SGANALOG: AN END-TO-END CIRCUIT BENCHMARK FROM OPEN-SOURCE SILICON TAPEOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing analog integrated circuit design benchmarks make two questions hard to answer: whether a model has learned transferable circuit skills rather than recalled familiar examples, and whether its output works under defined process and test conditions. We introduce SGAnalog, a benchmark built from human-designed, open-source circuits associated with Tiny Tapeout manufacturing shuttles. The collection holds 425 top-level designs spanning 273 distinct circuit topologies; the fixed evaluation sets hold 66 transcription tasks and 17 sizing tasks.

Every source is retrieved at the revision recorded for its shuttle submission and processed in a fixed containerized environment. The pipeline exports each schematic image and its SPICE netlist from the same source file, giving transcription an exact structural reference. Commit dates support model-specific training-cutoff analysis, while author testbenches provide the simulation context for sizing.

The benchmark evaluates schematic-to-netlist transcription and device sizing. Across seven models on the fixed transcription set, the strongest model reaches 56.1% exact graph isomorphism, and six of seven models drop sharply from the small to the medium tier. Removing author-chosen labels lowers exact matches for every model above floor, with larger drops for stronger models, while parameter accuracy holds, indicating that labels aid connectivity tracing. On the sizing set a different model leads. It converges on every proposal and reaches 91.2 out of 100 against the human reference (85.1 on the ten tasks with a scalar metric), while the two newest Claude models refuse 4 and 11 of the same prompts they transcribe without objection; a proposal without sizes scores zero. The two tasks produce different model rankings, exposing distinct visual and design capabilities and, in one family, a policy rather than capability limit.

1 INTRODUCTION

Analog integrated circuit design gives model outputs a concrete standard: the topology must be correct, and the sized circuit must work under defined process and test conditions. Evaluating these capabilities requires more than a collection of familiar circuit examples. The benchmark must also expose possible training overlap and provide an executable reference for function. Existing resources rarely provide both.

Separating capability from recall. Many analog benchmarks draw from canonical textbooks (Razavi, 2017; Gray et al., 2009), published circuit figures (Li et al., 2024; Gao et al., 2025), or synthetic enumeration (Dong et al., 2023). All three source families are public, so prior model exposure is difficult to rule out. Recent systems already approach saturation on the solved-task count of AnalogCoder (Lai et al., 2025; 2026). Without dated provenance and an exposure protocol, a high score does not reveal whether the model generalized or recalled a familiar design.

Verifying function. Large datasets often provide topology without device sizes, testbenches, or a process design kit (PDK). Simulator-backed benchmarks provide the missing context but cover a small set of base circuits and expand it through variants. The field therefore lacks a broad set of human designs that supports both structural comparison and execution under the original test conditions.

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.

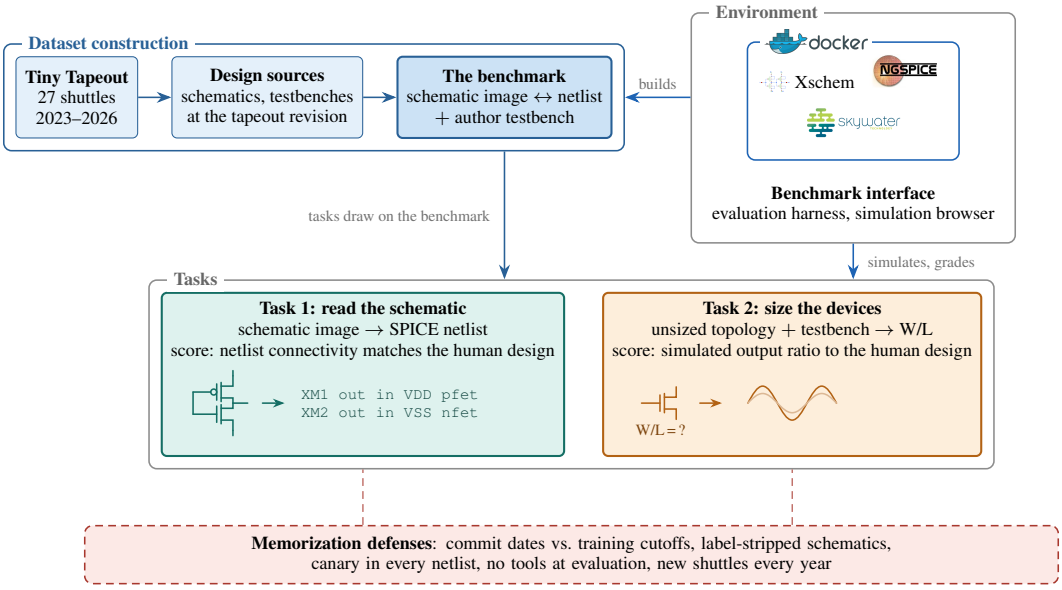


Figure 1: Benchmark overview. Versioned Tiny Tapeout sources are converted into matched schematic images and netlists, with author testbenches retained for simulation. Transcription is scored by device–net graph isomorphism, the the check that the netlist wires the same circuit; sizing is scored by simulation; and exposure is probed with commit-aware controls.

