SILICA-BENCH: EVALUATING LONG-HORIZON LLM AGENTS THROUGH VERIFIED RTL-TO-GDS IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon LLM agents must turn a sequence of decisions into a reliable result: interpret imperfect feedback, preserve requirements, and divide a finite budget between exploration and validation. We introduce SILICA-BENCH, which evaluates these abilities through physical improvement of digital circuits with a commercial RTL-to-GDS flow. Four properties make this a useful agent task: an open-ended, continuously graded power, performance, and area (PPA) objective with a fixed behavioral target; hard functional requirements; costly staged evidence over coupled RTL and flow choices; and 36 designs across eight domains and multiple scales. Agents receive reference RTL and a reference flow, work within a declared budget, and register what they deliver. Protected assessment binds RTL and post-route netlist replay to the nominated implementation, using local reconstruction or identity-matched reuse of its physical outputs. The protocol separates validity from improvement, early evidence from final quality, and discovery from delivery. A paired study of six model–harness systems and three edit authorities covers 240 nominations on 30 designs. Every configuration achieves at least 93% validity, yet no Joint system delivers a changed, verified gain above 5% on more than one third of tasks, and two Joint systems fall below unit failure-inclusive quality. Eleven of twelve valid regressions larger than 5% use the local reconstruction path. These results distinguish successful execution from reliable improvement, and make delivery of a verifiable engineering artifact an observable target for long-horizon agent evaluation.

1 INTRODUCTION

LLM agents increasingly solve tasks through repeated interaction with tools and files. Their usefulness depends on the decisions made along the way: which hypothesis to test, which result to trust, and when to stop exploring. Software evaluations already show why time and interaction matter. METR measures task horizons using expert human completion time, while RE-Bench compares systems under 2-, 8-, and 32-hour total budgets (Kwa et al., 2025; Wijk et al., 2025). These studies motivate evaluation beyond a single generated answer. An agent must use an evolving history of feedback, keep requirements intact, and reserve enough time to check and retain its work. A useful benchmark should make these choices affect both the delivered result and the resources spent obtaining it.

Digital physical implementation is a consequential setting for this question. Electronic design automation (EDA) transforms register-transfer-level (RTL) code through synthesis, placement, clock-tree synthesis (CTS), and routing into a physical layout.[1] The power, performance, and area (PPA) of an implementation determine the cost of performing its computation. Four properties make this environment useful for evaluating agents. **P1: open-ended objective, fixed target.** Many implementations preserve the same behavior, so there is no canonical answer to imitate. **P2: hard requirements with graded quality.** A smaller implementation is unacceptable if it changes required behavior. **P3: staged evidence over coupled choices.** RTL structure and flow settings jointly af-

[1] "RTL-to-GDS" denotes synthesis through routing and post-route analysis here; acceptance does not imply physical signoff (Appendix C).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

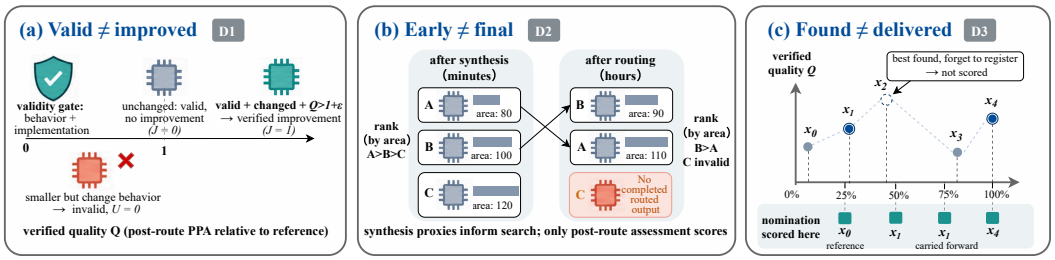

Figure 1: Three measurement distinctions (schematic). (a) Validity alone does not establish improvement: $J = 1$ also requires a changed candidate and $Q > 1 + \epsilon$. (b) Lower area is preferred ($>$ denotes preference); A/B reverse order, while C lacks a completed routed result. Values and time labels are illustrative. (c) Only completed registrations count; the shown x_1 registers after 25%, so x_0 remains the first nomination. Quality heights are retrospective illustrations, not feedback returned to the agent.

fect PPA, and measured gains depend on tools and evaluation stages (Lu et al., 2026; Fang et al., 2026). An early estimate therefore need not determine the final ranking. In our 36-design reference set, full-flow time ranges from 12.1 minutes to 7.67 hours, making evidence acquisition a meaningful part of the budget. **P4: diverse tasks at multiple scales.** Eight functional domains and a wide range of circuit sizes support comparisons beyond a single design. Together, these properties connect agent decisions to observable consequences: a structural hypothesis changes the circuit, an evidence request spends time, and a delivery decision determines which change is evaluated. The same physical objective gives these decisions a practical purpose: reducing the resources needed to carry out a fixed computation.

Executable software benchmarks and recent hardware studies already evaluate artifact changes, tool use, and physical optimization (Jimenez et al., 2024; Arnold et al., 2026; Liu et al., 2026; Zhou et al., 2026). The remaining measurement problem is to distinguish three outcomes (Figure 1). **D1: valid is not improved.** Keeping the reference may preserve behavior without reducing its cost. **D2: early is not final.** A favorable synthesis report need not survive post-route assessment. **D3: found is not delivered.** A useful candidate in the search history does not show that the agent retained and nominated it in time. Correctness, quality among successes, and a retrospective best candidate answer different questions; none alone describes reliable improvement over a whole run.

Our central idea is to fix the behavior to preserve and the reference cost to improve upon, while making the agent responsible for its implementation, its use of public evidence, and its nomination. SILICA-BENCH realizes this in a frozen commercial flow. The agent receives reference RTL, a callable flow, and a visible budget; it changes RTL or legal flow settings and registers an exact snapshot. Protected replay of both RTL and the corresponding post-route netlist, together with artifact-matched physical evidence, provides the basis for final quality. Only the latest completed registration is scored, preventing the evaluator from choosing a better artifact in hindsight. Matched tasks, budgets, harness, and evaluator allow differences to be studied at the system level. This design links open engineering search to a common test of what the agent actually delivers.

We make four contributions:

1. **A diverse engineering environment.** We organize 36 circuits from eight domains into reference-preserving RTL and flow improvement tasks, with measured physical references and time budgets (P1–P4).
2. **A verification-gated nomination protocol.** Protected assessment separates validity, continuous quality, and verified improvement; registered candidates determine what is scored (D1–D3; Algorithm 1).
3. **A controlled study of engineering freedom.** Six systems use Joint authority, and one anchor also uses RTL-only and Flow-only authority. Matched tasks, budgets, harness, and assessment yield 240 paired nominations on 30 designs (Section 6).
4. **Findings on reliable delivery.** High validity coexists with limited improvement: Joint EVQ ranges from 0.927 to 1.069, and at most 10/30 nominations per Joint system improve

by more than 5%. Added authority has mixed benefits, and large regressions concentrate in locally reconstructed nominations. These distinctions expose outcomes that tool activity or success-only quality can hide.

2 Related Work

Long-horizon agent evaluation. SWE-bench evaluates repository changes with tests, and SWE-agent studies how the interface affects performance (Jimenez et al., 2024; Yang et al., 2024). RE-Bench, PaperBench, and METR extend evaluation to research engineering, research replication, and tasks of different human duration (Wijk et al., 2025; Starace et al., 2025; Kwa et al., 2025). Cost-aware evaluation motivates reporting resources as well as outcomes (Kapoor et al., 2024). SILICA studies reference-preserving improvement with costly physical evidence and timed nominations.

RTL and implementation optimization. VerilogEval evaluates RTL generation, while RTL-Rewriter and RTL optimization benchmarks study changes to existing circuits (Liu et al., 2023; Yao et al., 2024; Lu et al., 2026). Dr. RTL uses tool feedback; RTLScout jointly explores RTL and synthesis choices; ARES relates PPA to model cost (Fang et al., 2026; Arnold et al., 2026; Cuyckens et al., 2026). These studies establish the value of feedback and implementation choices. SILICA builds on this setting to measure independently assessed, time-bounded delivery, rather than presenting joint optimization itself as a new idea.

Cross-stage closure and protected assessment. CLOSER-Bench studies budgeted cross-stage closure; PostEDA-Bench and TicTacBench address physical convergence, rule repair, and timing closure (Zhou et al., 2026; Liu et al., 2026; Wang et al., 2026). Commercial RTL-to-GDS agent studies already examine full-flow execution (Deng et al., 2026). EEBench uses simulation and hidden engineering checks in electronic design (atopile, 2026). SILICA focuses on open-ended improvement of a fixed digital computation: matched edit authority, protected behavioral checks, artifact-matched post-route evidence, and timed nominations connect difficulty to reliable improvement. Appendix J gives the scope comparison.

3 Problem Definition

We evaluate an agent system $\pi = (M, H)$: a model M with a harness H that supplies prompts, memory, tools, and control. Task i is

$$\mathcal{T}_i = (x_i^0, \mathcal{C}_i, \mathcal{E}_i, \mathcal{A}_i, B_i), \qquad x_i^0 = (r_i^0, \phi_i^0), \quad x = (r, \phi). \tag{1}$$

RTL r_i^0 with contract $\mathcal{C}_i$ fixes behavior; the reference RTL–flow pair x_i^0 fixes the comparison cost. Environment $\mathcal{E}_i$ fixes technology, tools, measurement, and resource caps. Edit authority $\alpha_i \in \{\text{RTL-only}, \text{Flow-only}, \text{Joint}\}$ restricts the legal action set $\mathcal{A}_i$; B_i is wall-clock search time.

At decision j, the system uses public history h_j to choose $a_j \sim \pi(\cdot \mid h_j)$: inspect, edit, request evidence, register, recover, or stop. All elapsed search time, including model and tool waits, counts against B_i. Long-horizon structure comes from dependent decisions: earlier actions determine later evidence, and checking one option reduces the time available for others. The output is a timed nomination $\hat{x}$, not the best artifact found afterwards. Independent evaluation returns

$$\mathcal{F}_{\mathcal{E}_i}(\hat{x}) = (V^{\text{func}}, V^{\text{impl}}, A, P, D), \qquad V = V^{\text{func}} V^{\text{impl}}. \tag{2}$$

The indicators describe protected RTL and netlist replay, and implementation completeness; A, P, D are area, power, and a delay proxy, undefined when not measurable. Section 5 defines quality Q, utility U, improvement J, and the aggregates R, EVQ, and VIR; Algorithm 1 states the evaluation loop.

