--- title: Round 6 of the skills optimization study: re-measuring the six noisiest cells project: portfolio date: 2026-06-01 kind: post type: research ---

Round 6 of the skills optimization study: re-measuring the six noisiest cells

Rounds 1-5 of my skills optimization study left six cells I did not trust: two script-backed auditor skills (`skills-quality` and `skills-freshness`) crossed with `opus`, `sonnet` and `haiku`. Each was a single draw per arm, and one cell (skills-quality on opus) read +5% in round 1 and -62% in round 5. Round 6 re-measured all six at depth: five draws per arm on both opus cells, three per arm on the four sonnet and haiku cells, arm A (cold) against arm B (with-skill), model held constant. All six came out positive: 54% to 85% saved on the cell medians. The volume-weighted aggregate is 75% saved: 748,768 tokens summed across the cold-arm medians against 188,474 with the skill, a net 560,294 saved over 6 cells. The two opus cells overturn their earlier verdicts outright, and in both the reversal is attributed to the cold arm having under-worked at N=1, not to the skill getting better; a third cell, skills-quality on sonnet, also overturns an earlier sign-flip.

The round ran as a fallback, which the record states plainly: "Triggered as the optim-rollout fallback: the audit/fix queue found 0 fixes to apply, so remaining quota went to firming up the noisiest existing cells." There is no markdown write-up of round 6. Two JSON files (`skills-optim-study-2026-06-01-replicates.json` and its `.input.json`) plus the PDF you can pull with `download --replicates` from the contact terminal at mikkonumminen.dev are the entire record. The skills themselves live in `github.com/MikkoNumminen/claude-skills`, and the source names them without the `mikko-` prefix they carry once installed.

Method

Arm A is cold: no `skills-quality` / `freshness` SKILL.md or script, but it *does* inspect the target skills, since auditing them is the task. Arm B reads the SKILL.md and runs its script read-only, no `--update`. The before-arm is "pinned by averaging >=5 cold draws (not a single N=1 draw)". Each cell reports a median plus spread, with the ratio computed on medians and again on the pinned mean. The aggregate is a "ratio of sum-of-per-cell-medians (volume-weighted), NOT mean-of-cell-percentages". Token accounting is input + output + `cache_creation_input_tokens`, deduped by `(sessionId, requestId)`, cache_read excluded: the same accounting as the round-1-5 study.

One departure from earlier rounds: `worktree_isolation` is recorded as "skipped - both arms read-only against ~/.claude/skills/".

The six cells

cell	model	N/arm	A median	A sd	B median	B sd	saved (median)	% (median)	% (pinned mean)
`skills-quality`	opus	5	125,778	5,019	30,015	2,148	95,763	76	76
`skills-freshness`	opus	5	191,519	10,675	45,913	5,417	145,606	76	76
`skills-quality`	sonnet	3	94,849	4,034	13,988	974	80,861	85	86
`skills-freshness`	sonnet	3	146,622	4,634	32,872	3,075	113,750	78	77
`skills-quality`	haiku	3	74,612	27,062	34,511	8,878	40,101	54	58
`skills-freshness`	haiku	3	115,388	19,000	31,175	5,928	84,213	73	71

Both percentages are published because they disagree in four of the six cells. The headline figures the verdicts quote are the medians (+76/+76/+85/+54/+78/+73); the pinned-mean column is the same measurement with the before-arm averaged instead, and it moves `skills-quality|haiku` from 54 to 58 and `skills-freshness|haiku` from 73 to 71. Neither is more true than the other, so both stay on the page.

What got overturned, and why

The two opus cells are the starkest reversals. `skills-quality|opus` had read +5% in round 1 and **-62%** in round 5; the round-5 note names the cause: "N=1 anomaly: cold arm short-circuited at 31K". With the cold arm actually doing the full audit (116-131K) and the script-backed skill running it "in ~0 LLM tokens (~30K total)", the verdict is "the save is large and tight... both arms stable, no N=1 fragility". `skills-freshness|opus` had read -21% / +1% / -1% across rounds 1, 2 and 5. "All near-zero/negative at N=1; R1 cold arm was shallow (135K/9 turns)". At depth its cold arm spends 172-200K over 16-31 turns verifying `file:line` citations against source, while the sha256 change-detection script short-circuits unchanged skills at ~46K.

`skills-freshness|haiku` is the cell the study calls "THE headline swing cell". Its round-1 -70% "was a pre-fix N=1 artifact (unusually cheap cold arm, 64K)"; post-fix it went +20% (R2) → +48% (R5) → +73% here, and the verdict elevates it to the one study-level claim in the file: "Confirms the central study claim: bounded script-backed guidance reliably saves haiku-class tokens."

Only `skills-quality|haiku` is described as "Consistent" rather than an overturn. It "matches R5 (+54%) closely", i.e. it was stable before this round. Its pattern gets its own name: "haiku scouts exhaustively when cold (54-118K), benefits most from a directed script."

The mechanism is identical in both skills, on every model. Cold, the audit runs in the LLM: reading all 14 SKILL.md (93-102K on sonnet), or verifying citations turn by turn. The sonnet freshness cold arm takes up to 109 turns. With the skill, the audit is handed to a deterministic script and the model reads the result.

What this round does not say

The file is explicit that its percentages have limits, and they belong next to the numbers:

- **The cold arm is the soft part of the ruler.** "cold-arm token cost is task-framing-sensitive (how thoroughly it audits); trust direction + magnitude, not the exact %." A bare "+76%" without that sentence overstates what was measured.
- **The depth is uneven.** N=5/arm on the two opus cells only; all four sonnet and haiku cells are N=3, and every non-opus verdict says so. The stated goal was "N>=5/arm on opus". Nothing more.
- **The haiku cells are still the noisy ones.** `skills-quality|haiku` has a cold-arm sd of 27,062 on a median of 74,612, with draws from 53,635 to 118,581: a better-than-2x range. `skills-freshness|haiku` has sd 19,000 on a median of 115,388. Only the opus cells are called stable.
- **This is not a fresh sample.** These six cells were selected *because* they were the noisiest. The 75% aggregate covers those six re-measured cells alone and is not comparable to the Spacepotatis aggregate or the per-model suite table in my summary post. It does not restate them.
- **The noise floor here is the file's own.** Round 6 records "cells within ~2% carry direction only", which is far tighter than the ±20% floor I apply to the calibration tables elsewhere in this line. The JSON does not reconcile the two, and I am not going to reconcile it retroactively.

The direction of every correction in this round is the same: the earlier cold arms had under-worked. That is a finding about the instrument, not about the skills.