Why open-source tapeouts. Tiny Tapeout (Venn et al.) provides a different foundation. Its machine-readable index links each shuttle submission to an exact repository revision. The analog repositories collectively contain transistor-level schematics, netlists, PDK context, and author testbenches. The recorded revisions make the pipeline reproducible, shuttle dates support model-specific cutoff analysis, and new shuttles provide a continuing refresh path. Verified open-source licenses allow the resulting artifacts to be redistributed. The collection spans amplifiers and comparators; analog-to-digital, time-to-digital, and digital-to-analog converter blocks; oscillators and timing circuits; power, bias, and reference circuits; RF front ends and filters; and sensing interfaces. It also includes smaller passive, switching, and memory-related subcircuits. The fixed evaluation sets are eligibility-filtered subsets of this broader collection rather than equal-sized samples of every family.

Scope. The benchmark targets schematic-level reasoning, not layout or subjective design style. Transcription measures whether a model can recover circuit structure from an image. Sizing measures whether model-selected device dimensions preserve simulated behavior in the author's testbench. Topology generation remains future work because it requires a uniform design brief and automatic testbench alignment.

Contributions. We call the benchmark SGAnalog, for silicon-grade analog: every reference design is a shuttle tapeout submission, and every score is measured against that taped-out design's own simulated behavior.[1]

1. A license-verified, commit-pinned benchmark from open-source Tiny Tapeout submissions, built with a deterministic containerized pipeline, grouping of duplicate circuit topologies, and aligned schematic images and netlists generated from the same source (§3).
2. Two end-to-end evaluations, source file to scored simulation: graph-scored schematic transcription and simulator-scored device sizing against the human reference under the same PDK and author testbench (§4).

[1]Anonymous review artifacts: code and data at https://anonymous.4open.science/r/sganalog-anon-7EF6/; a companion page with precomputed simulations of ten verified pairs at https://admirable-sunburst-858a7d.netlify.app/.

3. An exposure-aware protocol (§5): commit-date cutoff splits and blinded schematic images carry evidence in this paper, while closed-book evaluation, release canaries, and shuttle refreshes are infrastructure for the audits the collection keeps enabling.
4. A seven-model baseline showing a broad small-to-medium transcription drop and different rankings across transcription and sizing (§4).

2 Related work

Table 1: Analog benchmark landscape. Sized = devices carry numeric sizes. TB = author testbenches are included. PDK = an open process design kit simulation flow is included. "part" marks partial coverage. Audit = the benchmark ships an exposure audit. Scale follows the unit reported by each source and is therefore not directly comparable.

Benchmark	Scale	Provenance	Sized	TB	PDK	Audit
AnalogGenie (Gao et al., 2025)	3,350 topo.	textbooks + papers	no	no	no	no
CktGNN / OCB (Dong et al., 2023)	10k op-amps	synthetic	part	part	no	no
AICircuit (Mehradfar et al., 2024)	9 topo. / 361k pts	authored	yes	yes	part	no
OSIRIS (Chiari et al., 2026)	5 templ. / 87.1k layouts	authored	yes	yes	yes	no
MAGICAL-CIRCUITS (Xu et al., 2019)	26 blocks	UT Austin silicon	yes	part	no	no
AnalogCoder (Lai et al., 2025)	24 tasks	authored	part	yes	no	no
AMSnet 1.0 / 2.0 (Tao et al., 2024)	734 / 2,686	textbook scans	no	no	no	no
Masala-CHAI (Bhandari et al., 2025)	7,500	textbook scans	no	no	no	no
AnalogGym (Li et al., 2024)	30 topo.	published papers	part	part	part	no
SGAnalog (ours)	**273 distinct topologies**	**Tiny Tapeout repos**	**yes**	**part**	**yes**	**yes**

Prior image-based sets such as AMSnet (Tao et al., 2024) and Masala-CHAI (Bhandari et al., 2025) pair scanned textbook figures with netlists produced by manual or model-assisted extraction. In this benchmark, every eligible image–netlist pair is generated from one schematic source, avoiding a separate annotation step.