4 SILICA-Bench at a Glance

Figure 2 organizes the benchmark into three zones. **Task inputs** provide complete reference RTL, its behavioral contract, a callable reference flow, public reference PPA, and the time budget. **Adaptive**

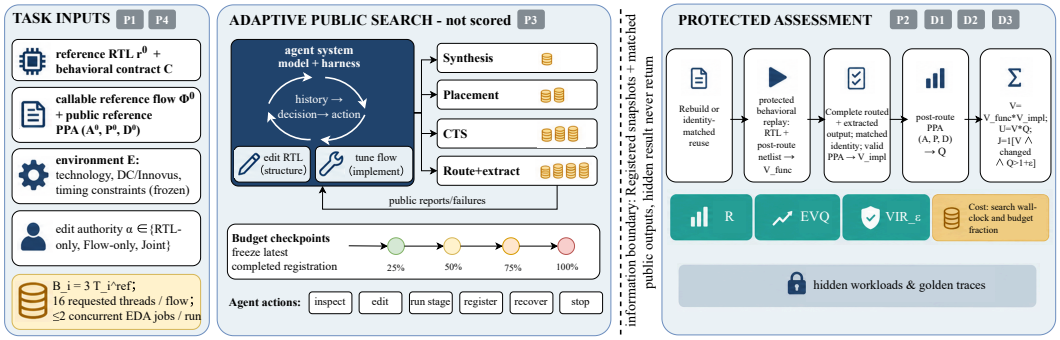

Figure 2: SILICA in three zones. Public reports guide search; only registered snapshots enter protected RTL and post-route netlist replay, with local reconstruction or identity-matched reuse of implementation outputs. Each agent run permits at most two concurrent EDA jobs. T_{flow} denotes T_i^{ref}; Φ denotes flow ϕ. Coin icons indicate evidence cost, not measured stage ratios. Implementation acceptance requires complete outputs and consistent identity; physical-rule checks remain diagnostics. Hidden results never return to search. This study reports search wall-clock and budget fraction (Section 6.4); intermediate checkpoints illustrate the protocol.

public search lets the model–harness system edit RTL and legal flow settings and request evidence from Design Compiler and Innovus, with at most two concurrent EDA jobs per agent run. Public reports guide decisions but never become final scores. **Protected assessment** binds each registered nomination to an implementation, checks its RTL and post-route netlist under hidden workloads, and obtains PPA under the reference measurement profile. Hidden workloads and assessment results never return to search. The same registration rule defines nominations at any budget fraction; Figure 2 marks 25, 50, 75, and 100%, while this study scores the terminal nomination. The following five design elements connect properties P1–P4 to distinctions D1–D3.

5 Benchmark Design

The environment creates choices; the protocol makes their consequences measurable. We first define the target and evidence available to the agent, then specify independent scoring, timed nominations, and matched comparisons.

5.1 Open-Ended Objective under a Hard Constraint (P1, P2)

SILICA starts from a working design and asks for a better implementation of the same computation. Reference RTL makes bit widths, arithmetic, and state transitions executable; the contract fixes interfaces, reset, legal inputs, and observable latency. This reduces the ambiguity of a prose-only starting point. The reference is a comparator, not an assumed optimum: there is no required patch and no reward for resembling one. RTL structure and legal flow choices are open; behavior, technology, clocks, and measurement remain fixed. Open-ended means that many answers receive graded quality, not that physical improvement has no upper limit.

The two references anchor different things: required behavior and physical cost. Narrowing legal inputs, changing observable latency, or relaxing the clock changes the task rather than improving its implementation.

For a valid nomination, we compare its area A_i, power P_i, and delay proxy D_i with reference values A_i^0, P_i^0, D_i^0:

$$Q_i = \left(\frac{A_i^0}{A_i}\frac{P_i^0}{P_i}\frac{D_i^0}{D_i}\right)^{1/3}, \quad U_i = \begin{cases} Q_i, & V_i = 1, \\ 0, & V_i = 0, \end{cases} \quad D_i = T_i^{\text{clk}} - s_i^{\text{setup}}. \tag{3}$$

$T_i^{\text{clk}} = 200$ ps and s_i^{setup} is signed minimum setup slack in the fixed path group. D_i is a constraint-relative proxy, not a measured maximum clock frequency. Positive PPA values are required. The geometric mean gives a dimensionless trade-off; $Q_i > 1$ need not improve every component, so

Table 1: Measured reference stage costs: median [min, max], 36 designs, 16 requested threads per flow. DC, PnR, and total reference time are minutes; budget is hours. The cost of later evidence is task-dependent.

Size (n)	DC	PnR	PnR/DC	T^{ref}	B_i (h)
Small (15)	7.0 [4.7, 12.8]	10.8 [6.8, 21.6]	1.27 [0.66, 4.62]	19.3 [12.1, 26.3]	0.96 [0.61, 1.31]
Medium (17)	11.4 [6.5, 22.8]	23.9 [16.5, 223.5]	2.06 [1.19, 27.91]	36.1 [24.3, 231.5]	1.80 [1.22, 11.58]
Large (4)	38.3 [12.9, 232.2]	201.7 [49.4, 414.3]	5.34 [0.94, 14.34]	324.7 [80.5, 460.0]	16.23 [4.03, 23.00]
All (36)	10.6 [4.7, 232.2]	19.8 [6.8, 414.3]	1.64 [0.66, 27.91]	30.2 [12.1, 460.0]	1.51 [0.61, 23.00]

component ratios are also reported. Invalid nominations receive $U_i = 0$ without inventing missing PPA.

Behavioral preservation is a hard target. Evidence comes from protected replay of candidate RTL and its corresponding post-route netlist against reference observations under a frozen finite workload. Both must pass. A rewrite that changes required outputs or reset behavior is unacceptable regardless of its area (Figure 1(a)). Finite replay supports preservation on the tested workload, not universal formal equivalence (Appendix B).

5.2 Staged Evidence and Coupled Choices (P3)

Long-horizon difficulty comes from how evidence changes decisions, not from slow tools alone. Reference synthesis takes a median 10.6 minutes, compared with 19.8 minutes for the combined physical flow. The per-design physical flow to synthesis ratio ranges from 0.66 to 27.91, with median 1.64. Thus later evidence can be substantially more expensive, but an orders-of- magnitude gap is not universal. Table 1 summarizes measured reference stage costs; separate placement, CTS, and routing times are not inferred from the combined measurement.

The search allowance is $B_i = 3T_i^{\mathrm{ref}}$, using measured complete reference-flow time with 16 requested threads per flow. Up to two concurrent EDA jobs share this deadline, without a minimum budget or truncation. This scales the opportunity to obtain evidence; it is not an attempt cap.

RTL changes what is implemented; flow choices change how it is implemented. Sharing arithmetic can reduce logic while adding multiplexing and wire delay. Synthesis gains need not survive routing (Fang et al., 2026), so only post-route evidence determines final quality. A failed check can still guide a useful revision, while another full run can consume time needed to retain and deliver a candidate. The benchmark rewards the outcome rather than a prescribed strategy or tool-call count. Algorithm 1 states this loop: public actions share one clock; completed registrations fix the artifact; protected assessment scores the terminal nomination.

5.3 Verification-Gated Scoring (D1, D2)

Public evidence informs search; protected assessment determines quality. Implementation takes one of two recorded forms: local reconstruction in the frozen environment, or identity-matched reuse of complete public flow outputs for the registered RTL–flow pair (Appendix C.4). In either case, the evaluator applies protected replay to both RTL and the corresponding netlist and reads PPA from that implementation's final reports. Agent-written summaries do not establish quality. Workloads and assessment outcomes remain inaccessible to search.

This boundary keeps evidence attached to the artifact delivered. A passing check for one RTL version and favorable PPA for another pair do not jointly validate a nomination. Matching implementation identity and evaluating its behavior prevents such evidence from being combined. Assessment independence comes from protected checks and controlled measurement; it does not require rerunning every eligible physical output.

$V_i^{\mathrm{func}} = V_i^{\mathrm{RTL}} V_i^{\mathrm{net}}$ requires both RTL and corresponding post-route netlist observations to match the protected reference; each indicator is one on a pass and zero on an assessed failure. $V_i^{\mathrm{impl}} = 1$ requires a complete routed and extracted result with consistent identity and valid post-route PPA. The relaxed physical profile records setup, hold, design-rule, and connectivity diagnostics but does

Algorithm 1 One SILICA task: public search, registration, protected assessment

Require: Task $\mathcal{T}_i$, system π, margin ϵ; protected workload W_i
Register x_i^0 at $t = 0$; initialize public history h and ledger $\mathcal{L}$
while $\mathrm{elapsed}() < B_i$ and π has not stopped **do**
  $a \sim \pi(\cdot \mid h)$ ▷ inspect, edit, run, register, recover, stop
  Enforce $\mathcal{A}_i$; admit at most two concurrent EDA jobs
  Execute admitted action; append public evidence and elapsed time to h
  if a registers a legal pair x and copying completes at $t \le B_i$ **then**
    Append immutable snapshot, identity, and completion time (x, t) to $\mathcal{L}$
  end if
end while
Close public execution; $\hat{x} \leftarrow$ last registration with $t \le B_i$ ▷ no hindsight selection
Obtain implementation of $\hat{x}$: rebuild or identity-matched reuse
Replay $\hat{r}$ and its post-route netlist on W_i against reference observations
Set $V^{\mathrm{func}} = V^{\mathrm{RTL}} V^{\mathrm{net}}$; check implementation to obtain V^{impl}
$V \leftarrow V^{\mathrm{func}} V^{\mathrm{impl}}$; initialize $U \leftarrow 0$, $J \leftarrow 0$
if $V = 1$ **then**
  Compute Q from matched PPA; $U \leftarrow Q$; $J \leftarrow \mathbf{1}[\chi = 1 \ \wedge \ Q > 1 + \epsilon]$
end if
return $(V, U, J,$ wall-clock cost$)$ ▷ protected results never enter h

not require them to pass. Consequently, $V_i = 1$ is acceptance under this profile, not timing closure or physical signoff. Figure 3(b) separates these checks from quality; Appendix C specifies measurement units and evidence requirements.

Let $\chi_i = 1$ if normalized RTL or legal flow differs from the reference, and zero otherwise. $J_i = 1$ only when $V_i = 1$, $\chi_i = 1$, and $Q_i > 1 + \epsilon$; otherwise $J_i = 0$. We report $\epsilon = 0.05$ as the primary analysis margin and 0.01 as a sensitivity analysis, without treating either as a statistical test or a preregistered noise bound (Appendix G). For N tasks,

$$R = \frac{1}{N} \sum_i V_i, \qquad \mathrm{EVQ} = \frac{1}{N} \sum_i U_i, \qquad \mathrm{VIR}_\epsilon = \frac{1}{N} \sum_i J_i. \tag{4}$$

Validity Rate R measures accepted nominations; Expected Verified Quality (EVQ) includes the quality cost of failures; Verified Improvement Rate (VIR) counts changed, accepted gains. All use the full task denominator. An unchanged reference can be valid with $Q_i \approx 1$ but has $J_i = 0$. A few strong successes may coexist with low EVQ if many nominations fail. Reporting all three resolves D1; independent post-route evidence resolves D2. Cost is reported separately, never hidden inside quality.

When at least one nomination is accepted, EVQ equals R times the arithmetic mean of quality among accepted nominations. Joint reporting distinguishes better implementations, higher reliability, and unchanged retention.

