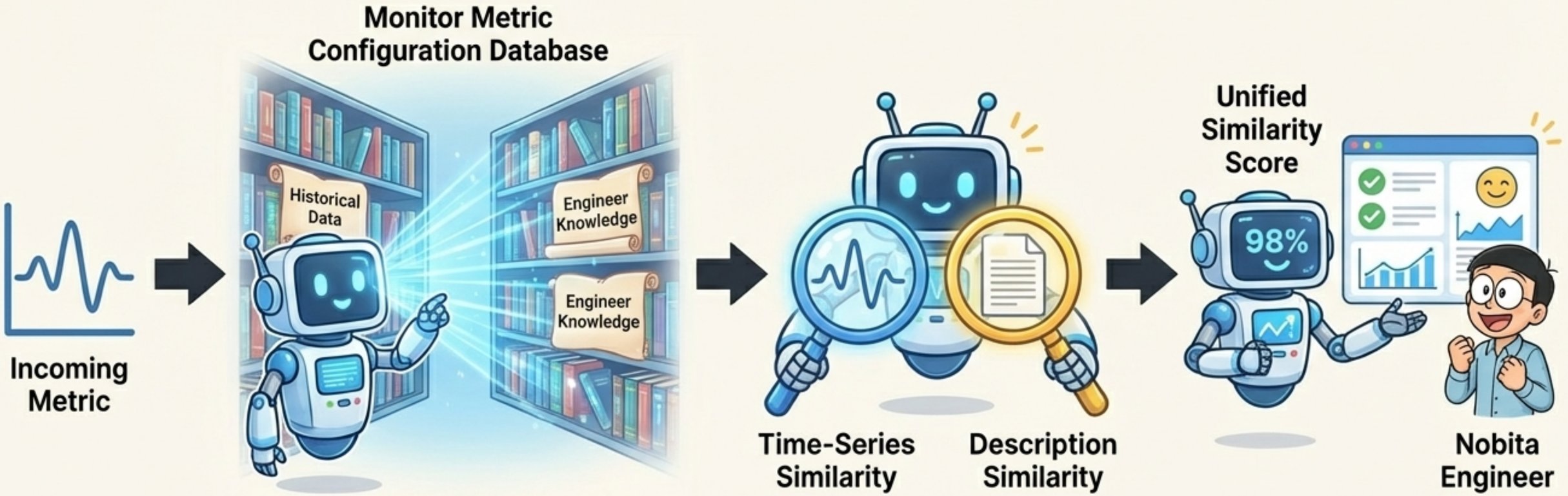

**“There’s a huge gap between what’s powerful and what’s practical. We’re stuck in the middle!”**

# “Leave it to me! Ta-da! The MonitorAssistant!”

1. ***Smart Setup (Configuration Recommendation):***
   Automatically picks the best monitoring model and settings for you. It inherits knowledge from past incidents.
2. ***Smart Reports (Anomaly Alert):***
   Tells you what's wrong, why it might be happening, *and* gives you a troubleshooting guide. It's more than just an alarm.
3. ***Smart Learning (Feedback Loop):***
   You can talk to it in plain English to fix false alarms. It listens, learns, and gets better over time!

Powered by Large Language Models (LLMs), **`MonitorAssistant`** bridges the gap between simple rules and complex AI.