AnalogGym (Li et al., 2024) is the closest simulator-backed benchmark in intent. Its amplifier topologies come from published literature, whereas our sizing reference is the submission author's design evaluated under the same PDK and testbench as the model. Concurrent work, NotSoTiny (Ghorab et al., 2025), uses the digital side of Tiny Tapeout (Venn et al.) for RTL code generation; our benchmark targets the analog artifacts and their simulation context.

3 Dataset construction

Harvest funnel. Table 2 walks the index down to the fixed task sets; Appendix B holds the full triage, licensing, and composition inventory. Blobless clones triage every candidate by file tree and license without downloading design data. Licensing is verified from the license texts as fetched at the pinned commits rather than assumed, and design source is fetched at the commit the shuttle index records, never at HEAD.

Circuit identification. Counting circuits naively fails in both directions: counting every schematic file overcounts (symbols, testbenches, sub-blocks); counting only hierarchy tops undercounts, because the standard xschem (Schippers) idiom places the interesting circuit under its testbench. We classify by role instead. A *testbench* contains stimulus and simulation control, or matches the naming conventions for one. A *circuit* is any non-testbench schematic. A *top-level design* is a circuit that no other non-testbench schematic instantiates. Table 5 (Appendix B) gives the resulting composition by role, tier, and era.

Redundancy. Repository counts overstate the diversity of the tapeout-derived collection because blocks are resubmitted and forks preserve earlier circuits. We compute a strict Weisfeiler–Lehman hash of each top-level device–net bipartite graph after removing net names. Repeated pad models, inverters, tutorial circuits, and R-2R blocks account for most shared hashes. The resulting candidate

Table 2: Dataset spine: from the shuttle index to the fixed task sets. Appendix B gives the full triage, licensing, composition, and redundancy inventory.

From index to task sets	n
Index projects (27 shuttles)	4,572
Entries with analog pins	237
Repositories contributing circuits	125
Schematics fetched at pinned commits	1,448
Circuits	933
Top-level designs	425
Unique device–net graphs	273
Fixed tasks: transcription / sizing	66 / 17

duplicate groups are released for split control. Task 1 scoring uses explicit graph isomorphism rather than hash equality.

Reproducibility contract. Every stage, from harvest through simulation, executes inside one pinned container image (IIC-OSIC-TOOLS (Pretl & Zachl, 2026), digest `sha256:ae380e2b...df38`), which pins ngspice (Vogt et al.) and supplies the open PDKs; the host requires only Docker. Duplicate file stems and archived copies can make hierarchy resolution depend on filesystem order, so candidates resolve by depth and then full path. Classification tables from two complete runs are byte-identical, and simulation tables regenerate identically up to their per-run timing columns.

Task-set construction. A curated core spans circuit classes, difficulty tiers, and eras. For transcription, leaf schematics with every symbol resolved are eligible; deterministic round-robin sampling over era and tier fixes 66 tasks. For sizing, we audit 412 author-testbench runs after deterministic include-path repair, without changing circuits or stimuli. A run is eligible only if it completes within 600 seconds, contains real device cards, and resolves the schematic hierarchy; 190 runs meet this gate, browsable through the release's simulation interface (Appendix E). End-to-end device-under-test mapping and the primitive-only hierarchy requirement then yield 17 tasks. Both sets remain fixed across models. The 66 tasks span 53 distinct device–net graphs: ten duplicate groups, inherited from collection-level redundancy, hold the other 13 members. Restricting every model to the 53 representatives moves no score by more than 3.6 points and changes no ordering (Appendix D). Hierarchical sizing tasks await the measurement layer described in Section 6.

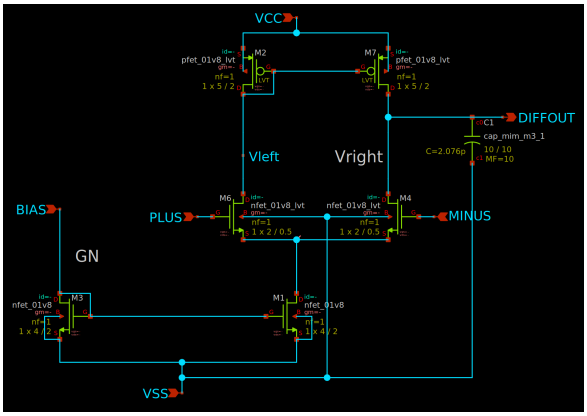
4 Tasks and Evaluation

The benchmark defines two complementary tasks. Transcription tests schematic reading; sizing tests device choices under simulation.

Task 1: schematic-to-netlist transcription. Input: the schematic image (Figure 2). Output: a netlist. Scoring is by graph isomorphism, never by text match. We build a device-net bipartite multigraph from each netlist and test it for isomorphism against the ground truth, treating the drain and source of a MOSFET as interchangeable and ignoring net names. We report exact isomorphism, a structural F1 over device and net signatures for partial credit, and W/L parameter accuracy. Because image and netlist derive from the same schematic source file, each image corresponds exactly to its reference netlist, and drawing style becomes a controllable nuisance variable.