5.4 Nomination-Based Evaluation (D3)

A candidate counts only when the agent registers it in time. Let $x_{i,\ell}$ be snapshot ℓ, indexed in registration-completion order, and $t_{i,\ell}$ its completed-registration time. The protocol requires a reference registration at time zero, $x_{i,0} = x_i^0$ and $t_{i,0} = 0$. At fraction k of the budget,

$$\ell^*(i, k) = \max\{\ell : t_{i,\ell} \le k B_i\}, \qquad \hat{x}_{i,k} = x_{i,\ell^*(i,k)}, \quad k \in [0, 1]. \tag{5}$$

The rule applies to any k; this study scores $k = 1$. Intermediate checkpoints in Figures 2 and 3 illustrate the protocol. Without a new registration, the previous nomination carries forward. Recovering a workspace does not replace a nomination; the restored candidate must be registered if it is to replace the current one. Registration proves identity and timing, not validity. Figure 3(a) contrasts this rule with selecting a promising artifact after seeing hidden outcomes.

At a timeout, the latest nomination remains the evaluated object. A valid reference fallback can retain unit quality and zero improvement; an invalid later nomination is not replaced by an earlier successful one. Only the terminal registered artifact contributes to the empirical results; checkpoint trajectories are an extension of the same rule.

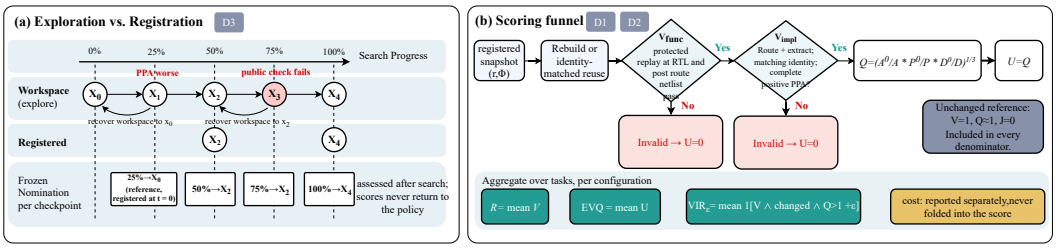

Figure 3: Registration and verification-gated scoring (schematic). (a) This separate example freezes x_0, x_2, x_2, x_4 at successive checkpoints; X labels denote candidates x, and "PPA worse" is public evidence. Restoring the workspace does not change the registered nomination. (b) Both RTL and post-route netlist replay must pass before quality counts. Invalid nominations receive zero utility without imputing PPA. Means use one nomination per task, with cost separate; this study scores the terminal nomination ($k = 1$).

5.5 Controlled Comparison and the Task Suite (P4)

A result describes an agent system under declared conditions. We match tasks, public tools, harness, evaluator, time allowances, and edit rules; model identity, provider route, prompt, sampling, and memory policy identify each system. RTL-only permits structural changes, Flow-only permits legal flow changes, and Joint permits both. The anchor comparison asks whether added freedom helps the same system; it does not by itself prove interaction between particular edits. Harness choice is fixed here, not a separately measured experimental axis.

The suite contains 36 designs: eight NoC/bus/DMA, seven DSP/FEC/arithmetic, seven cryptographic, five memory/controller, four I/O, two CPU, two ML, and one graphics task. It spans eight domains. For descriptive size stratification, we use the larger of each design's historical generic cell count and current DC count: Small below 10,000, Medium from 10,000 to below 50,000, and Large at least 50,000. This gives 15/17/4 designs. These mixed-source counts are a descriptive size proxy, not comparable NAND2-equivalent gates or a replacement for current physical area. Appendix D reports one selected total, DC sequential cells, reference PPA, and runtime, with count provenance retained in the audit.

Every task has equal weight; size does not enter PPA normalization. New designs or tool environments define separately frozen versions. Diversity supports comparisons across functions but does not establish immunity to training contamination or unlimited physical headroom.

6 Experiments

We ask three questions of the same 240 assessed nominations: do agents deliver verified improvement (D1), what does edit authority buy, and does budget spent become improvement delivered (D2, D3)?

6.1 Experiment Setup

Six model–harness systems use Claude Code under Joint authority: GPT-6 Sol, Claude Opus 5, Qwen3.8-Max, GLM-5.3, Kimi-K3, and Grok-4.7. The Qwen anchor also uses RTL-only and Flow-only, giving eight configurations and 288 scheduled runs over the 36-design suite (Table 2). Our complete-case rule admits a design to paired comparison only when every configuration has an assessed nomination, regardless of its validity or quality. This yields 30 designs and excludes six affected by campaign availability disruptions during long-running service and network operation (Appendix D.4). The paired set spans all eight domains and 579–63,435 reported cells, but contains only one Large design. All three assessed functional failures remain in the comparison with $U = 0$.

Budgets are identical across configurations for a task and total 45.54 hours per configuration on the paired set. We evaluate one run per task and configuration ($S = 1$), with the profile in Table 2. Task-paired 95% bootstrap intervals describe variation across these designs, not repeated-run stability; they neither remove availability bias nor adjust for multiple comparisons. Cost is search wall-clock

Table 2: Experimental setup. The 36-design suite and 30-design paired analysis have distinct roles; all main scores use the latter.

Item	Setting
Systems / authority	Six Joint systems; Qwen also RTL-only and Flow-only; Claude Code harness
Comparison set	30 designs, eight domains, 14 Small / 15 Medium / 1 Large; one run per configuration and task
Tools / measurement	Design Compiler + Innovus; ASAP7; 200 ps; vectorless power; post-route extraction
Search allowance	$B_i = 3T_i^{\mathrm{ref}}$; 0.61–4.03 h, median 1.40 h; at most two EDA jobs per run, 16 requested threads per flow
Assessment	RTL + post-route netlist replay; complete physical outputs; setup/hold/DRC/connectivity as diagnostics
Statistics	Terminal nomination; $\epsilon = 0.05$ and 0.01; 10,000 task-paired bootstrap draws, seed 20260926

Table 3: Terminal Joint results on the same 30 designs. EVQ includes invalid nominations as zero; component ratios are geometric means over valid nominations (lower is better). Improvement requires a changed artifact; the 5% margin is a declared analysis threshold, with 1% sensitivity in Appendix G.

System	n_{valid}/N	EVQ [95% CI]	$\mathrm{VIR}_{.05}$	A/A^0	P/P^0	D/D^0
GLM-5.3	30/30	1.0694 [1.0255, 1.1209]	10/30	0.9526	0.9422	0.9301
GPT-6 Sol	30/30	1.0688 [1.0333, 1.1145]	8/30	0.9284	0.9305	0.9620
Qwen3.8-Max	30/30	1.0329 [1.0166, 1.0508]	7/30	0.9762	0.9730	0.9583
Claude Opus 5	30/30	1.0319 [1.0033, 1.0591]	10/30	0.9846	1.0004	0.9323
Kimi-K3	30/30	0.9866 [0.9722, 0.9984]	0/30	1.0194	1.0432	0.9815
Grok-4.7	28/30	0.9273 [0.8222, 1.0049]	4/30	1.0157	1.0335	0.9805

and fraction of the reference-derived allowance. Appendix F gives per-design utilities, component quality, and cost.

6.2 RQ1: Do Agents Deliver Verified Improvement?

Validity and improvement separate. Table 3 shows at least 28/30 valid nominations for every Joint system, while $\mathrm{VIR}_{.05}$ ranges from 0/30 to 10/30. GLM and GPT have the largest observed EVQ, but their paired difference (GPT minus GLM) is -0.0006 with interval $[-0.0276, 0.0249]$, leaving their ordering unresolved. Kimi retains an unchanged reference on 22 tasks and improves no changed nomination by more than 5%. Here *retention* means an unchanged nomination, not an inferred recovery strategy; a *regression* is a valid, changed nomination with $Q < 1$.

Failures and trade-offs matter. Grok's valid-only mean $Q = 0.9935$ becomes EVQ $= 0.9273$ when its two functional failures count as zero. Component ratios reveal a different distinction: Claude reduces geometric mean delay to 0.9323 of reference while power is 1.0004. Across all eight configurations, 71 of 133 valid nominations with $Q > 1$ worsen at least one PPA component. Thus a scalar gain is not a Pareto improvement. These results instantiate Figure 1(a): valid output, high quality among successes, and reliable improvement answer different questions.

6.3 RQ2: What Does Edit Authority Buy?

Table 4 compares the same anchor on the same tasks. Joint exceeds Flow-only by 0.0247 EVQ, with interval $[0.0109, 0.0406]$, but its difference from RTL-only, 0.0167 with interval $[-0.0468, 0.1017]$, has unresolved sign. RTL-only has the highest valid-only geometric mean quality among the three authorities (1.0475), yet its one functional failure lowers EVQ below Joint. Flow-only is valid on all tasks and exceeds a 5% gain on one. Added freedom therefore changes both opportunity and risk; these anchor-specific contrasts do not establish universal RTL–flow synergy.

Table 4: Edit authority for the Qwen anchor; budgets, harness, tasks, and assessment match. Component ratios condition on validity; all other metrics retain 30 tasks.

Authority	n_{valid}/N	EVQ [95% CI]	$\text{VIR}_{.05}$	A/A^0	P/P^0	D/D^0
Joint	30/30	1.0329 [1.0166, 1.0508]	7/30	0.9762	0.9730	0.9583
RTL-only	29/30	1.0162 [0.9315, 1.0789]	11/30	0.9587	0.9686	0.9368
Flow-only	30/30	1.0082 [1.0000, 1.0177]	1/30	1.0035	0.9947	0.9786

Paired ΔEVQ (95% CI): Joint $-$ RTL-only 0.0167 $[-0.0468, 0.1017]$; Joint $-$ Flow-only 0.0247 $[0.0109, 0.0406]$; RTL-only $-$ Flow-only 0.0081 $[-0.0771, 0.0707]$.

Table 5: Budget use and delivered artifacts ($N = 30$ per row). Stops: completed/deadline/agent error. Levers: both/RTL/flow/neither. Regressions are valid changed nominations with $Q < 1$; gains use J at $\epsilon = .05$. Reuse denotes an identity-matched public implementation.

System	Budget	Stops	Levers	Changed	$> 5\%$	Regress.	Invalid	Reused
GLM-5.3	92.5%	23/6/1	10/9/7/4	26	10	5	0	20
GPT-6 Sol	95.1%	16/14/0	4/21/1/4	26	8	1	0	19
Qwen3.8-Max	92.5%	21/8/1	6/7/9/8	22	7	1	0	14
Claude Opus 5	94.1%	23/7/0	12/6/9/3	27	10	5	0	18
Kimi-K3	99.1%	2/28/0	2/1/5/22	8	0	6	0	0
Grok-4.7	67.9%	21/9/0	12/3/1/14	16	4	7	2	8
Qwen RTL-only	92.7%	20/10/0	0/23/0/7	23	11	4	1	13
Qwen Flow-only	92.8%	18/12/0	0/0/20/10	20	1	6	0	11

6.4 RQ3: Does Budget Spent Become Improvement Delivered?

Table 5 links elapsed search to delivered artifact. Kimi consumes 99.1% of its task allowance on average and reaches the deadline on 28/30 tasks, yet EVQ is below one. GLM and GPT consume 92.5% and 95.1% and deliver higher quality. Time spent is insufficient evidence of progress. Terminal registrations, rather than tool-call counts or a retrospective best candidate, define the outcome.