# It learns from the past to prepare for the future!

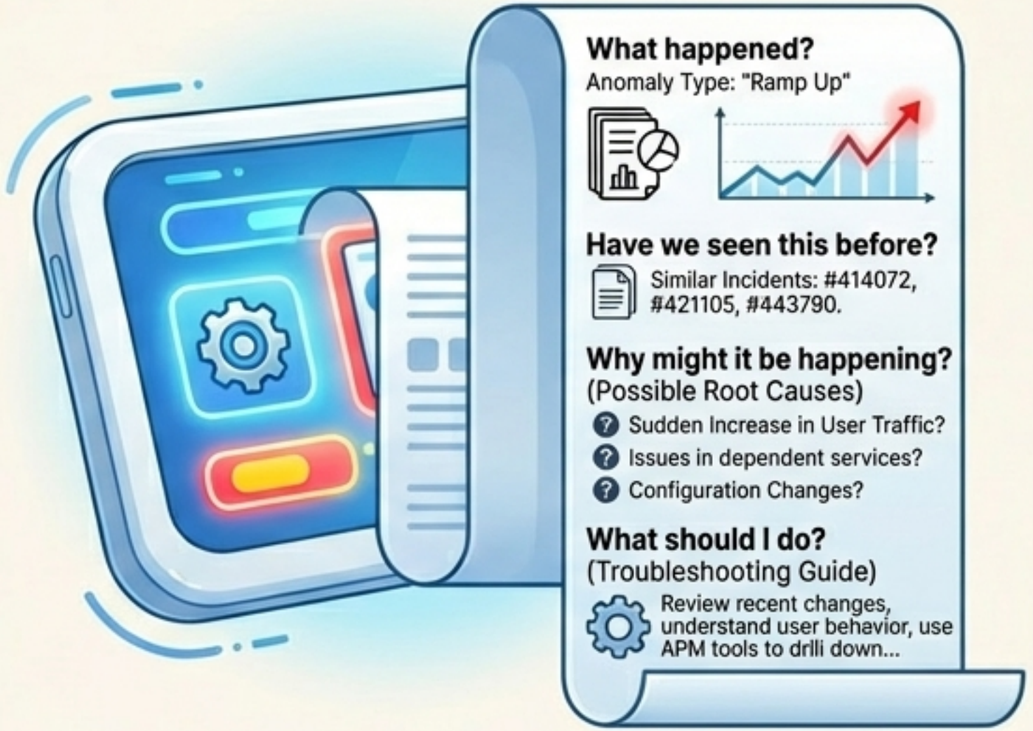



`MonitorAssistant`'s `Monitor Configuration Infusion` technique uses a unique **Unified Similarity score**. It combines time-series data patterns (using shapelets) and natural language descriptions to find the most similar past metric and recommend its proven monitoring setup.

# It doesn’t just say "BEEP!" It gives you a full report!

## The Old Way


## The MonitorAssistant Way




By analyzing similar historical incidents, `**MonitorAssistant**` generates practical, guidance-oriented reports that help engineers troubleshoot faster.

# You can talk to it, and it actually understands!

I found some false alarms. There's a spike at 2023-11-19 08:05:20. And we're seeing some dip-type false alarms.

Why do you think they are all false alarms? Please tell me the reasons behind your decision.

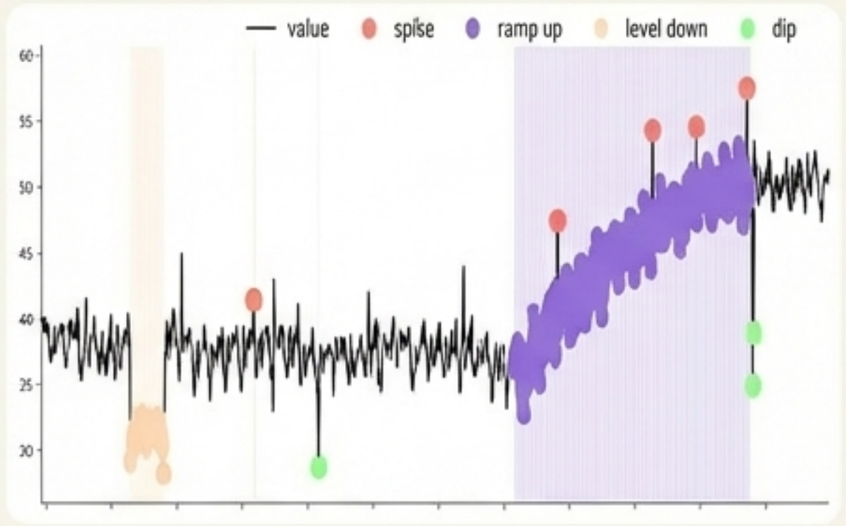
In this metric, we do not care about any dips. The magnitude of these spikes are too small to consider them outliers.

Understood! I will apply a patch to ignore all dips and any spikes that are too small. Here's what the alerts look like now. Is this correct?


The **`LLM-Engineer-In-The-Loop` workflow** allows engineers to provide feedback in natural language. The system understands the intent and applies rule-based patches, tuning the model without **requiring any algorithm expertise.**