Transcription is an established reading task rather than a design task; AMSnet (Tao et al., 2024) and Masala-CHAI (Bhandari et al., 2025) study the same extraction problem. Here it also provides a common evaluation across circuits regardless of testbench state, enabling the commit-date analysis below.

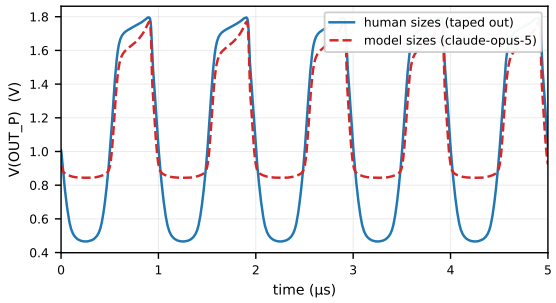
The baseline below uses the default schematic image. The release also ships a blinded image for each schematic. Device terminals, size annotations, and PDK model names remain because reading them is part of the task. Universal supply names also remain because they carry little circuit-specific




```
* Differential amplifier
* (5T OTA with output cap)

XM2 Vleft Vleft VCC VCC
+ sky130_fd_pr__pfet_01v8_lvt
+ L=2 W=5 nf=1
XM7 DIFFOUT Vleft VCC VCC
+ sky130_fd_pr__pfet_01v8_lvt
+ L=2 W=5 nf=1
XM6 Vleft PLUS TAIL VSS
+ sky130_fd_pr__nfet_01v8_lvt
+ L=0.5 W=2 nf=1
XM4 DIFFOUT MINUS TAIL VSS
+ sky130_fd_pr__nfet_01v8_lvt
+ L=0.5 W=2 nf=1
XM3 BIAS BIAS VSS VSS
+ sky130_fd_pr__nfet_01v8 L=2 W=4 nf=1
XM1 TAIL BIAS VSS VSS
+ sky130_fd_pr__nfet_01v8 L=2 W=4 nf=1
XC1 DIFFOUT VSS
+ sky130_fd_pr__cap_mim_m3_1
+ W=10 L=10 MF=10 C=2.076p
```

Scored: graph-isomorphic; W/L accuracy 1.00.



Device	Human	Model
input pair	3.6/0.38 ×2	20/0.5 ×10
loads (PMOS)	3.6/0.38 ×2	5/0.5 ×2
tail	3.6/0.38 ×2	7.2/0.38 ×4
bias mirror	3.6/0.38 ×2	3.6/0.38 ×2
bias resistor	L=1.85 ×10	L=30 ×1

Sizes as W/L in µm × nf. In the same testbench the human sizing swings ±665 mV at the output; the model's reaches ±463 mV with a compressed lower rail. The delta is what simulation-based scoring measures.

Figure 2: Both tasks, end to end. Top: transcription. The input schematic image (left) and claude-opus-5's verbatim output netlist (right) for a five-transistor OTA. Its net names differ from the drawing, but its graph is isomorphic to the reference. Bottom: sizing for a fully differential amplifier. The model receives the topology with sizes removed, then its proposal and the human reference run in the same author testbench. Aggregate results appear in Table 3.

retrieval value. Author-chosen net labels and instance names are removed as plausible retrieval handles; removal is driven by each circuit's netlist rather than by pattern matching. The blinded task stays well posed because the scorer ignores names.

Blinding lowers exact isomorphism for every model above floor, and the drop grows monotonically with default-image accuracy, from nothing for claude-haiku-4-5 at floor to 18.2 points for claude-fable-5; the per-model deltas are in Appendix C. Structural F1 falls by at most 0.11 and parameter accuracy holds or improves. Because the scorer ignores names, this pattern is consistent with labels serving as visual anchors for connectivity rather than as part of the answer itself: the models that trace wiring best lose the most when the anchors disappear. The drop is individually significant for the three leading models (paired McNemar $p \leq 0.04$) and a consistent trend below them. The stripped names themselves discriminate: on the 53 tasks with idiosyncratic hidden labels, no model emits a stripped name on more than six, the recurring hits are mixer idioms such as LO_N and IFOUT_P that any designer would coin, and the same models repeat visible labels on 21 to 29 tasks in the default condition (Appendix D). Blinded outputs almost never reproduce invisible names, which is the memorization signature this diagnostic exists to catch. It does not rule out prior exposure; the cutoff analysis addresses that separate channel.