Of 168 changed nominations, 103 reuse an identity-matched public implementation and 65 use local reconstruction. Valid regressions occur in 8/103 and 27/65, respectively; 11 of the 12 with $Q < 0.95$ use reconstruction. Conversely, 40 of 51 changed gains above 5% reuse matched outputs. These associations motivate checking the exact delivered artifact, but the assessment path is not randomized and also depends on output availability. It cannot establish that reconstruction caused regression or that an agent never ran a matching flow. Two functional failures occur even with reused physical outputs: completed implementation does not replace behavioral verification.

6.5 Failure Modes and Interpretation

The three failures are output mismatches: Grok on Golay decoding and RAM testing, and Qwen RTL-only on rasterization. All have complete physical outputs. Another 35 changed nominations pass verification but regress, including 12 below $Q = 0.95$ (Appendix H). The relaxed physical profile accepts negative slack; these are improvement results, not 5 GHz timing-closure claims. Taken together, the study tests a specific competence: preserve behavior, use costly evidence, and deliver a checked artifact whose measured benefit survives final assessment.

7 Conclusion

SILICA-Bench evaluates whether long-horizon LLM agents deliver verified circuit improvements. Fixed behavior, open implementation choices, staged evidence, and timed nominations connect decisions to checked artifacts. Across 240 paired nominations, high validity coexists with limited gains; neither added edit authority nor elapsed search ensures improved delivery. Separating validity, quality, and changed improvement exposes these differences. Future work can extend this evaluation to larger circuits and stricter verification and physical-acceptance profiles, further testing the capabilities of LLM agents against real-world industrial demands.

AI USE STATEMENT

Generative AI tools were used only to edit the manuscript text and assist with LaTeX formatting, including migration to the ICLR 2027 template. They were not used to develop research ideas, implement research code, conduct experiments, or analyze results. The authors take responsibility for the final content of the manuscript.

REFERENCES

Felix Arnold, Ryan Amaudruz, Dimitrios Tsaras, Renzo Andri, and Lukas Cavigelli. RTLScout: Joint Agentic Code and Synthesis Optimization for Efficient Digital Circuits. *arXiv preprint arXiv:2606.06530*, 2026. URL https://arxiv.org/abs/2606.06530v3. Version 3, 14 September 2026.

atopile. EEBench: Methodology, 2026. URL https://www.eebench.org/methodology.html. Developer-published benchmark methodology, accessed 25 September 2026.

Stef Cuyckens, Mihaela Jivanescu, Jun Yin, Chao Fang, and Marian Verhelst. ARES: Adaptive Reasoning-Effort Steering for PPA- and Cost-Aware RTL Optimization with LLM Agents. *arXiv preprint arXiv:2607.27879*, 2026. URL https://arxiv.org/abs/2607.27879v1.

Jinyuan Deng, Zhengrui Chen, Xufeng Wei, Tianyu Xing, Chenyi Wen, Qi Sun, and Cheng Zhuo. Can AI Agents Really Complete RTL-to-GDS? Lessons from Benchmarking Tool-Interactive EDA Workflows. *arXiv preprint arXiv:2607.17528*, 2026. URL https://arxiv.org/abs/2607.17528v3. Version 3, 23 July 2026.

Wenji Fang, Yao Lu, Shang Liu, Jing Wang, Ziyan Guo, Junxian He, Fengbin Tu, and Zhiyao Xie. Dr. RTL: Autonomous Agentic RTL Optimization through Tool-Grounded Self-Improvement. *arXiv preprint arXiv:2604.14989*, 2026. URL https://arxiv.org/abs/2604.14989v2. Version 2, 26 April 2026.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI Agents That Matter. *arXiv preprint arXiv:2407.01502*, 2024. URL https://arxiv.org/abs/2407.01502v1.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI Ability to Complete Long Software Tasks. *arXiv preprint arXiv:2503.14499*, 2025. URL https://arxiv.org/abs/2503.14499v4. Presented at NeurIPS 2025; version 4, 10 July 2026.

Mingjie Liu, Nathaniel Pinckney, Brucek Khailany, and Haoxing Ren. VerilogEval: Evaluating Large Language Models for Verilog Code Generation. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 1–8, 2023. doi: 10.1109/ICCAD57390.2023.10323812.

Pengju Liu, Nuo Xu, Jinwei Tang, Yu Cao, and Caiwen Ding. Bridging the Last Mile of Circuit Design: PostEDA-Bench, a Hierarchical Benchmark for PPA Convergence and DRC Fixing. *arXiv preprint arXiv:2605.06936*, 2026. URL https://arxiv.org/abs/2605.06936v3. Version 3, 21 July 2026.

Yao Lu, Shang Liu, Hangan Zhou, Wenji Fang, Qijun Zhang, and Zhiyao Xie. A New Benchmark for the Appropriate Evaluation of RTL Code Optimization. *arXiv preprint arXiv:2601.01765*, 2026. URL https://arxiv.org/abs/2601.01765v1.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848*, 2025. URL https://arxiv.org/abs/2504.01848v3.

Bowei Wang, Zhigang Fang, Zhijie Yang, Renzhi Chen, Shanshan Li, and Lei Wang. TicTacBench: Benchmarking Timing Closure Capabilities of Coding Agents. *arXiv preprint arXiv:2609.23363*, 2026. URL https://arxiv.org/abs/2609.23363v1.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. *arXiv preprint arXiv:2411.15114*, 2025. URL https://arxiv.org/abs/2411.15114v2. Version 2, 27 May 2025.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. *arXiv preprint arXiv:2405.15793*, 2024. URL https://arxiv.org/abs/2405.15793v3.

Xufeng Yao, Yiwen Wang, Xing Li, Yingzhao Lian, Ran Chen, Lei Chen, Mingxuan Yuan, Hong Xu, and Bei Yu. RTLRewriter: Methodologies for Large Models Aided RTL Code Optimization. In *2024 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 98:1–98:7, 2024. doi: 10.1145/3676536.3676775.

Peilong Zhou, Zhirong Chen, Cangyuan Li, Haoyu Gao, Kaiyan Chang, Ziming Qu, and Ying Wang. CLOSER-Bench: Evaluating Budgeted Cross-Stage Design Closure for Hardware Agents. *arXiv preprint arXiv:2607.16632*, 2026. URL https://arxiv.org/abs/2607.16632v1. Version 1, 18 July 2026.

A Notation and Task Contract

A.1 Symbols

Table A.1 defines the quantities used in the protocol. Scores refer to one nominated RTL–flow pair; a favorable report for a different pair cannot supply its behavioral or physical evidence.

Table A.1: Notation. All task averages use the same paired design set.

Symbol	Meaning
$\pi = (M, H)$	Model and harness evaluated as one deployed system
$x = (r, \phi)$; x_i^0	Candidate RTL and legal flow; task reference pair
$\mathcal{C}_i, \mathcal{E}_i, \mathcal{A}_i$	Behavioral contract, fixed environment, and legal actions
T_i^{ref}, B_i	Measured reference-flow time and search allowance $3T_i^{\mathrm{ref}}$
h_j, a_j	Public history and the next action
$\mathcal{L}, t_{i,\ell}, \hat{x}_{i,k}$	Registration ledger, completion time, and nomination at budget fraction k
W_i	Protected finite workload and observation policy
$V^{\mathrm{RTL}}, V^{\mathrm{net}}$	Pass indicators for candidate RTL and post-route netlist replay
$V^{\mathrm{func}}, V^{\mathrm{impl}}, V$	Dual replay, physical completeness, and their conjunction
A, P, D; A^0, P^0, D^0	Candidate and reference area, power, and delay proxy
$T_i^{\mathrm{clk}}, s_i^{\mathrm{setup}}$	Fixed 200 ps period and signed minimum setup slack
Q, U	Valid-only PPA quality and utility including invalidity as zero
χ, J, ϵ	Artifact changed, verified improvement indicator, and improvement margin
$R, \mathrm{EVQ}, \mathrm{VIR}_\epsilon$	Mean validity, mean utility, and mean improvement indicator
s_i	Selected total cell count for inventory size; distinct from setup slack
α_i	Edit authority: RTL-only, Flow-only, or Joint
$\mathbf{1}[\cdot]$	Indicator equal to one when its condition holds and zero otherwise
N, S	Paired designs (30) and runs per design/configuration (1)

A.2 Fixed behavior and permitted edits

The reference RTL and task contract fix the top interface, parameterization, clock and reset behavior, legal inputs, observable outputs, and latency. RTL-only edits implementation code while retaining the reference flow; Flow-only edits allowed synthesis and implementation choices while retaining reference RTL; Joint permits both. An edit cannot relax the clock, change the scoring workload or technology, redefine outputs, or move behavior outside the observation window. These rules define the optimization task, not a preferred solution.

A.3 Public and protected files

The agent sees reference RTL, callable flow scripts, task instructions, reference metrics, and its own public reports. Candidate RTL and authorized flow settings are writable; contracts and measurement policy are fixed. Protected testbenches, patterns, golden observations, and final assessment records are outside the agent workspace. Recovering a working directory does not register an artifact. Only a completed immutable registration changes what the evaluator selects.

B Behavioral Preservation and Replay Evidence

B.1 Dual replay and the behavioral target

The functional target is preservation of reference observations under a frozen workload. The evaluator uses VCS and the protected replay adapter to run candidate RTL and the corresponding post-route netlist against the reference observations. With Obs denoting contract-defined observations and $\mathrm{net}(\hat{x})$ its corresponding netlist,

$$V_i^{\mathrm{func}} = \prod_{z \in \{\hat{r}, \mathrm{net}(\hat{x})\}} \mathbf{1}[\mathrm{Obs}(z, W_i) = \mathrm{Obs}(r_i^0, W_i)]. \tag{6}$$

This is the conjunction $V^{\mathrm{RTL}}V^{\mathrm{net}}$. Matching the RTL alone does not establish that the implemented netlist preserves the same behavior.

B.2 Observation and comparison policy

Top module, parameter overrides, simulation options, reset, seeds, patterns, and observation windows belong to the task contract. The reference and candidate use the same settings. Comparison concerns defined output values at required events, not textual simulator banners, absolute file paths, or irrelevant log formatting. Masks are fixed before evaluation and apply only to contractually unobservable values; they do not ignore an unexpected unknown, remove a difficult output, or shift required latency after seeing a mismatch. Empty output or premature termination cannot pass simply because no unequal line was found. Netlist replay uses the fixed ASAP7 cell simulation models; behavioral replay does not itself establish timing closure.

