สรุปผลทูน AI ถอดแบบก่อสร้าง — t01 → t02 → t03 → t04/t44 → t05 (Destrier)

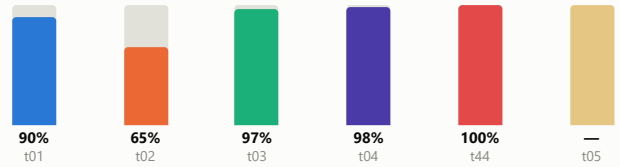
รวบรวมเต็มจาก workmen's diary + t01/t02/t03_workflow.md + t04_Purson/t05_Destrier docs · โมเดลฐาน Qwen3.6-35B-A3B (MoE) ทุกรอบยกเว้น t04/t44 (InternVL3-78B)

t01	t02	t03	t04 Purson	t44 Voldemort	t05 Destrier
21-24 ก.ค.	28 ก.ค.-21 ส.ค.	24-25 ส.ค.	26-29 ส.ค.	29-30 ส.ค.	30 ส.ค.-1 ก.ย.

Element recall (id ตรงกับเฉลย) — ยิ่งสูงยิ่งดี

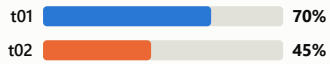
t05 วัดแบบให้อภัยการรวม/แยก element (mark_base) — เทียบใกล้เคียง t01-t03 ที่สุด (strict เข้มสุด = 11.8%)

JSON valid — parse ได้ไหม



JSON สวยไม่ได้แปลว่าอ่านถูก — t04/t44 ได้ JSON valid สูงสุดแต่ recall 0% (บทเรียนแพงที่สุดของโปรเจกต์)

View exact match

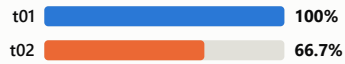


วัดเฉพาะ t01/t02 (20 ตัวอย่างเดียวกัน)

element_type exact

ในบรรดา element ที่หาเจอแล้ว ตอบ type ถูกกี่ %

count exact



นับจำนวนตัวถูกไหม (เช่น "4-DB12")

Over-predict — element มโนเพิ่ม ยิ่งน้อยยิ่งดี

วัดเฉพาะ t01/t02 บน 20 ตัวอย่างเดียวกัน — t03/t04/t05 เปลี่ยนวิธีวัดเป็น per-subtask ไม่บันทึกตัวเลขนี้ตรงๆ

ตัวเลขดิบเทียบ t01 vs t02 (ทุก metric ที่วัดคู่กัน)

ตัวชี้วัด	T01	T02	ชนะ
JSON valid	90%	65.0%	t01
view ตรงเป๊ะ	70%	45.0%	t01
element recall	28.2%	20.6%	t01
element_id exact	100%	100%	เท่ากัน
element_type exact	87.5%	74.3%	t01
count exact	100%	66.7%	t01
over-predict	78	70	ใกล้เคียง

t03 — id-recall แยกตาม subtask (บ้าน08, 33/33 งาน, 2026-08-25)

SUBTASK	N	JSON VALID	PRED EL / GT EL	ID-RECALL แต่ละงาน
gridline	1	1/1	0/0	— (schema ไม่มี element_id)
notes	4	4/4	12/0	— (GT ใช้ sections[] แทน)
schedule	5	5/5	11/32	43%, 29%, 0%, 0%, 0%
section	13	13/13	34/23	0% ×8, บาง n/a
plan_footing	4	4/4	3/6	0%, 100%, 0%, 0%
plan_beam	3	3/3	2/54	**0%, 0%, 0%**
plan_slab	2	1/2	1/6	0%, 50%

รวม id-recall เฉลี่ย **10.1%** (22 งานที่ GT มี id ให้วัด) — จุดอ่อนหลักคือ **plan_beam อ่านคานทั้งหน้าไม่ออกเลย** (2/54) ตรงกับที่ overlay ภาพยืนยันไว้ (t02 หน้า26 ยุบคาน 45 ตัวเหลือ 2 zone notation)

ไทม์ไลน์บั๊ก/เหตุการณ์สำคัญที่เจอระหว่างทาง (จาก diary)