Task 2: device sizing. Input: the human topology with W/L stripped to placeholders, the testbench, the PDK, and target metrics. Output: sizes; the topology is fixed. The reference is the simulated performance of the human-authored sizing under the same PDK, testbench, and simulator. Figure 2 (bottom) shows the task executed end to end on one circuit: the model's proposed sizes beside the reference sizes, and both simulated in the author's unmodified testbench.

The task runs on the fixed sizing set (§3); scoring is author-testbench-relative. Both sizings run the author's analyses. The reference selects the output vector with the largest transient swing or absolute low-frequency AC gain, and the model is evaluated on that same vector. For each eligible metric, let r be the model-to-reference ratio; the metric score is $\min(r, 1/r)$, averaged across available metrics. A nonconvergent run scores zero, while a convergent task without an eligible scalar metric receives a survival score of one; ten of the 17 tasks expose a scalar metric, and the remaining seven score survival only. A proposal that provides no sizes also scores zero: the unmodified deck holds the human sizes, so simulating it would grade the reference against itself.

Table 3 reports the mean scores, Figure 3b shows every cell behind them, and Appendix D decomposes each mean into the tasks with a scalar metric, the survival-only tasks, the answered-only mean, and the copy-robust variant, with bootstrap intervals. Every Claude or Gemini proposal that provides sizes converges, whereas both GPT generations produce several invalid device geometries. Convergence alone is too weak a bar. On one task, two proposals converge with output swing below one percent of the reference, a failure the convergence count cannot see and the simulated metric does. Refusal is a third failure mode, and it is confined to the newest models: claude-opus-5 returns an empty API-level refusal on 4 of the 17 prompts and claude-fable-5 on 11, while both transcribe the same circuits without objection. A full repeat run of claude-fable-5 refuses 12 prompts. Ten are refused in both runs and three change sides, so the refusal set is stable in bulk but not per prompt. Across three runs claude-opus-5's refusal count falls from 4 to 1 and its 17-task mean moves from 74.3 to 90.7. The leaderboard's largest single source of variance is the safety layer, not the sizing. On the tasks they answer, the two score 97.2 and 90.2. The sizing ranking therefore diverges from transcription's (Figure 3c) for two reasons: sizing knowledge is text-borne and present at every Claude tier, and on the newest models a safety policy, not capability, sets the score. Standardized per-class measurement templates are the next refinement (§6). Because the human sizes are public, Section 5 treats answer-key exposure explicitly.

Evaluation protocol. The released harness evaluates any model behind an API in a single invocation, and every request declares no tools, so evaluations are closed book by construction. Outputs produced outside the harness are scored from a standard submission format against the fixed task list, and our own published runs pass through that same path, so submitted and published results are graded by identical code. The scorer is deterministic: identical inputs produce identical result tables across processes and hash seeds, a property we corrected the parameter metric to guarantee.

Baseline. Table 3 reports aggregate results, and Figure 3 shows the per-task evidence behind them. Every run uses the provider's default decoding with a 16,000-token output cap. The three leading models ran three independent times on both tasks: exact isomorphism moves by at most 4.5 points, the fable-over-opus ordering holds in every run, and the largest sizing swing traces to refusal-count changes, not capability (Appendix D). On 66 tasks a 95% Wilson interval spans roughly ± 11 points and adjacent intervals overlap. Paired McNemar tests separate only the two steps the prose leans on, the gap above claude-sonnet-5 ($p = 0.001$) and the floor under gpt-5 ($p = 0.008$); orderings inside the frontier are point estimates, not separations (Appendix D).

Six of seven models drop sharply between the small and medium tiers (Figure 4, Appendix A); Haiku is already at floor on the small tier. The failure among the remaining models is wiring, not reading: a model that recovers a topology also reads its W/L annotations nearly perfectly, so a single blended score would hide which skill is missing. Tier difficulty is also not strictly monotonic: Opus and Gemini score higher on small circuits than on trivial ones. The trivial tier is dominated by passive networks and device models, whose vocabulary produces errors in models that transcribe a five-transistor amplifier cleanly.

Failure mode separates the leading models as much as exact accuracy does (Figure 3a). Claude-fable-5 posts the best exact-isomorphism and structural F1 point estimates, though McNemar against claude-opus-5 does not separate the two ($p = 0.45$). Behind them the frontier splits: claude-opus-5 is more often fully correct than gpt-5.6-sol, but its failures are severe, while gpt-5.6-sol produces repairable near misses and holds the second-best structural F1 despite a lower exact-isomorphism rate. The preferable model therefore depends on whether a downstream system can repair a netlist.