B.3 What this evidence establishes

A pass supports agreement on the tested workload, not universal equivalence for every legal input sequence. Historical coverage in Table D.3 documents task provenance, rather than coverage of each candidate assessment. Large coverage percentages do not turn finite replay into a proof. The reported three output mismatches remain invalid regardless of attractive physical measurements.

C Commercial Flow and Measurement Policy

C.1 Environment and measurement boundary

The physical profile uses Design Compiler, Innovus, ASAP7, a 200 ps main clock, and the recorded post-route extraction settings. Each flow requests 16 threads; at most two EDA jobs run concurrently for an agent task under one deadline. A thread request is distinct from measured CPU consumption. The task boundary covers synthesis, placement, clock-tree synthesis, routing, and post-route analysis; "RTL-to-GDS" does not imply fabrication signoff. No open-source flow is an experimental condition in this study.

C.2 PPA definitions

Area is the reported post-route cell area in $\mu\mathrm{m}^2$; power is mW under the fixed vectorless activity and measurement profile. With signed minimum setup slack s^{setup} in ps, $D = 200 - s^{\mathrm{setup}}$ ps. The path group and analysis conditions match the reference. This is a constraint-relative delay proxy, not an empirical $F_{\max}$. Area, power, and delay must be positive and finite for quality to be defined. Ratios use the reference measurement for the same task, rather than historical size or another parameterization. A geometric mean permits trade-offs, so each component remains visible alongside Q.

C.3 Implementation validity and diagnostics

$V^{\mathrm{impl}} = 1$ requires a complete routed, extracted implementation, consistent artifact identity, and valid PPA. Setup, hold, design-rule, and connectivity statuses are diagnostic under the relaxed physical profile; they are not additional pass gates. All 240 paired nominations meet the implementation gate, while three fail functionality. Negative setup slack is accepted, so neither $V = 1$ nor $Q > 1$ means that 5 GHz was achieved.

C.4 Local reconstruction and identity-matched reuse

The assessment record distinguishes `LOCAL_REBUILD` (137 paired nominations) from `AGENT_PUBLIC_EDA_REUSED` (103). Reconstruction implements the nominated pair in the assessment environment. Reuse accepts complete physical outputs matched to that pair and measurement profile; the evaluator reads the implementation reports and applies protected dual replay. A natural-language agent summary cannot substitute for this evidence. The flow identity is not replaced by a more favorable unrelated flow.

Reuse reduces duplicate implementation cost, but it is a weaker independence claim than recomputing every physical output. Source reporting flags that actual execution-driver source is not packaged for reused outputs; identity matching does not establish a complete audit of that driver. A local rebuild may also arise from packaging or output availability. It does not prove that the agent never attempted a full flow. Results grouped by PPA source are therefore observational. The extra GLM/BMB record carries a requested-flow versus executed-output semantic warning and is not used in paired inference (Table F.9).

D THE 36-DESIGN TASK SUITE

D.1 INVENTORY AND SIZE

Table D.1 lists the 36 reference designs in eight domains: NOC (bus/NoC/DMA), DSP (arithmetic/FEC), SEC (cryptography), MEM (memory and controllers), IO (peripherals), CPU, ML, and GPU (rasterization). These are functional categories, not claims that every circuit accelerates an LLM. The archived task provenance is retained in the source inventory.

Table D.1 reports one total cell count s_i per design, selected as the larger applicable count in the historical design guide and reference implementation overview, and the DC sequential-cell count. Small is $s_i < 10^4$, Medium is $10^4 \leq s_i < 5 \times 10^4$, and Large is $s_i \geq 5 \times 10^4$: 15/17/4 designs. The source audit retains the original counts. These descriptive counts do not replace physical area in scoring.

D.2 REFERENCE PHYSICAL MEASUREMENTS

Table D.2 records PPA, synthesis time, physical-flow time, total reference time, and budget. Total wall-clock includes flow orchestration, so it can exceed DC plus PnR time. $B_i = 3T_i^{\mathrm{ref}}$ uses unrounded seconds without a floor or clipping. All 36 reference builds are accepted under the relaxed profile; 35 have negative setup slack.

D.3 HISTORICAL COVERAGE

Table D.3 preserves archived coverage and stage exceptions. An N/A entry retains its structural meaning and is not an empty result or 100% coverage. Historical coverage and candidate replay answer different questions (Appendix B).

Table D.1: The 36-design suite: total cell count, sequential cells, and size stratum. Total cells s_i select the larger applicable count in the historical design guide and the reference implementation overview. Sequential cells are from the DC reference report.

No.	Design ID	Domain	Total cells (s_i)	Seq. (DC)	Size
1	alex-axi-vfifo-full	MEM	1,539,631	395,114	L
2	kemet-ptahconv-full	ML	144,439	24,676	L
3	spinal-bmb-ooo-full	NOC	70,091	14,063	L
4	G2R10-rpeters-rasterizer	GPU	63,435	10,049	L
5	G2R02-fec-rs-dec	DSP	46,906	9,829	M
6	tachyum-fma-full	DSP	42,324	2,266	M
7	spinal-mac-eth-full	IO	37,541	8,633	M
8	secworks-sha512	SEC	29,813	3,673	M
9	secworks-aes	SEC	24,601	2,987	M
10	E14-verilog-axi	NOC	22,985	4,710	M
11	tensorwright-convolution-engine	ML	22,830	1,942	M
12	E25-secworks-blake2s	SEC	20,883	1,875	M
13	G2R22-freecores-des	SEC	19,924	1,984	M
14	secworks-chacha	SEC	17,408	1,098	M
15	secworks-sha256	SEC	15,526	1,806	M
16	G2R05-fec-golay24-dec	DSP	15,451	1,108	M
17	G2R11-ultraembedded-ddr3-axi	MEM	14,356	3,047	M
18	E18-nocrouter	NOC	13,997	2,175	M
19	wb2axip	NOC	14,287	2,244	M
20	dawson-double-multiplier	DSP	19,506	573	M
21	E15-verilog-axi	NOC	10,487	1,910	M
22	E24-secworks-sha1	SEC	9,957	1,525	S
23	G2R04-wb2axip-wbxbar	NOC	8,702	1,280	S
24	E05-openmsp430	CPU	7,339	685	S
25	ultra-ram-tester-full	MEM	5,853	945	S
26	litedram-controller-full	MEM	5,833	1,070	S
27	dawson-double-adder	DSP	5,377	530	S
28	ptp-td-leaf-full	IO	5,109	724	S
29	G1R01-wb2axip-axlite2wbsp	NOC	4,114	760	S
30	r8051-full	CPU	3,479	101	S
31	litespi-full	MEM	2,644	378	S
32	hardfloat-divsqrt-full	DSP	1,960	99	S
33	i2c-slave-axil-master-full	NOC	1,256	133	S
34	E27-satlab-rs	DSP	936	155	S
35	verilog-i2c	IO	749	71	S
36	spinal-spi-xdr-full	IO	579	43	S

Small (S): $s_i < 10^4$; Medium (M): $10^4 \leq s_i < 5 \times 10^4$; Large (L): $s_i \geq 5 \times 10^4$ (15/17/4 designs).
Domains: CPU; NOC: NoC/bus/DMA; SEC: cryptography; DSP: arithmetic/FEC; ML: machine learning; GPU: rasterization; MEM: memory/controllers; IO: peripherals.

Table D.2: Measured reference PPA and wall-clock budgets for the designs in Table D.1. Area and power are post-route values; $D_i^0 = 200\,\mathrm{ps} - s_i^{\mathrm{setup}}$ uses signed setup slack. Each flow requests 16 threads; $B_i = 3T_i^{\mathrm{ref}}$. Display values are rounded; experiments use the original seconds. These are reference measurements, not agent results.

No.	A_i^0 (μm^2)	P_i^0 (mW)	D_i^0 (ps)	DC (min)	PnR (min)	T_i^{ref} (min)	B_i (h)
1	141,530.334	953.945	175.066	232.24	218.07	450.94	22.547
2	18,620.176	235.384	2282.696	45.54	414.34	459.97	22.999
3	8,778.866	85.510	492.498	12.93	185.40	198.38	9.919
4	7,689.696	76.210	390.930	31.11	49.37	80.52	4.026
5	6,784.541	63.668	356.131	8.01	223.47	231.53	11.576
6	5,088.260	76.121	560.796	17.43	29.44	46.92	2.346
7	5,127.159	38.112	305.965	10.76	67.88	78.67	3.934
8	4,178.803	58.668	579.124	13.01	27.52	40.56	2.028
9	3,474.822	38.821	468.671	11.43	28.93	40.39	2.020
10	3,120.587	29.729	346.904	22.76	37.11	59.90	2.995
11	3,132.207	42.929	506.457	12.26	23.77	36.05	1.803
12	3,021.355	72.285	827.319	22.43	26.68	49.15	2.457
13	2,570.469	33.446	319.170	13.09	19.68	32.80	1.640
14	2,436.376	27.700	724.847	9.70	22.02	31.74	1.587
15	2,200.982	29.054	478.083	11.76	18.26	30.05	1.502
16	1,779.737	19.008	332.738	7.76	16.51	24.30	1.215
17	1,969.539	20.251	318.325	10.84	34.19	45.06	2.253
18	1,924.443	18.463	513.600	10.26	20.68	30.97	1.548
19	1,939.738	18.832	295.519	11.68	19.01	30.72	1.536
20	1,875.804	38.052	592.361	9.67	19.93	29.63	1.481
21	1,331.912	13.150	348.022	6.46	23.86	30.34	1.517
22	1,487.524	15.865	419.782	7.01	15.60	22.62	1.131
23	1,178.122	11.201	272.520	6.76	11.09	17.87	0.894
24	1,001.471	11.617	671.917	4.67	21.60	26.29	1.314
25	792.540	7.144	384.990	10.51	12.84	23.37	1.169
26	785.425	7.534	285.143	6.46	11.76	18.24	0.912
27	647.031	6.000	276.502	8.76	11.09	19.87	0.993
28	652.178	5.943	281.189	6.42	10.42	16.87	0.843
29	544.461	5.481	253.121	10.84	11.09	21.95	1.098
30	382.492	3.529	566.371	11.02	10.76	21.79	1.089
31	376.354	3.418	262.306	6.12	9.76	15.90	0.795
32	247.875	2.603	406.602	9.84	9.43	19.29	0.964
33	136.790	1.303	227.319	12.84	8.42	21.28	1.064
34	169.157	1.914	248.804	7.51	9.09	16.62	0.831
35	105.501	0.819	234.006	5.37	6.76	12.15	0.607
36	76.428	0.661	279.100	5.34	8.09	13.45	0.672

PnR includes placement, clock-tree synthesis, routing, extraction, and final reports. T_i^{ref} additionally includes orchestration within the successful attempt, but not earlier failed attempts or hidden replay. Hold, DRC, and connectivity are diagnostic fields under this release. Positive setup slack occurs for the AXI Virtual FIFO reference; a 5 GHz constraint is not a claim of timing closure or signoff.

Table D.3: Historical simulation coverage for the 36 designs, indexed as in Table D.1. Values are archived percentages, not new candidate verification results. N/A retains its original structural meaning; it is not 100%. The historical labels comprise 29 strict records and seven stage exceptions.