# Putting the MonitorAssistant to the test at Microsoft!

## Part 1: Initial Run

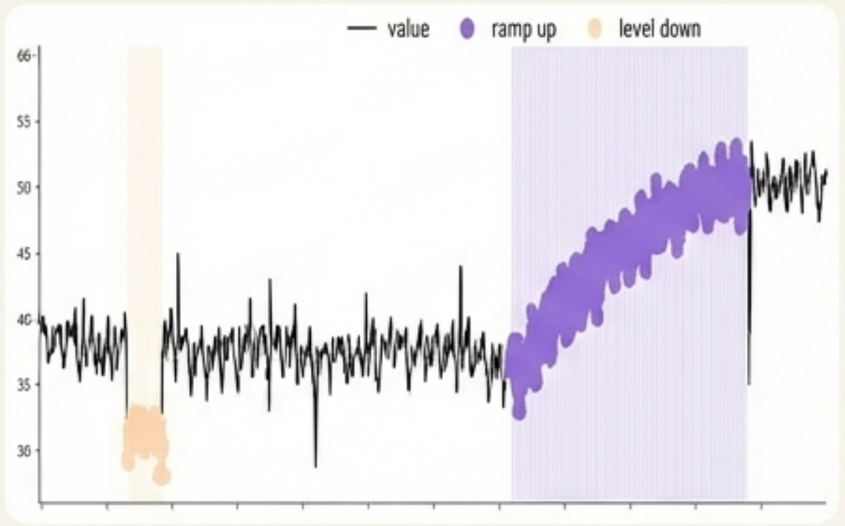



The first recommendation detected the real issues, but also some false alarms (spikes and dips).

## Part 2: Feedback

The engineer provided simple feedback in plain English.

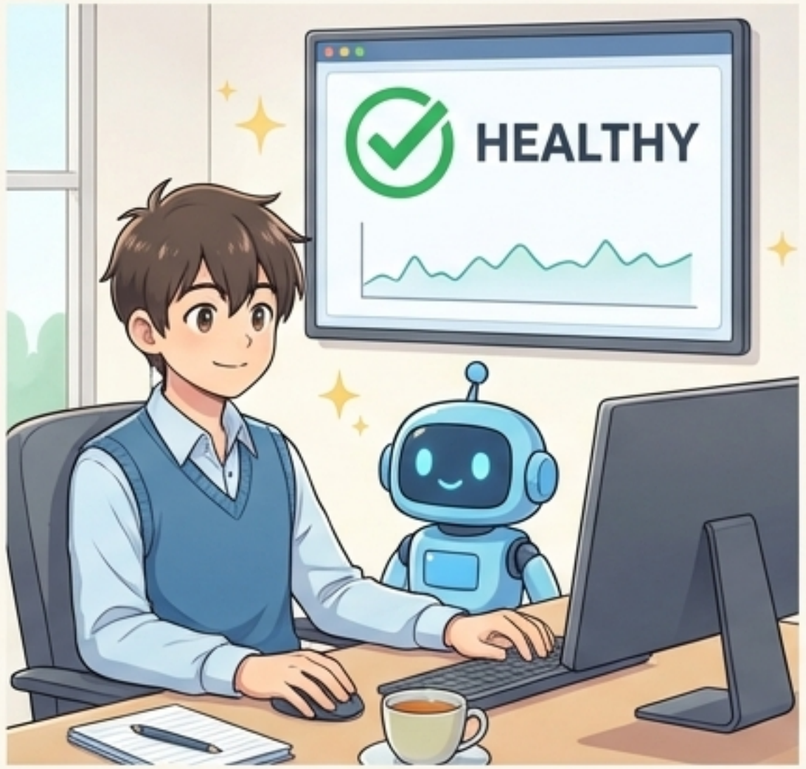
## Part 3: Final Result




Success! After learning from the feedback, the false alarms are gone.

In a real-world Microsoft service, `**MonitorAssistant**` successfully identified all practical anomalies and, after a single round of natural language feedback, eliminated all false alarms.

# A Happy Ending for the Nobita Engineer!




✅ **Automated Expertise:** Intelligently recommends the right monitoring model by inheriting knowledge from a database of past metric-incident pairs.

✅ **Actionable Insights:** Generates anomaly reports that explain the anomaly's shape, list similar historical incidents, and provide a troubleshooting guide.

✅ **Effortless Improvement:** Allows any engineer to refine the system using plain English, no data science degree required.

**'MonitorAssistant'** makes advanced anomaly detection practical, interpretable, **and user-friendly**, successfully bridging the gap for large-scale cloud services.

# The Future of Monitoring is Here!

**Paper Title:** MonitorAssistant: Simplifying Cloud Service Monitoring via Large Language Models

**Authors:** Zhaoyang Yu, Minghua Ma, Chaoyun Zhang, Si Qin, Yu Kang, Chetan Bansal, et al.

**Conference:** FSE '24: 32nd ACM International Conference on the Foundations of Software Enginieering