The scorer itself required two corrections, reported because each would have distorted a leaderboard. Parameter accuracy once scored against one arbitrary isomorphism of each symmetric circuit; it

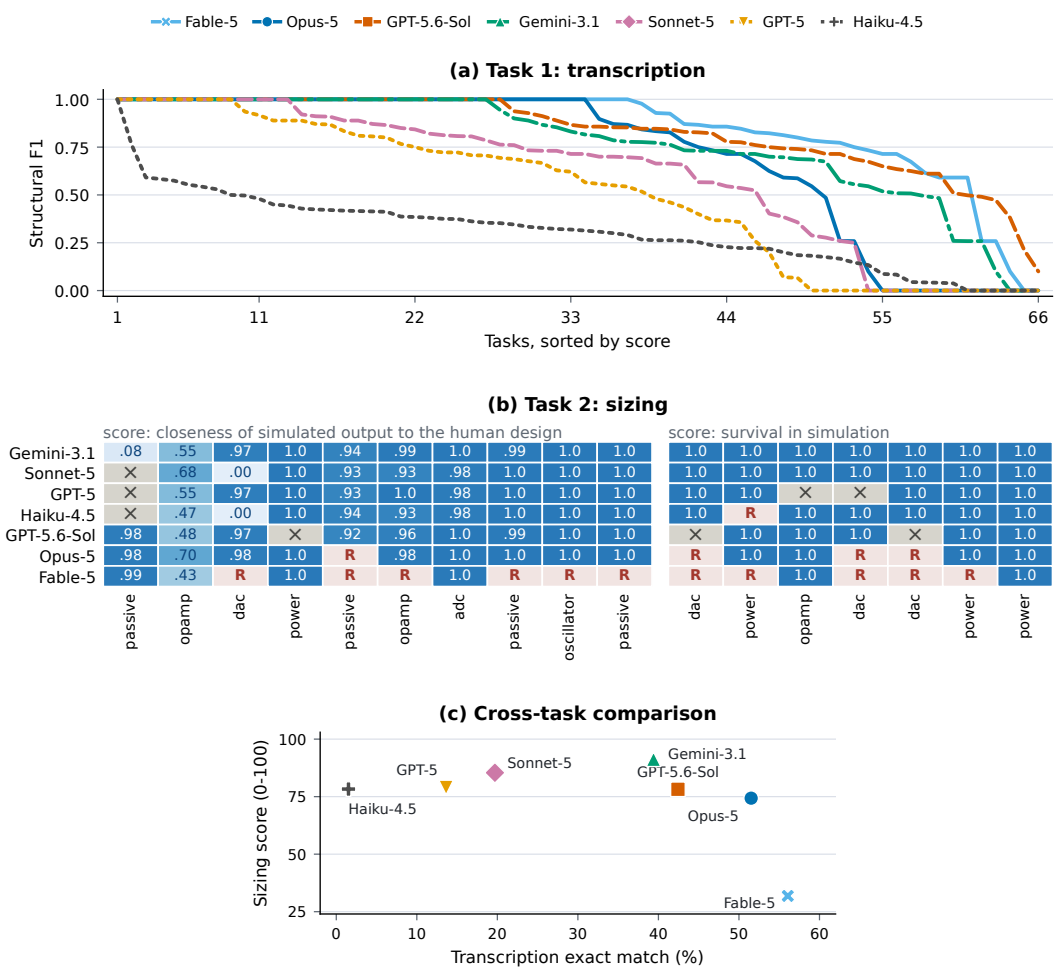


Figure 3: Main results for both tasks. (a) Task 1: each model's 66 structural F1 scores sorted in descending order. The width of the plateau at 1.0 equals the exact-isomorphism count, and the tail shape is the failure mode: Opus falls steeply once it fails, while Sol declines gradually. The encoding uses no banding thresholds. (b) Task 2: the complete score matrix, seven models by 17 sizing tasks, one cell per proposal. Each cell prints its score, with blue depth as a redundant encoding; a gray cross marks a nonconvergent sizing, and a red R marks a proposal that provided no sizes, which scores zero. The left group is scored by how close the proposal's simulated output comes to the human design's, hardest answered task first; in the right group the author analyses expose no usable scalar, so a convergent sizing scores one. The two .00 cells, on a DAC task, are convergent sizings whose output collapsed. (c) Transcription accuracy versus sizing score: model rankings differ across the two capabilities. The sizing axis is truncated to keep the seven points legible. The difficulty-tier breakdown of Task 1 is in Appendix A.

now pairs devices under the mapping with maximum parameter agreement among structurally valid isomorphisms, selected deterministically rather than by hash-seed order. PDK resistor terminals were once ordered pins, although the symbol draws no orientation and pin order cannot be read from the image. The corrected scorer still rejects genuinely misread connections, such as visually explicit bulk ties replaced by rail defaults. Table 3 also reports model-specific cutoff splits; Section 5 states what can and cannot be inferred from them.