No.	Line	Condition	Toggle	Branch	FSM transition	Historical label
1	95.604	85.697	98.586	91.525	100.000	Strict
2	90.262	87.603	99.147	97.333	80.000	Strict
3	94.954	93.141	94.489	94.026	97.143	Strict
4	99.849	90.837	88.399	97.627	95.161	Strict
5	100.000	100.000	91.034	98.230	100.000	Strict
6	97.011	90.959	90.551	94.089	N/A	Strict
7	96.071	89.899	98.806	95.652	83.333	Strict
8	100.000	100.000	97.460	98.830	80.000	Strict
9	99.270	100.000	95.460	93.330	100.000	Strict
10	92.163	84.699	92.587	80.380	88.889	Exception
11	98.413	86.667	87.333	94.118	100.000	Strict
12	99.053	100.000	93.383	93.878	100.000	Strict
13	100.000	100.000	99.990	98.701	N/A	Strict
14	99.357	N/A	86.390	96.078	100.000	Strict
15	99.740	91.670	95.110	97.690	100.000	Strict
16	100.000	90.000	95.251	98.058	N/A	Strict
17	99.219	89.968	95.624	98.611	81.818	Strict
18	100.000	88.527	96.258	100.000	100.000	Strict
19	94.024	87.259	89.667	88.669	100.000	Strict
20	99.160	95.238	100.000	97.368	100.000	Strict
21	94.516	87.356	79.100	85.135	80.000	Exception
22	100.000	88.889	100.000	97.183	100.000	Strict
23	95.243	88.889	99.002	89.713	N/A	Strict
24	99.666	98.940	78.067	99.180	80.000	Exception
25	98.872	92.188	85.125	97.727	100.000	Strict
26	96.851	85.413	90.338	89.792	79.130	Exception
27	93.182	95.833	100.000	92.857	100.000	Strict
28	97.418	96.774	93.169	89.385	100.000	Strict
29	96.154	89.005	93.089	89.349	N/A	Strict
30	100.000	95.000	88.129	99.533	N/A	Strict
31	94.651	79.433	76.813	84.615	72.000	Exception
32	100.000	97.378	73.407	100.000	50.000	Exception[†]
33	97.368	93.056	90.832	92.000	81.818	Strict
34	100.000	80.952	97.321	100.000	40.000	Exception
35	97.213	86.957	96.833	94.118	98.387	Strict
36	99.798	100.000	92.808	99.531	100.000	Strict

The archived strict thresholds are Line $\geq 90\%$, Condition/Toggle/Branch $\geq 85\%$, and FSM transition $\geq 80\%$ where applicable. The historical stage exceptions retain their original labels; current reference implementation success does not revise these coverage records. [†]HardFloat coverage comes from the archived 8/24 simulation; the incompatible 3/3 generic count does not imply that the coverage used the wrong configuration. Coverage and fixed-workload replay do not prove universal functional equivalence.

D.4 PAIRED INCLUSION AND AVAILABILITY EXCLUSIONS

The analysis uses a complete paired set: design i enters comparison exactly when each of eight configurations has an assessed terminal nomination. The rule uses availability, not scores; validity failures still count as $U = 0$. Table D.4 lists the six design-level exclusions. The authors attribute campaign availability losses to service and network instability during long-running operation. The archived status categories also distinguish unimported, queued, running, and infrastructure-interrupted records; they are not recoded as functional failures.

This complete-case analysis is not a claim of a preregistered exclusion rule. The set excludes all four designs with reference flows above three hours, and retains only one of four Large designs. It therefore limits inference about the longest and largest tasks. The retained set has 14 Small, 15 Medium, and one Large design, with domain counts 7/6/7/3/3/2/1/1 in the order listed above. Its budgets range from 0.607 to 4.026 hours, with median 1.398 hours and sum 45.536 hours per configuration. All 27 assessed nominations outside this set remain in the supplementary record, without entering rankings.

Table D.4: Design-level exclusions from paired comparisons. All eight configurations must have an assessed nomination; exclusions depend on availability, not validity or quality. Assessed results on these designs remain in the supplementary archive.

Design	Size	T^{ref} (min)	B (h)	Assessed
`G2R02-fec-rs-dec`	Medium	231.53	11.58	6/8
`alex-axi-vfifo-full`	Large	450.94	22.55	0/8
`kemet-ptahconv-full`	Large	459.97	23.00	2/8
`litespi-full`	Small	15.90	0.79	7/8
`spinal-bmb-ooo-full`	Large	198.38	9.92	6/8
`spinal-mac-eth-full`	Medium	78.67	3.93	6/8

E SYSTEMS AND DEPLOYMENT IDENTITY

All configurations use the Claude Code harness. Table E.1 records the provider identifiers used in the campaign. A reported model name is an endpoint identity, not an independently verified weight or model snapshot. The comparisons describe these deployed systems, including the provider route and harness, rather than isolated base-model capability. Kimi is retained because it has complete paired assessments; reported service degradation does not justify removing a lower-scoring system after observing its outcomes. The anchor's Joint runs are reused for its authority contrasts.

Table E.1: Reported system identities and edit authorities. Each configuration has one run on each paired task; provider strings are retained verbatim.

System	Provider identifier	Authority
GPT-6 Sol	`azure.gpt-6-sol`	Joint
Claude Opus 5	`azure.claude-opus-5`	Joint
Qwen3.8-Max	`dashscope.qwen3.8-max`	Joint, RTL-only, Flow-only
GLM-5.3	`dashscope.glm-5.3`	Joint
Kimi-K3	`kimi-k3`	Joint
Grok-4.7	`grok-4.7`	Joint

The evaluated resource contract fixes $B_i = 3T_i^{\mathrm{ref}}$, up to two concurrent EDA jobs, and 16 requested threads per flow. Model response time, public-tool waits, edits, and registrations all consume the search allowance. The descriptive budget fractions use task-package allowances; they do not independently certify remote timeout enforcement. Exact harness builds and provider-side revisions are not inferred from model aliases. This bounds reproduction at the deployed-system level.

F Full Results and Cost

F.1 Per-design terminal utilities

Table F.1 gives all 240 utilities on the same 30 designs. Failure-inclusive mean utility is EVQ; all three assessed failures contribute zero. Design IDs join the reference inventory, whose PPA normalizes each nomination. The reported per-run values are rounded to four decimals; aggregate estimates retain the source report's precision.

Table F.1: Complete paired utility matrix. IDs join the 36-design inventory. G: GLM; P: GPT; Q: Qwen Joint; C: Claude; K: Kimi; X: Grok; R/F: Qwen RTL-/Flow-only. Bold marks a highest Joint utility above one; † marks a functional failure ($U = 0$).

ID	Design	G	P	Q	C	K	X	R	F
24	E05-openmsp430	1.0000	1.0000	**1.0174**	1.0141	1.0000	0.9761	1.0000	1.0068
10	E14-verilog-axi	1.1296	**1.1603**	1.0000	1.0221	1.0000	1.0000	0.9613	0.9897
21	E15-verilog-axi	**1.1087**	1.0496	1.0197	0.9613	1.0000	1.0220	1.2321	1.0033
18	E18-nocrouter	1.0039	1.1128	1.1130	**1.1257**	1.0000	1.1123	1.1170	1.0000
22	E24-secworks-sha1	**1.1590**	1.0173	1.0416	1.0746	1.0000	1.0000	1.1264	1.0028
12	E25-secworks-blake2s	**1.2018**	1.0000	1.1898	1.0333	1.0358	1.0974	1.0786	1.0338
34	E27-satlab-rs	0.9835	1.0035	0.9833	**1.0246**	1.0000	0.9952	1.0000	0.9733
29	G1R01-wb2axip-axlite2wbsp	0.9980	1.0375	1.0138	**1.0400**	0.9934	0.9965	1.0242	1.0081
23	G2R04-wb2axip-wbxbar	0.9956	**1.0072**	1.0000	1.0065	1.0000	1.0000	0.9941	1.0000
16	G2R05-fec-golay24-dec	**1.5680**	1.5625	1.0000	0.7908	1.0000	0†	1.3946	0.9870
4	G2R10-rpeters-rasterizer	1.0220	1.0286	1.0136	**1.1066**	1.0000	1.0000	0†	1.0000
17	G2R11-ultraembedded-ddr3-axi	1.0130	1.0153	**1.0850**	1.0000	1.0000	0.9697	1.0147	1.0021
13	G2R22-freecores-des	1.0136	1.0129	1.0120	**1.0324**	0.9555	0.9416	0.9956	1.0000
27	dawson-double-adder	1.1221	**1.1773**	0.9591	1.1099	1.0000	1.0000	1.0956	0.9784
20	dawson-double-multiplier	**1.3458**	1.2394	1.0261	1.0419	0.9814	0.9479	1.0631	1.0140
32	hardfloat-divsqrt-full	1.0431	1.0467	1.0314	**1.0633**	1.0024	0.9159	1.0003	1.0487
33	i2c-slave-axil-master-full	**1.1706**	1.0622	1.0164	1.0768	0.8469	1.0653	1.0337	0.9775
26	litedram-controller-full	0.9385	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
28	ptp-td-leaf-full	0.8404	0.9761	0.9982	1.0000	**1.0083**	1.0051	1.0189	0.9926
30	r8051-full	1.0269	**1.0281**	1.0000	1.0173	0.9204	1.0000	1.0512	1.0000
9	secworks-aes	1.0279	1.0369	1.0405	**1.0539**	1.0040	1.0000	1.0187	1.0222
14	secworks-chacha	1.0000	**1.0232**	1.0076	0.8916	0.8881	0.6807	0.8881	1.0204
15	secworks-sha256	1.2481	1.1886	1.0778	**1.2546**	1.0000	1.0320	1.1773	1.0396
8	secworks-sha512	1.0324	1.0066	**1.1226**	1.0675	1.0000	1.0000	1.0364	1.0025
36	spinal-spi-xdr-full	1.0000	1.0147	**1.0697**	0.9907	1.0000	1.0566	1.0000	1.0163
6	tachyum-fma-full	1.0195	1.0000	1.0000	**1.0256**	0.9614	1.0000	1.0000	1.0000
11	tensorwright-convolution-engine	1.0538	**1.2044**	1.1043	1.1431	1.0000	1.0238	1.1050	1.1031
25	ultra-ram-tester-full	0.9895	1.0214	**1.0241**	1.0158	1.0000	0†	1.0596	1.0000
35	verilog-i2c	**1.0186**	1.0133	1.0000	0.9700	1.0000	0.9806	1.0000	1.0301
19	wb2axip	1.0092	1.0192	**1.0204**	1.0016	1.0000	1.0000	1.0000	0.9927
	EVQ	1.0694	1.0688	1.0329	1.0319	0.9866	0.9273	1.0162	1.0082

F.2 VALID QUALITY, TRADE-OFFS, AND IMPROVEMENT MARGINS

Table F.2 distinguishes conditional mean, geometric mean, median, and range from failure-inclusive EVQ. Of 133 valid $Q > 1$ nominations, 71 worsen at least one PPA component. Neither aggregate quality nor a fast delay proxy implies an across-component improvement. There are 72 unchanged nominations and 168 changed nominations; unchanged outputs never receive $J = 1$, including those with slightly different physical measurements.

