Does reasoning_effort improve arena performance? — predeclared plan

Date: 2026-08-17 (declared BEFORE launch) **Question:** Does raising `reasoning_effort` make an LLM perform better on the desk's golden-workflow arena — and is the relationship monotonic?

Design

- **Model (fixed):** `gpt-5-6-luna` (`openai/gpt-5.6-luna`, zenmux route). Measured ladder (2026-08-14 probe): `none, low, medium, high, xhigh, max`; `minimal` rejected.
- **Arms (4):** `none`, `low`, `high`, `max` — both ends of the ladder plus the common contrast pair. One contestant per arm (`(model_id, reasoning_effort)` identity, migration 0058).
- **Workflows (5, all current):** `risk-manager-control-day`, `high-board-portfolio-review-day`, `trader-rfq-booking-day`, `risk-limit-breach-day`, `ops-settlement-day`. Each workflow is a paired block: every arm faces the same fixtures, truth values, and scorer.
- **Trials:** 2 per (workflow, arm) → 40 trials, folded per-cell by `fold_trial_breakdowns`. Run #107 vs #108 showed a 6.6-point swing at identical config, so single trials cannot support a conclusion.
- **Jury:** off (default). Objective-only scoring — deterministic.
- **Route sanity (pre-launch, measured):** trivial prompt at `high` spends 0 reasoning tokens (adaptive); a snowball-payoff question spends 0 at `none` (answer muddled) vs 1,034 at `high` (answer correct). The knob is live on this route.

Predeclared metrics

1. **Primary:** objective score (%) per (workflow, arm) cell, compared across arms **paired by workflow**. Hypothesis under test: score is non-decreasing in effort.
2. **Secondary:** ability-card axes (GRD/ADH/SYN/PRC/EFF, OVR), tool calls per trial, per-cell trial stdev (consistency), error counts, wall-clock per trial (match timestamps + transcript timing), reasoning/completion token spend where recorded.
3. **Mechanism:** per-check pass-rate tally across arms (the validity-audit instrument) to locate WHERE effort helps or hurts (grounding vs adherence vs synthesis vs procedural vs efficiency).

Predeclared validity rules

- Trials that die infra (`invalid` , `_is_infra_blank` / infra-contaminated) are excluded; a cell with 0 clean trials is reported **missing, never imputed**.
- If more than ~20% of trials are infra-dead, stop and `--resume` rather than conclude from a swept board.
- This board is pinned-effort and therefore **not comparable to historical boards** (#8–#104) on EFF/CON; all comparisons are within-board.
- Arms run sequentially in one process, so arm order is confounded with time-of-day/provider load; we report per-trial timing so a drift would be visible, and we do not interpret sub-point score differences.
- n = 5 workflow blocks × 2 trials. We report signs and means of paired differences; no significance theater beyond what n supports.

Cost accounting (predeclared as a first-class result)

Higher effort that buys +1 point at 3× wall-clock and token cost is a finding, not a footnote. Cost per arm is reported beside score.
