CryptoLens — Week 2

ครบ Track 1–3 แล้ว — AI จริง (ไม่ mockup), มี eval + fallback + cost metrics, design system ใช้ต่อได้, production คิดไว้แต่ยังไม่ deploy

Track 1 — AI Implementation (4 ข้อ)

1 · Technical implementation คืออะไร

เลือก **5 technique ผสม** เพราะ product ขาย trust ให้ retail trader ที่กังวล ไม่ใช่ signal bot:

Technique	ใช้ทำอะไร
System prompt design	summarize_coin, market_mood, daily_digest, ask_coin — grounded generation จาก snapshot จริง ห้าม hallucinate / สั่งซื้อขาย
Guardrails	prompt rules + _clean_ask_response() + post-parse digest แทนที่ verdict ถ้า LLM ขัด market direction
Deterministic metrics	Heat, Deviation (z-score 30 วัน), facts ใน heat.py — LLM ไม่แตะเลข
Portfolio triage	triage.py — hero verdict จาก |z| × weight + macro mode + false-alarm filter
Custom benchmark	grounding eval + deviation backtest

ไม่ใช้: RAG, Agent, OCR, fine-tune — ไม่มี KB / เอกสาร user, ตอบจาก market data สด

Provider chain: gemini-2.5-flash flash-lite 3.1-flash-lite Groq llama-3.3-70b + rate_limit.py คุม RPM

Flow:

```
holdings -> Binance/news -> Heat+Deviation (deterministic) -> triage verdict
    -> LLM narrative (grounded) -> DailyDigest / Chat / Fact cards
```

2 · รู้ได้ยังไงว่ามันดี

Grounding eval (eval/grounding_eval.py) — วัด % ผ่าน 3 checks:

- เลข 100 trace กลับ snapshot ได้
- ไม่มีคำสั่งซื้อ / ขาย (TH + EN)
- Digest มี disclaimer

```
python -m eval.grounding_eval          # mock 100% (5/5) - runs in CI, no key
python -m eval.grounding_eval --live   # live: must run yourself before present
```

Deviation backtest (eval/backtest_deviation.py) — วัด signal deterministic (ไม่เกี่ยว LLM): calm-day accuracy, spike recall 85%, thresholds mild_z = 1.75, abnormal_z = 1.9

3 · ถ้า AI พัง — fallback

ชั้น	พฤติกรรม
Provider chain	quota / 503 / timeout provider ถัดไป
Stale cache	คืน cache เก่า + stale: true
Digest AI ตายทั้งหมด	triage.portfolio_verdict() + degraded: true — ไม่ cache digest พัง
Digest parse ผิด	ใช้ fb_verdict / fb_narrative
Frontend	inferVerdictFromCoins() + error state ใน DailyDigest

ชั้น	พฤติกรรม
Deterministic รอด	Heat / Deviation / facts ยังขึ้นครบโดยไม่พึ่ง LLM

ช่องว่างที่รู้: /api/ask provider หมด 502 + ChatPanel “ลองอีกครั้ง” (ยังไม่มีคำตอบ deterministic)

4 · Cost & latency

metrics.py + GET /api/metrics วัดทุก call แยก op

Metric (daily_digest)	ค่า
Tokens / call	~1,563
Latency	~2.9s
Cost / call (paid projection)	~$0.00023
Cost จริงวันนี้	$0 (free tier)

คุม cost: digest cache ต่อ UTC ชม. · summarize_coin lazy (ไม่ generate ตอน insights load) · max_tokens cap · WATCHLIST_MAX = 10

ระหว่างรอ: skeleton / spinner · digest จาก cache ขึ้นทันที · ยังไม่มี token streaming

Unit economics: user ~30–60 call/เดือน $0.006–0.012/user — driver จริงตอน scale = hosting + data API มากกว่า LLM

Track 2 — Design System

มี system ที่ AI generate UI ต่อได้แล้ว:

- **Tokens:** frontend/app/globals.css @theme — 11 color tokens + Sora / IBM Plex Sans Thai + JetBrains Mono
- **กติกา:** frontend/design.md — iron rules (heat-* สงวน HeatBar, ทิศทางราคา mint/coral/muted เท่านั้น)
- **Feed AI:** AGENTS.md, frontend/AGENTS.md, CLAUDE.md
- **Components ใช้ system แล้ว:** HeatBar, DeviationBadge, CoinFactCard, DailyDigest, HoldingsPanel

ตัวอย่าง: DeviationBadge ใช้ coral/warn/muted ไม่แตะ heat-* · HeatBar เป็นเจ้าของ ramp เดียว

Track 3 — Production (คิดไว้ · ยังไม่ทำ)

หัวข้อ	ปัจจุบัน	ขั้นถัดไป
Deployment	local dev	FE Vercel · BE Render / Railway
Scaling 100–1k	in-memory cache, sync AI, SQLite	Redis · async queue · Postgres
Reliability	provider chain + stale + digest degraded	health check UI · ask fallback
Database	SQLite 7 ตาราง	Postgres + index (user_id, bucket)
Privacy	trade key AES-GCM (vault.py)	บังคับ CRYPTOLENS_MASTER_KEY ใน prod

ไม่มี OCR/RAG บน PII surface จำกัด (watchlist / holdings / trade key)

Journal — สิ่งที่ mockup ไม่เห็น

- AI คืน Heat ไม่ได้ ดึงเป็น pure function ใน heat.py
- RPM free tier พังเร็ว ยก summarize_coin ออกจาก insights load
- Digest false alarm เยอะ แยก triage.py + filter |z| × weight 40
- LLM ตาย digest ยังใช้ได้ deterministic verdict ไม่ใช่จอขาว
- Cost จริงต่ำกว่าที่กลัว

สถานะก่อน present

พร้อมแล้ว	ต้องทำก่อน present
AI จริง + 4 ข้อ Track 1	รัน grounding_eval --live ได้ % จริง
Design system	sync design.md §6 กับโค้ด
Deviation backtest ผ่าน	/api/ask fallback (optional แต่ควรพูดว่ารู้ gap)
Digest degraded fallback	token streaming (nice-to-have)

Action: ถ้าจะ present วันนี้ — รัน `python -m eval.grounding_eval --live` ก่อน แล้วใส่ % ใน slide; เตรียม demo scenario “ปิด API key digest ยังขึ้นด้วย triage fallback”