PROOF OF SUCCESS METRICS · PRD §04

Log Analysis AI — เอกสารพิสูจน์ผลตามเป้าหมาย

ผู้ช่วยวิเคราะห์ Log และเฝ้าระวังระบบด้วย AI

ผู้จัดทำ: ณัฐปคัลภ์ กันทะศร (Natpakan Kanthasorn) | **วันที่:** 10 Jun 2026 | **ชุดทดสอบ:** 540 sessions (held-out test set)

1. บทสรุป (Executive Summary)

ผลลัพธ์: ผ่านครบทุกเกณฑ์ (ALL PASS) — ระบบถูกวัดผลบนข้อมูลทดสอบ 540 sessions ที่โมเดล **ไม่เคยเห็นตอนเทรน** และผ่านตัวชี้วัดทุกตัวที่กำหนดไว้ใน PRD §04 โดยตัวเลขทั้งหมดคำนวณจากการเทียบคำทำนายกับเฉลย (ground truth) ด้วยสูตรมาตรฐาน จึง **ตรวจสอบย้อนกลับและรันซ้ำได้**

2. ผลการวัดเทียบเป้าหมาย

ตัวชี้วัด (Metric)	เป้า PRD	ผลที่ได้	สถานะ	ความหมาย
Recall — high-severity	≥ 0.90	**1.000**	PASS	ตรวจจับ incident รุนแรงได้ครบ (เป้าหลักของ PRD)
Precision — overall	≥ 0.70	**0.831**	PASS	alert ที่ส่งออกไป เป็นของจริงกี่ %
F1 — overall	≥ 0.77	**0.791**	PASS	สมดุลระหว่าง precision กับ recall
AI จับได้ที่ rule พลาด	> 0	**37**	PASS	พิสูจน์ว่า AI ดีกว่า rule-based เดิม

ตัวเลขเสริม: Accuracy = 0.928 | MTTD (เวลาเฉลี่ยกว่าจะตรวจเจอ) = 0.15 หน้าต่าง | เทียบ baseline แบบ rule-based เดิม: Recall เพียง 0.418 ขณะที่ AI ได้ 0.755

3. วิธีพิสูจน์ (Methodology)

3.1 มีเฉลย (Ground Truth): ทุก session มีป้ายกำกับความจริงในไฟล์ `data/sample/labels.csv` ว่าผิดปกติจริง (1) หรือปกติ (0) — เราเอาคำทำนายของ AI ไปเทียบกับเฉลยนี้

3.2 ทดสอบบนข้อมูลที่โมเดลไม่เคยเห็น: แบ่งข้อมูลเป็น Train 1260 sessions (ใช้สอน) และ Test 540 sessions (ซ่อนไว้ ใช้วัดผล) — การวัดบนข้อมูลใหม่ที่โมเดลไม่เคยเจอ ป้องกันการ "ดูข้อสอบก่อนสอบ" ทำให้ผลสะท้อนความสามารถจริง

3.3 นับเป็นตาราง Confusion Matrix แล้วคำนวณด้วยสูตรตายตัว:

	AI ทำนาย: ผิดปกติ	AI ทำนาย: ปกติ
จริง: ผิดปกติ	**74** TP (จับถูก)	**24** FN (พลาด)
จริง: ปกติ	**15** FP (เตือนผิด)	**427** TN (ปกติถูก)

3.4 สูตร + ตรวจด้วยมือ:

Precision = TP / (TP+FP) = 74/89 = **0.831**

Recall = TP / (TP+FN) = 74/98 = **0.755**

F1 = 2×P×R / (P+R) = **0.791**

Accuracy = (TP+TN) / รวม = 501/540 = **0.928**

ใครก็คำนวณตามนี้แล้วได้เลขเดียวกัน — ไม่มีพื้นที่ให้ปรับแต่งตัวเลข

4. ตรวจสอบซ้ำได้ด้วยตัวเอง (Reproducibility)

พิสูจน์ได้ 3 ชั้น — ใครรันก็ได้ผลเดียวกัน (เพราะ fix random seed):

1. **รันการวัดผล** → เห็นตาราง PASS/FAIL ทุกตัวชี้วัด
2. **ตรวจเลขด้วยมือ** จาก TP/FP/FN/TN ตามสูตรข้อ 3.4 → ต้องตรงกัน
3. **รันซ้ำกี่ครั้งก็ได้ผลเดิม** → ยืนยันว่าไม่ใช่ฟลุ๊ค

```
# ชั้น 1 + 3: รันวัดผล (ได้เลขเดิมทุกครั้ง)
python scripts/evaluate.py --dataset sample

# โค้ดคำนวณ metric + เทสต์ตรวจสูตร
src/loganalysis/metrics/evaluation.py   tests/test_evaluation.py
```

5. ข้อจำกัด และการทำให้แข็งแกร่งขึ้น

ตัวเลขชุดนี้พิสูจน์บน **ข้อมูลจำลอง (synthetic)** ที่ออกแบบให้สมจริง — จึงพิสูจน์ว่า **ระบบและโมเดลทำงานถูกต้องตาม logic** แต่ยังไม่ใช่ข้อมูล production จริง

เพื่อยกระดับความน่าเชื่อถือสู่การใช้งานจริง ควรทำเพิ่ม:

1. รันบน **ชุดข้อมูลจริง** (HDFS_v1 จาก LogHub — โค้ดรองรับด้วย `--dataset hdfs`)
2. ทำ **k-fold cross-validation** (แบ่งหลายรอบแล้วเฉลี่ย) แทนการแบ่งครั้งเดียว
3. ทดสอบกับ log จริงของระบบที่จะนำไปใช้ และติดตามผลหลัง deploy อย่างต่อเนื่อง

6. สรุป

"เอาคำทำนายของ AI บนข้อมูลที่ไม่เคยเห็น ไปเทียบกับเฉลย แล้วนับด้วยสูตรมาตรฐาน (Precision / Recall / F1) — รันซ้ำได้ผลเดิมทุกครั้ง และตรวจเลขย้อนกลับด้วยมือได้" นี่คือเหตุผลที่ผลลัพธ์ ALL PASS เชื่อถือได้ ไม่ใช่การกำหนดตัวเลขเอง