Run #110 — does reasoning_effort improve arena performance? (gpt-5.6-luna)

Date: 2026-08-17 → 2026-08-18 (single detached process, ~13.5 h) **Design:** predeclared in `2026-08-17-luna-reasoning-effort-plan.md` before launch. One model (`gpt-5-6-luna`, zenmux), four pinned arms (`none / low / high / max`), all 5 golden workflows as paired blocks, 2 trials per cell → 40 trials. **Board health: 20/20 arms scored, 0 invalid, 0 infra deaths** — every predeclared validity rule held; no cell was excluded.

Headline answer

No — "higher effort ⇒ better agent" is not what the data shows. For luna on this arena, reasoning_effort behaves as two different dials at once:

1. **Zero → some reasoning is the real jump.** Every reasoning arm beats `none` by ~5 points mean objective (`none` 88.9 vs `low` 93.8 / `high` 93.3 / `max` 94.8). And `low` achieves that while being *cheaper than none* (29.1 calls, 10.2 min/trial vs 30.8 calls, 11.1 min/trial): a little thinking *reduces* total work by cutting flailing.
2. **Above `low`, quality is flat but cost explodes.** `max` buys +1.0 objective point over `low` for **3.3× wall-clock** (33.8 vs 10.2 min/trial) and **+50% tool calls** (43.9 vs 29.1). `high` buys nothing over `low` (93.3 vs 93.8). Within-ladder differences are inside per-cell trial noise (stdev 1–4); the cost differences are not.

By mean OVR (which prices efficiency in), **`low` is the best operating point: 86.9**, vs 81.8 (`none`), 81.6 (`high`), 82.7 (`max`).

The board

workflow	none	low	high	max
risk–manager–control–day (flagship)	**94.8**	93.6	91.1	91.0
high–board–portfolio–review–day	71.4	84.3	78.6	**90.0**
trader–rfq–booking–day	94.5	93.7	**96.8**	**96.8**
risk–limit–breach–day	97.4	97.3	**100.0**	96.1
ops–settlement–day	86.3	**100.0**	**100.0**	**100.0**
mean objective	88.9	93.8	93.3	**94.8**
mean OVR (card)	81.8	**86.9**	81.6	82.7
mean calls/trial	30.8	**29.1**	38.6	43.9
mean wall min/trial	11.1	**10.2**	14.2	33.8
mean trial stdev	3.98	**1.06**	2.20	1.72

Effort's value depends on the workflow's shape

- **Procedural, well-routed loops (flagship): effort *hurts*.** Monotone decline 94.8 → 91.0. At `max`, luna skipped reading the `read-risk-result` skill doc entirely (0/2 vs 2/2 at `none`) — more internal reasoning substituted for following the documented procedure. Calls balloon (29 at `low` → 47–62), EFF collapses (67 → 1.5–17).
- **Judgment/selection-heavy (high-board): effort *pays*, a lot.** +18.6 points `none` → `max` (71.4 → 90.0). The `none` arm failed exactly the inference work: step-6 retrieval of the prior governed valuation (0/2), the final governance artifact (1/2), the required tool ordering (0/2).
- **Numeric solving (trader-rfq): only high+ clears the hard check.** The step-1 achieved-price-within-tolerance check went 0/2 at `none` and `low`, 2/2 at `high` and `max` — a genuine reasoning win no amount of procedure-following buys.
- **Near-ceiling ops workflows: `low` is already enough.** ops-settlement is 100.0 at every reasoning level; `none` dropped one trial to 72.7 by skipping settlement reads and fumbling the reopen-refusal answer (diligence failures, not capability ones).

Consistency and tails

- `none` is the least consistent arm (mean trial stdev 3.98; ops-settlement trials split 72.7 / 100.0). Adaptive zero-reasoning makes performance a coin flip on judgment moments.
- `max` has extreme latency tails: one high-board trial ran **107 minutes** (median max trial ≈ 24 min).
- A parallel probe study on this same run found trap/prohibition compliance **8/8 at all four efforts** — effort moves exploration volume, never policy.

Mechanism (check-level tally)

Losses at `none` are *diligence* failures — steps skipped, answers given without fetching evidence. Losses at `high` / `max` are *over-execution* failures — extra exploration displacing the documented procedure (skill-doc skips, occasional grounding slips at 2× par call volume), plus the EFF collapse the golf curve is designed to price. Correctness axes (GRD/ADH/SYN) barely move above `low`; the call count moves ~50%.

Replicates the cross-model pattern: the "Grok inversion" (EFF crash at high effort) is not grok-specific — luna shows the same shape once effort is pinned, inverting the #107/#108 deepseek observation (fewer calls at high) — the direction of the effort→calls relationship is model-family-specific.

Limitations (predeclared)

- n = 2 trials × 5 blocks per arm; within-ladder ordering (low vs high vs max) is not resolved — only `none` vs `any` is outside noise.
- Arms ran sequentially (same process), so arm order correlates with time-of-day; per-trial wall times show no drift step, and 0 infra deaths.
- Token spend is not recorded in transcripts; wall-clock and call counts are the cost proxies.
- One model, one provider route. The deepseek contrast above says do not generalize the calls-vs-effort direction across families.

Desk recommendation

Run luna at **`low`** for routine desk work (best OVR, cheapest, most consistent); escalate to **`high`** / **`max`** **only for judgment-heavy or solve-type tasks** (high-board-style reviews, price solving), accepting 2–3× wall-clock. Never run `none` for anything with judgment moments: the mean looks acceptable but the variance is where it bites.

Artifacts

- Run: arena run **#110** (matches + per-trial transcripts under `artifacts/arena/110/`).
- Analysis script: session scratchpad `analyze_run110.py` (read-only over the live DB; board, card means, paired deltas, per-check tally).
- Plan (predeclared): `docs/arena/2026-08-17-luna-reasoning-effort-plan.md`.
