



NOBODY ON
THE CHANNEL
NOTICED.

THE RADIO
LOG DID.

Checks each read-back of a radioed safety order against the plant registry; wrong ones, and any it cannot check, reach the supervisor.

MISMATCH

H-01 · CH-01 · English

tag: XV-203 ≠ XV-208 XV-208 is on line A (registry)

OPERATOR H-01-0

Close valve two-oh-three, line Bravo.

Copy, closing two-oh-eight, line Bravo.

FIELD TECH H-01-1

extraction Gemma-4-31B-IT · AMD-hosted API

```
zsh · build/
$ jq -r '...' bench/results/text/llm-judge-r1-reaudit.json
llm_judge false-VERIFIED 2 false alarms 1/48
$ jq -r '...' bench/results/text/ladder-r1-reaudit.json
Gemma-4-31B-IT false-VERIFIED 0 false alarms 2/48
regex false-VERIFIED 0 false alarms 0/48
# 3 more rungs, false-VERIFIED 0 on each
```

TWO LINES. NOBODY FLAGS THEM.

OPERATOR Close valve two-oh-three, line Bravo.

FIELD TECH Copy, closing two-oh-eight, line Bravo.

no flag · no card · the crew walks to XV-208

LINE A · CLARIFICATION

LINE B · PRESS STATION



In the registry the demo runs on, XV-203 is on line B at the press station; XV-208 is on line A, in clarification.

READ-BACK IS THE CHECK. NOTHING CHECKS IT.

WRONG READ-BACKS CAUGHT, OF EVERY 100 • AVIATION, HELMKE 2021

TOWER



63

APPROACH



50

filled: caught by the controller • empty: went through

NO
RATE

published for plant radio

- 1 the diff, checked against the registry
- 2 the read-back word that differs
- 3 a human confirms or dismisses

THE MODEL POINTS. CODE DECIDES.

1 · THE EXCHANGE THE MODEL READ

ORDER
Close valve two-oh-three,
line Bravo.

READ-BACK
Copy, closing two-oh-eight,
line Bravo.

H-01 · written words, text-scripted



2 · WHAT GEMMA-4-31B-IT RETURNED

```
{
  "commands": [
    {
      "action_span": "closing",
      "line_span": "line Bravo",
      "state_span": null,
      "tag_span": "two-oh-eight"
    }
  ],
  "kind": "readback",
  "retracted_spans": []
}
```

AMD Token Factory developer.amd.com.cn

median latency 4.304 s, first pass

cost USD 4.8937e-06



3 · WHAT CODE DECIDED

- 1. instruction: grounded + covered (1 command(s))
- 2. read-back: grounded + covered (1 command(s))
- 3. audio: no gate (no token probabilities, no unprompted pass)
- 4. instruction: XV-203 line B (registry ok)
- 5. compare: tag: XV-203 ≠ XV-208

○ **MISMATCH**
tag: XV-203 ≠ XV-208

speech-to-text: not run yet · span extraction: AMD Token Factory · referee: local, plain Python · queue: local
receipt, as stored · request_id c9647a04cfcc4eeaadbfa965a4d74c6d · model_served self-dploy/Gemma-4-31B-IT (sic, as served) · host
developer.amd.com.cn · latency_s 35.4972 (a slow call; the first pass's median was 4.304 s) · cost_usd 4.8937e-06

JUDGING IS NOT THE MODEL'S JOB.

LLM-as-judge
DeepSeek-V4.1-Flash · no registry

~~VERIFIED~~

Order and read-back match; **the valve does not exist.**

It also passed **2** wrong read-backs.

◦ ABSTAIN

R-01 · CH-01 · English

ABSTAIN · unknown tag cannot clear this exchange: a human must check it

OPERATOR R-01-0 *no recording*

XV-205 not in registry

Close valve two-oh-five, line Charlie.

Copy, closing two-oh-five, line Charlie.

FIELD TECH R-01-1 *no recording*

Confirm: needs follow-up

Dismiss...

2 → 0 WRONG READ-BACKS PASSED.

The LLM-as-judge baseline passed 2. Each of the five rungs passed 0.

rung	false-VERIFIED	false alarms on 48 correct read-backs · dashed: 5 % budget	rule
regex rung 0 · no model	0	0/48	meets
MiniCPM5-2B smallest on the ladder	0	48/48	all 48 to a human
Gemma-4-31B-IT thinking off · re-audit	0	2/48	meets
DeepSeek-V4-Flash thinking off · re-audit	0	4/48	over budget
Qwen3.8-27B thinking on · re-audit	0	4/48	over budget
LLM-as-judge baseline · DeepSeek-V4.1-Flash · no registry	2	1/48	passed XV-205

On written words, regex meets the rule too; audio decides.

WHO PAYS

PRICED PER CHANNEL.. BOUGHT BY HSE.

1,318

mills, Indonesia 2022 (Trase)

× 30 %

addressable*

× 6

channels per mill*

× USD 2,400

per channel-year*

≈ USD 5.7 M

a year, mills only (estimate)

★ assumption

WEDGE

FREE 30-DAY
LISTEN-ONLY
BASELINE

WHO ELSE LISTENS TO PLANT RADIO

LOGGERS TRANSCRIBE. NONE WE CHECKED COMPARES.

	TRANSCRIBES RADIO	COMPARES A READ-BACK	PLANT TAG REGISTRY
NICE Inform AI	yes	—	—
Eventide NexLog + Speech Factor AI	yes	—	—
Motorola Assist	yes	—	—
IntelliPERMIT	—	radioed codes, not speech	yes
HAAWAI (air-traffic research)	yes	yes, aviation	no
SayAgain	planned	yes	yes

— not in the public documentation we read, 2026-10-09 · sources: evidence/sources.json

WHAT WE ABSTAIN ON.

◦ **ABSTAIN**

“No recorded speech has been measured.”

“The plant, its 120 tags and every exchange are fictional...”

“...what a language model adds is unproven until the audio rows exist.”

NO AUDIO EVALUATION IN THIS SUBMISSION

Audio row

the rule on recorded speech

Failure case

−6 dB look-alike tags

Kill test

English, Bahasa or auto

Blind auditor

by ear vs the queue

THE ASK

PUT IT ON ONE MILL'S RADIO.

ONE MILL. THIRTY DAYS. LISTEN ONLY.

Q1 2027 paid pilot, one palm-oil mill

Q3 2027 DCS check: did it close?

Q4 2027 Malaysian palm belt



github.com/edycutjong/sayagain
[Release v0.4.3](#)

EDY CU • JAKARTA