Table F.2: Quality distribution and threshold sensitivity. Quality statistics condition on validity. Both improvement rates retain all 30 tasks.

System	Mean Q	GM Q	Median Q	Min	Max	$Q > 1/\simeq 1/< 1$	$\mathrm{VIR}_{.01}$	$\mathrm{VIR}_{.05}$
GLM-5.3	1.0694	1.0621	1.0207	0.8404	1.5680	21/3/6	19/30	10/30
GPT-6 Sol	1.0688	1.0636	1.0223	0.9761	1.5625	25/4/1	22/30	8/30
Qwen3.8-Max	1.0329	1.0318	1.0169	0.9591	1.1898	20/7/3	19/30	7/30
Claude Opus 5	1.0319	1.0288	1.0251	0.7908	1.2546	22/3/5	20/30	10/30
Kimi-K3	0.9866	0.9858	1.0000	0.8469	1.0358	4/19/7	1/30	0/30
Grok-4.7	0.9935	0.9904	1.0000	0.6807	1.1123	8/11/9	7/30	4/30
Qwen RTL-only	1.0513	1.0475	1.0189	0.8881	1.3946	18/7/4	17/30	11/30
Qwen Flow-only	1.0082	1.0079	1.0011	0.9733	1.1031	15/8/7	9/30	1/30

Table F.3 separates all-component non-worsening from scalar quality gains and records reference retention.

Table F.3: Component trade-offs and reference retention on the paired set. A gain in this table means valid $Q > 1$, without an improvement margin. All-component non-worsening also requires at least one strict improvement.

System	Gain, component worse	All components no worse	Unchanged	Unchanged $Q \neq 1$
GLM-5.3	13	8	4	1
GPT-6 Sol	12	13	4	0
Qwen3.8-Max	9	11	8	2
Claude Opus 5	10	12	3	0
Kimi-K3	4	0	22	3
Grok-4.7	4	4	14	3
Qwen RTL-only	9	9	7	1
Qwen Flow-only	10	5	10	2

F.3 TASK-PAIRED CONTRASTS

Table F.4 includes all 15 Joint-system pairs and the three anchor-authority pairs. Counts and magnitudes can disagree: GPT wins on 18 designs against GLM and loses on 11, but the mean difference is -0.0006. Claude wins 16 against GLM but has lower mean utility. A task win does not measure the size of the physical gain or the cost of a failed nomination. No multiplicity-corrected superiority claim follows from these intervals.

Table F.4: Task-paired contrasts retained from the report. Differences are A minus B; intervals are descriptive and not multiplicity-corrected.

A	B	ΔEVQ	95% CI	Win/tie/loss
Kimi-K3	Qwen3.8-Max	-0.0463	[-0.0657, -0.0282]	3/5/22
Kimi-K3	GPT-6 Sol	-0.0823	[-0.1293, -0.0445]	2/2/26
Kimi-K3	Grok-4.7	0.0593	[-0.0180, 0.1649]	12/8/10
Kimi-K3	Claude Opus 5	-0.0453	[-0.0750, -0.0153]	6/2/22
Kimi-K3	GLM-5.3	-0.0828	[-0.1369, -0.0357]	5/2/23
Qwen3.8-Max	GPT-6 Sol	-0.0359	[-0.0864, 0.0046]	11/2/17
Qwen3.8-Max	Grok-4.7	0.1056	[0.0296, 0.2079]	20/5/5
Qwen3.8-Max	Claude Opus 5	0.0011	[-0.0264, 0.0292]	11/1/18
Qwen3.8-Max	GLM-5.3	-0.0365	[-0.0904, 0.0087]	14/0/16
Qwen3.8-Max	Qwen RTL-only	0.0167	[-0.0468, 0.1017]	13/3/14
Qwen3.8-Max	Qwen Flow-only	0.0247	[0.0109, 0.0406]	22/4/4
GPT-6 Sol	Grok-4.7	0.1416	[0.0437, 0.2776]	24/2/4
GPT-6 Sol	Claude Opus 5	0.0370	[-0.0044, 0.0997]	14/1/15
GPT-6 Sol	GLM-5.3	-0.0006	[-0.0276, 0.0249]	18/1/11
Grok-4.7	Claude Opus 5	-0.1046	[-0.1978, -0.0365]	5/1/24
Grok-4.7	GLM-5.3	-0.1421	[-0.2789, -0.0439]	6/0/24
Claude Opus 5	GLM-5.3	-0.0376	[-0.1046, 0.0123]	16/0/14
Qwen RTL-only	Qwen Flow-only	0.0081	[-0.0771, 0.0707]	18/2/10

F.4 Edit use and implementation provenance

Table F.5 partitions nominations by the actual changed objects, not merely the authority granted. Of 168 changed nominations, 103 reuse matched implementation outputs and 65 use reconstruction. The associated regression counts are 8 and 27; gain counts above 5% are 40 and 11. The association can reflect agent behavior, design difficulty, and availability of eligible outputs. Physical-output reuse alone is not a validity test. Table F.6 gives the per-configuration breakdown by evidence path.

Table F.5: Edit use and terminal outcomes. Both/RTL/flow/neither partition all 30 nominations. Gain requires $V = 1$, a changed artifact, and $Q > 1.05$; regression requires $V = 1$, a changed artifact, and $Q < 1$.

System	Both	RTL	Flow	Neither	Gain	Regress.	Invalid
GLM-5.3	10	9	7	4	10	5	0
GPT-6 Sol	4	21	1	4	8	1	0
Qwen3.8-Max	6	7	9	8	7	1	0
Claude Opus 5	12	6	9	3	10	5	0
Kimi-K3	2	1	5	22	0	6	0
Grok-4.7	12	3	1	14	4	7	2
Qwen RTL-only	0	23	0	7	11	4	1
Qwen Flow-only	0	0	20	10	1	6	0

Table F.6: Changed nominations by physical-evidence path. Every paired count is reused/rebuilt. These paths are observational, not randomized interventions.

System	Changed	Valid regressions	Verified gains $> 5\%$
GLM-5.3	20/6	1/4	9/1
GPT-6 Sol	19/7	1/0	7/1
Qwen3.8-Max	14/8	1/0	4/3
Claude Opus 5	18/9	0/5	9/1
Kimi-K3	0/8	0/6	0/0
Grok-4.7	8/8	2/5	3/1
Qwen RTL-only	13/10	2/2	7/4
Qwen Flow-only	11/9	1/5	1/0

F.5 Search and assessment cost

Table F.7 reports wall-clock time. Search totals sum the 30 task runs in each configuration; parallel scheduling means that these totals are not elapsed campaign duration. Assessment totals include the last recorded attempt and its waits, not the unarchived cost of earlier attempts. Search EDA CPU time, token use, cache use, and provider charges are outside this accounting; multiplying wall-clock by the thread request would invent CPU usage. A short assessment can reflect reuse of an already produced physical implementation, not a cheap end-to-end flow.

F.6 Reference-duration strata

Table F.8 uses reference-flow time to form matched descriptive strata. The longest stratum contains only the rasterizer: Qwen RTL-only fails there, while the other configurations are valid. Short-task Grok validity is 14/16. Other stratum validity rates are one. These counts describe the paired sample and do not establish how performance scales with duration.

Table F.7: Wall-clock accounting on the paired set. Search hours sum task time, not parallel campaign duration. Assessment is the last recorded attempt including its waits; it excludes earlier attempts. Neither is CPU-hours or token cost.

System	Search (h)	Med. (min)	P90 (min)	Assess. (h)	Med. (min)	Reuse/rebuild
GLM-5.3	42.1724	70.7792	136.8690	11.6473	0.6290	20/10
GPT-6 Sol	43.4694	76.0079	129.7043	13.2012	0.7638	19/11
Qwen3.8-Max	41.9496	71.0787	132.0788	18.9032	28.2671	14/16
Claude Opus 5	42.4124	75.6761	130.2894	14.9105	0.7840	18/12
Kimi-K3	45.0077	83.8771	141.4147	38.7083	74.0031	0/30
Grok-4.7	29.1563	49.1486	122.3604	27.2289	59.1925	8/22
Qwen RTL-only	41.5411	75.7219	132.3884	19.2905	27.1459	13/17
Qwen Flow-only	42.2596	70.2751	134.7470	25.7917	66.4406	11/19

Table F.8: EVQ by reference-flow duration; each column uses the same designs across configurations. The longest stratum has one design and cannot support a scaling conclusion.

System	$\leq$ 30 min ($n = 16$)	30–60 min ($n = 13$)	$>$ 60 min ($n = 1$)
GLM-5.3	1.0750	1.0663	1.0220
GPT-6 Sol	1.0754	1.0638	1.0286
Qwen3.8-Max	1.0113	1.0610	1.0136
Claude Opus 5	1.0148	1.0471	1.1066
Kimi-K3	0.9845	0.9881	1.0000
Grok-4.7	0.8712	0.9907	1.0000
Qwen RTL-only	1.0539	1.0481	0.0000
Qwen Flow-only	1.0022	1.0161	1.0000

F.7 ADDITIONAL ASSESSED RECORDS OUTSIDE PAIRED COMPARISON

Table F.9 retains 27 assessed records on five of the six excluded designs. They do not share the same available configurations and therefore supply provenance rather than a second ranking or an estimate of long-task headroom. No unavailable result is imputed. The GLM/BMB semantic warning is retained explicitly instead of using that result as a mechanism or performance claim.

Table F.9: The 27 additional assessed nominations, retained for provenance only. They use different available system subsets and are excluded from all cross-system rankings. The GLM BMB record has a reuse-versus-requested-flow semantic warning; its reported value is retained without treating it as confirmed optimization evidence.