Table 3: Model results on the fixed transcription set. Cutoffs are vendor-reported; post n counts tasks whose pinned revision postdates the cutoff, ruling out that exact revision from the model's pretraining window. Post/pre columns give exact isomorphism on each side. Sizing is the Task 2 score on a 100-point scale, with a proposal that provides no sizes scoring zero; No ans. counts those proposals (refusals, one unparseable answer) out of 17.

Model	Cutoff	Iso	Struct	W/L	Post n	Post / pre iso	Sizing	No ans.
claude-fable-5	2026-01	56.1%	0.853	0.921	13	30.8% / 62.3%	31.8	11
claude-opus-5	2026-05	51.5%	0.711	0.929	0	—	74.3	4
gpt-5.6-sol	2026-02	42.4%	0.831	0.942	13	30.8% / 45.3%	78.2	0
gemini-3.1-pro-preview	2025-01	39.4%	0.767	0.977	36	38.9% / 40.0%	91.2	0
claude-sonnet-5	2026-01	19.7%	0.604	1.000	13	7.7% / 22.6%	85.4	0
gpt-5	2024-09	13.6%	0.512	0.988	38	15.8% / 10.7%	79.0	0
claude-haiku-4-5	2025-02	1.5%	0.307	1.000	35	floor	78.3	1

5 MEASURING EXPOSURE

Public GitHub repositories are standard pretraining data, so this benchmark cannot assume independence. It instead provides circuit-level evidence for probing exposure: pinned commit dates, blinded images, release canaries, and a closed-book protocol.

Commit-date analysis. Each circuit retains its shuttle and pinned commit date; we compare these dates with each model's reported training cutoff. Table 3 reports the raw split per model; claude-opus-5 has no post-cutoff tasks in the fixed set. The raw split confounds exposure with two other variables, and Table 4 applies the matching controls. First, 23 of the 66 tasks belong to graph-duplicate groups (§3); a duplicate pair that straddles a model's boundary lets memorization of the earlier copy appear as post-cutoff skill, so tasks whose duplicate lies across the boundary are excluded per model (13 tasks at the 2026 cutoffs, 7 at January 2025, none at September 2024). Second, tier composition differs across the boundary: the post-cutoff side of the three 2026 cutoffs is dominated by seven medium-tier tasks from 2026 shuttles, and no evaluated model solves any of the seven, including claude-opus-5, whose May 2026 cutoff places all seven on its pre-cutoff side. The apparent pre-cutoff advantage of the 2026-cutoff models in Table 3 therefore tracks tier composition, not exposure. Within the trivial and small tiers, the two models with balanced splits score no worse post-cutoff than pre-cutoff (gemini-3.1-pro-preview 11/20 against 7/20; gpt-5 6/27 against 2/20), and the 2026-cutoff models retain three post-cutoff tasks each, too few to support a direction. Fisher exact tests place both splits well inside chance ($p = 0.34$ and $p = 0.44$): the reading is consistent with no memorization advantage, not evidence of its absence, and for the 2026-cutoff models the instrument returns no verdict until new shuttles land post-cutoff tasks. The same split applied to the 17 sizing tasks shows no post-cutoff deficit for any model; every model with post-cutoff sizing tasks scores at least as well on them (Appendix D). This remains a descriptive comparison, not an exposure estimate: an earlier version of a design may predate its pinned revision. The commit metadata supports the same analysis for future models, while new shuttles expand the post-cutoff side.

Table 4: Controlled cutoff splits: exact isomorphism as solved/n per side, after excluding tasks whose graph duplicate lies across the model's boundary. The medium-tier post-cutoff cells are dominated by seven 2026-shuttle tasks that no evaluated model solves. claude-haiku-4-5 is at floor overall and is omitted.

	Trivial + small		Medium	
Model	pre	post	pre	post
claude-fable-5	22/35	3/3	6/10	0/5
gpt-5.6-sol	16/35	2/3	4/10	0/5
claude-sonnet-5	10/35	0/3	1/10	0/5
gemini-3.1-pro-preview	7/20	11/20	4/8	0/11
gpt-5	2/20	6/27	1/8	0/11

Training-time and inference-time channels. Commit dates probe exposure through training data. A model evaluated with tool access has a second channel: the upstream repositories are public, so a search-enabled agent can retrieve an answer at inference time regardless of era. The protocol separates the two. Baseline scores are closed book (§4), so the model has no retrieval channel by construction. Systems with search or repository access belong to a separate open-book track and must be labeled as such. The blinded image removes author-chosen labels from the image. Reproducing an invisible internal name is evidence of memorization or external retrieval, so the blinded condition also provides an exposure diagnostic for submitted results.