21 ก.ค.	**t01 DAY OF SHAME** — เสีย output ทั้งชุดเพราะไม่ backup ก่อนทำลาย instance เช่า
24 ก.ค.	**t01 chain-of-thought bug** — โมเดลพิมพ์ "เหตุผล" ยาวก่อน JSON กิน token budget จนดูเหมือนค้าง, recall วัดได้ 0/3 ก่อนแก้ enable_thinking
28 ก.ค.	**t02 image-token bug** — t01 เทรนด้วยภาพย่อเหลือ ~266 token/ภาพ (ผิดปกติ, fallback 512px) ทำให้ผล "ดูดี" แต่เทียบกับรอบอื่นไม่แฟร์เลย
21 ส.ค.	**t02 trailing-comma bug** — lm-format-enforcer ปล่อย trailing comma หลุด JSON parse, แก้ด้วย regex ก่อนเจอ xgrammar ที่ผลิต comma แบบนี้ไม่ได้เลยตั้งแต่ต้นทาง
21-22 ส.ค.	**xgrammar พิสูจน์บน GPU จริง** — JSON valid 30/31 หน้า (96.8%) เทียบ 57.9% แบบไม่มี grammar
26-29 ส.ค.	**t04 Purson พัง** — ลืมเปิด `crop_to_patches`, รูปย่อเหลือ 256 token/รูป → recall 0% ทั้งที่ JSON valid 96-100% (บั๊กคลาสเดียวกันครั้งที่ 3: t01→t02→t04)
29-30 ส.ค.	**t44 Voldemort พัง** — บีบอัด 4-bit ทำลาย ViT (ตา), ตอบเห็น Barbie/BB-8 ในแบบก่อสร้าง; ประตูตรวจ $2 ก่อนเทรนเต็มจับได้ (เซฟ $60-70), แต่รอบแรกเกือบโดนหลอกเพราะ prompt ทดสอบใส่คำตอบเฉลยปนไปเอง
30-31 ส.ค.	**t05 Destrier** กลับไปใช้สาย Qwen3.6 ที่พิสูจน์แล้ว (t01/t03) แทน InternVL3, รวม 4-fold soup, dataset 4 pass (pass0+pass1+2.4+3)
1 ก.ย.	Destrier ต่อ pass1/1.5/2.5/3 เข้า worker.py จริงครั้งแรก, พิสูจน์ pass3 กับข้อมูลจริง 191/191, จ่าลองรันบ้านทั้งหลังไม่ใช้ GPU เจอบั๊ก Constistant 9 ตัว แก้หมด
2 ก.ย.	รันบ้าน09 ผ่านระบบ pass+Destrier บนการ์ดจอเช่าจริงครั้งแรก, เพิ่ม checkpoint ทุก pass + ปุ่ม "ทำต่อ" กันงานพังกลางทาง

ตายระหว่างทาง (InternVL3-78B — คนละสถาปัตยกรรม)

☠ t04 Purson
ลืมเปิด crop_to_patches — รูปย่อเหลือ 256 token/รูป → recall 0%

☠ t44 Voldemort
บีบอัด 4-bit ทำลาย ViT — เห็น Barbie/BB-8 ในแบบก่อสร้าง

สิ่งที่เปลี่ยนแต่ละรอบ (สรุปย่อ)

t01	ภาพย่อผิดพลาด (บั๊ก) — ผลดีเทียม
t02	ภาพจริง — คานทั้งหน้า recall 0%, เจอ+แก้ trailing comma
t03	เพิ่ม pass0/2.4/3, กริดมาสเตอร์ฝังใน prompt
t04	เปลี่ยนโมเดล InternVL3-78B — ลืม config รูป, ตาย

t44	4-bit quant ตัวเดียวกัน — บีบตาโมเดลจนบอด, ตาย
t05	กลับ Qwen3.6, รวม 4-fold soup, รันจริงบน GPU แล้ว (2 ก.ย.)

สรุปจาก workmen's_diary (ก.ค.-ก.ย. 2026) + t01/t02/t03_workflow.md + t04_Purson/t05_Destrier docs · เกณฑ์วัดเดียวกันข้ามรอบ แต่ไม่ควบคุมตัวแปรเต็มที่ (คนละหน้า/prompt/decoding stack)