Design	System	V	U	PPA path
G2R02-fec-rs-dec	GLM-5.3	1	1.2684	reused
G2R02-fec-rs-dec	GPT-6 Sol	1	0.9697	reused
G2R02-fec-rs-dec	Qwen3.8-Max	1	0.9948	reused
G2R02-fec-rs-dec	Claude Opus 5	1	1.0095	rebuilt
G2R02-fec-rs-dec	Grok-4.7	1	0.9571	reused
G2R02-fec-rs-dec	Qwen RTL-only	1	1.0380	reused
kemet-ptahconv-full	GLM-5.3	1	1.0000	rebuilt
kemet-ptahconv-full	Grok-4.7	1	1.1683	reused
litespi-full	GLM-5.3	1	0.9871	rebuilt
litespi-full	GPT-6 Sol	1	1.0049	reused
litespi-full	Qwen3.8-Max	1	1.0049	reused
litespi-full	Claude Opus 5	1	0.9593	rebuilt
litespi-full	Kimi-K3	1	1.0023	rebuilt
litespi-full	Qwen RTL-only	1	0.9949	rebuilt
litespi-full	Qwen Flow-only	1	1.0095	rebuilt
spinal-bmb-ooo-full	GLM-5.3	1	1.5604	reused
spinal-bmb-ooo-full	GPT-6 Sol	1	1.0570	reused
spinal-bmb-ooo-full	Qwen3.8-Max	1	1.0431	reused
spinal-bmb-ooo-full	Claude Opus 5	1	1.1968	reused
spinal-bmb-ooo-full	Grok-4.7	1	1.0643	rebuilt
spinal-bmb-ooo-full	Qwen RTL-only	1	1.0852	reused
spinal-mac-eth-full	GLM-5.3	1	1.0661	reused
spinal-mac-eth-full	GPT-6 Sol	1	1.0394	reused
spinal-mac-eth-full	Claude Opus 5	1	1.5150	reused
spinal-mac-eth-full	Kimi-K3	1	1.1784	rebuilt
spinal-mac-eth-full	Grok-4.7	1	0.9645	reused
spinal-mac-eth-full	Qwen RTL-only	1	1.0322	reused

G Statistics, Budgets, and Registration Semantics

G.1 Task-paired uncertainty

R, EVQ, and VIR_ϵ use all $N = 30$ paired tasks, including invalid nominations. Component geometric means and Q statistics condition on $V = 1$ and disclose the valid count. For each contrast, the difference is computed within design before averaging. The source analysis uses 10,000 bootstrap resamples of design IDs, seed 20260926, and percentile 95% intervals. Each sampled design carries all configurations together. With $S = 1$, these intervals measure task variation, not run-to-run stability or uncertainty over all possible engineering tasks. If every observed $V = 1$, the empirical bootstrap interval is $[1, 1]$; that is not proof of a population success probability of one. The analysis is descriptive and exploratory, with no multiple-comparison correction.

G.2 Improvement thresholds and unchanged-reference variation

The primary improvement margin is $\epsilon = 0.05$; 0.01 is a sensitivity analysis (Table F.2). Both require a changed artifact and $V = 1$. Of 72 unchanged nominations, 60 reproduce $Q = 1$ at report precision. The remaining values are 0.9833, 0.9927, 0.9934, 0.9952, 0.9965, 0.9980, 0.9982, 1.0003, 1.0024, 1.0051, 1.0068, and 1.0083. Thus the maximum observed absolute deviation is 0.0167. This motivates distinguishing small numerical changes from material improvements, but these are not controlled repeats: implementation and measurement conditions can also contribute. The margins are declared analysis choices, not a preregistered calibration, a guarantee against measurement variation, or a significance threshold. Unchanged measurements never count as improvement at either margin.

G.3 Registration and terminal delivery

A registration is an immutable RTL–flow snapshot with completion time. The reference is registered at time zero. At budget fraction k, the latest registration completed by kB_i determines the nomination; without a new registration, it carries forward. Restoring a workspace does not change this ledger. A later invalid nomination is not replaced by a better earlier one. This study evaluates $k = 1$. Checkpoints at $0.25, 0.5, 0.75$ in the conceptual figures describe the same rule; they are not empirical progress curves or independent experiments with smaller announced budgets.

G.4 Infrastructure and denominators

The complete-pair rule in Appendix D.4 treats absence of an assessed result separately from an assessed failure. Excluding a design removes it for every configuration. This preserves pairing but selects on availability, which correlates with duration and size. Neither bootstrapping nor inclusion of the 27 auxiliary records removes that selection effect.

H Failure Attribution and Case Notes

H.1 Behavioral failures

Table H.1 lists all three functional failures. Each has physical outputs, illustrating why the functional gate precedes quality. Two reuse public physical outputs; the third is reconstructed locally. Even attractive ungated PPA cannot make their utility positive.

Table H.1: The three assessed functional failures; complete physical outputs do not establish preserved behavior.

System	Design	Failure	V^{impl}	U	PPA path
Grok-4.7	`G2R05-fec-golay24-dec`	Value mismatch	1	0	reused
Qwen RTL-only	`G2R10-rpeters-rasterizer`	Value mismatch	1	0	reused
Grok-4.7	`ultra-ram-tester-full`	Value mismatch	1	0	rebuilt

H.2 VALID BUT WORSE IMPLEMENTATIONS

The 35 valid changed regressions are distinct from functional failures. Table H.2 lists all 12 whose $Q < 0.95$. Eleven use local reconstruction and one uses reuse. This motivates artifact-matched validation as a delivery discipline, without assigning causality to the assessment path.

Table H.2: All twelve valid changed nominations with $Q < 0.95$. Eleven used local reconstruction; this is an association with the recorded assessment path, not proof that the agent never ran a matching flow.

System	Design	A/A^0	P/P^0	D/D^0	Q	PPA path
Claude Opus 5	`G2R05-fec-golay24-dec`	1.555	1.615	0.806	0.791	rebuilt
Grok-4.7	`G2R22-freecores-des`	1.094	1.101	0.995	0.942	rebuilt
Grok-4.7	`dawson-double-multiplier`	1.214	1.165	0.830	0.948	rebuilt
Grok-4.7	`hardfloat-divsqrt-full`	1.094	1.131	1.051	0.916	rebuilt
Kimi-K3	`i2c-slave-axil-master-full`	1.272	1.391	0.930	0.847	rebuilt
GLM-5.3	`litedram-controller-full`	1.122	1.116	0.965	0.939	reused
GLM-5.3	`ptp-td-leaf-full`	1.253	1.436	0.936	0.840	rebuilt
Kimi-K3	`r8051-full`	1.109	1.166	0.992	0.920	rebuilt
Claude Opus 5	`secworks-chacha`	0.904	1.831	0.853	0.892	rebuilt
Kimi-K3	`secworks-chacha`	0.909	1.848	0.850	0.888	rebuilt
Grok-4.7	`secworks-chacha`	1.261	2.934	0.857	0.681	rebuilt
Qwen RTL-only	`secworks-chacha`	0.909	1.848	0.850	0.888	rebuilt

H.3 CONTRASTING DESIGNS

On Golay decoding, GPT and GLM obtain areas of 942.6 and 939.3 $\mu\mathrm{m}^2$, about 47% below reference, and roughly 43% lower power; Claude's valid nomination increases area by about 55%. The same fixed computation therefore exposes both useful freedom and harmful decisions. On ChaCha, Kimi and Qwen RTL-only report the same measured PPA triple (2215.577 $\mu\mathrm{m}^2$, 51.18 mW, 615.86 ps). Equality of these measurements does not prove identical code. Both reduce area and delay but increase power to about 1.85 times reference, yielding $Q = 0.8881$. The observed trade-off, not an inferred internal strategy, explains their lower utility.

H.4 STOPPING AND REFERENCE RETENTION

Two paired runs have an agent-error stopping label: Qwen Joint and GLM Joint. Both retain a valid reference nomination. This is why stopping labels are separate from validity and why an abnormal exit is not automatically $U = 0$. An unchanged nomination establishes retention; it does not establish that the system deliberately recovered an earlier candidate.

H.5 PHYSICAL DIAGNOSTICS

All 240 paired records have failing setup diagnostics. Hold reports 9 pass, 199 fail, and 32 unassessed diagnostic entries; design-rule checks report 213 pass and 27 unassessed entries; connectivity reports 240 pass. An unassessed diagnostic is neither a pass nor a functional failure. These statuses describe the relaxed profile, not completed timing or physical signoff. Their treatment is fixed independently of the candidate's score.

I DISCUSSION AND LIMITATIONS

What reliable improvement adds to agent evaluation. A fixed executable reference makes the behavioral target precise without prescribing a unique implementation. Continuous quality distinguishes retention from gain; protected checks separate public hints from acceptance; and registration binds evidence to what the agent delivers. P1–P4 explain why the task creates meaningful decisions. D1–D3 explain which measurements are needed to distinguish their outcomes. The empirical separation between validity, physical improvement, and budget use supports this joint view, without identifying a model's private reasoning process.

Interpretation of the paired study. The primary sample has 30 of 36 designs and excludes the four longest reference flows. It represents eight domains but only one Large design; results do not establish competence on the largest tasks. One run per condition cannot determine stochastic reliability. Hardware costs and tool access also constrain reproduction. The fixed harness, provider adapters, and deployment conditions are part of each system. Authority effects are measured for one anchor and do not establish a universal interaction law. The 103 reused physical outputs reduce recomputation but share physical execution with public search; independence rests on identity matching and protected assessment, with the driver-audit limits of Appendix C.4.

Verification and measurement limits. Finite RTL and netlist replay can miss behavior outside the workload and is not a universal equivalence proof. The relaxed physical profile records hold diagnostics and accepts negative setup slack. Vectorless power is a controlled estimate, not silicon workload power. The mixed-source size index is descriptive. Improvement margins are retrospective analysis choices with sensitivity reporting, rather than controlled noise calibration. Wall-clock cost excludes model-token and actual CPU accounting. Open solutions reduce dependence on a canonical patch, but do not eliminate training contamination or physical performance limits; new task lineages and environments require separately frozen releases.

Value for chip development and relation to self-improvement. A verified PPA change at fixed computation identifies potentially useful agent-produced implementation decisions. SILICA measures this external engineering step, not saved human labor, fabrication yield, silicon speed, or end-to-end LLM inference throughput. Production qualification has additional obligations. Improving computing infrastructure could be part of a broader feedback loop in which AI helps improve the resources supporting later AI systems. Recursive self-improvement additionally requires that feedback to improve the agent's subsequent capabilities and support further improvement. This study evaluates the artifact step and does not claim to close that loop.

J Positioning Against Prior Work

Table J.1 summarizes the scope supported by the cited sources. It does not infer missing features from an abstract or treat every previous agent benchmark as a short, cheap question-answer task. Independent checks, continuous quality, budgeted interaction, and timed evaluation all have precedents. SILICA combines them around reference-preserving physical improvement and the candidate an agent actually delivers.

Table J.1: Relevant precedents and their relation to SILICA's measurement focus. Entries describe scope rather than asserting feature absence.

Work	Established scope and relation
SWE-bench / SWE-agent	Executable repository outcomes and the effect of the agent interface (Jimenez et al., 2024; Yang et al., 2024).
RE-Bench / METR	Extended work, budget-dependent performance, and reliability over task duration (Wijk et al., 2025; Kwa et al., 2025).
RTLScout / ARES	Joint RTL/synthesis optimization and the balance between PPA gain and model cost (Arnold et al., 2026; Cuyckens et al., 2026).
CLOSER-Bench	Budgeted cross-stage closure with progress, cost, and recovery (Zhou et al., 2026).
PostEDA-Bench	PPA convergence and physical-rule repair (Liu et al., 2026).
TicTacBench	RTL-level timing closure evaluated after placement and routing (Wang et al., 2026).
EEBench	Simulation-based electronic design with protected engineering checks (atopile, 2026).
SILICA	Fixed reference behavior and open implementation choices; artifact-matched post-route evidence; explicit nominations; matched authority and terminal delivery.