Answer-key exposure differs by task. For transcription the artifact given to the model is an image that does not contain the netlist. For sizing the upstream repository contains the human sizes, and the score rewards matching the reference's simulated behavior, so a copied reference sizing would score well. Scoring runs the author's unmodified analyses (§4); the countermeasure is a regurgitation diagnostic, released per proposal, that compares proposed and reference sizes device by device. In the baseline, 30% of 782 proposed device sizes match the reference exactly, and 12 of 103 answered proposals reproduce every reference size. Six of the twelve are one op-amp task whose reference leaves most of its transistors at the default 1/0.15 size, so exact agreement there reflects the default rather than retrieval. The diagnostic flags answer-key reuse; it does not prevent it. Scoring flagged full reproductions as zero moves gemini-3.1-pro-preview from 91.2 to 73.6 and no other model by more than twelve points (Appendix D); on two-device tasks a full match is also reachable by independent choice, so the diagnostic bounds answer-key reliance rather than convicting it.

Canary. Every released netlist carries the BIG-bench canary as an ignored SPICE comment, and the release check verifies its placement. The canary marks generated artifacts; it cannot protect upstream sources that were already public. Commit-date splits address that channel, while blinded images address inference-time retrieval. Per-circuit memorization probes remain an archival milestone.

6 Limitations and roadmap

Most entries target Sky130 (SkyWater Technology Foundry & Google); broader coverage of GF180 and IHP SG13G2, including SiGe, is planned from MPW-mirror, IHP, and ISHI-KAI sources. Schematics with unresolved host paths are excluded from image-sensitive splits. Circuit-class labels are keyword-derived and will be reclassified from netlist graphs. Top-level designs without a parseable netlist remain outside the redundancy analysis.

Three milestones remain for the archival version. First, because author testbenches expose heterogeneous measurements, we will add per-class templates over the eligible set, starting with gain, gain-bandwidth product, and phase margin for the amplifier class. This layer widens sizing coverage beyond the leaf-subcircuit set. Second, we will add the per-circuit memorization probes described in Section 5, joining the delivered blinded-image comparison. Third, topology generation as a task. We do not claim it in this version because two of its components are unsolved: a uniform representation for the design brief, and automatic pin-matching between a generated circuit and the author testbench. Both build on the measurement layer; claiming the task before its scoring exists would overstate readiness.

AI use statement

We used an LLM assistant to help build the harvest and evaluation pipeline code and to draft and edit the manuscript text. The authors reviewed all AI-assisted code and text, verified every reported number against the committed result tables, and the release ships the consistency checker that enforces the match. The models evaluated in Section 4 are objects of study, not authoring tools. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

All redistributed designs are Apache-2.0, verified from license texts archived at the pinned commits; the released attribution file credits every repository and its authors, and per-circuit manifests carry source URL, commit, license, and extraction method. The Tiny Tapeout index (Venn et al.) is CC0; the PDKs are Apache-2.0. Benchmark code and released data use Apache-2.0, matching the source licenses. Designer names appear only as attribution of their public, licensed work.

Reproducibility statement

Every stage, from harvest through simulation and scoring, runs inside one digest-pinned container image, and classification tables regenerate byte-identically across complete runs (§3). Each design is fetched at the commit its shuttle submission records, so the dataset rebuilds from public sources. The anonymized repository at `https://anonymous.4open.science/r/sganalog-anon-7EF6/` holds the pipeline, the evaluation harness, the fixed task sets, and the consistency checker; Appendices A–E give the tier breakdown, the full inventory, and the simulation browser.

References

J. Bhandari, V. Bhat, Y. He, H. Rahmani, S. Garg, and R. Karri. Masala-CHAI: A large-scale SPICE netlist dataset for analog circuits by harnessing AI, 2025. arXiv:2411.14299.

G. Chiari, M. Piccoli, and D. Zoni. OSIRIS: Bridging analog circuit design and machine learning with scalable dataset generation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.19439.

Z. Dong, W. Cao, M. Zhang, D. Tao, Y. Chen, and X. Zhang. CktGNN: Circuit graph neural network for electronic design automation. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2308.16406.

J. Gao, W. Cao, J. Yang, and X. Zhang. AnalogGenie: A generative engine for automatic discovery of analog circuit topologies. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2503.00205.

R. M. Ghorab, E. Parisi, C. Gutierrez, M. Alberti-Binimelis, M. Moreto, D. Garcia-Gasulla, and G. Kestor. NotSoTiny: A large, living benchmark for RTL code generation, 2025. arXiv:2512.20823.

P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits*. Wiley, fifth edition, 2009.

Y. Lai, S. Lee, G. Chen, S. Poddar, M. Hu, D. Z. Pan, and P. Luo. AnalogCoder: Analog circuit design via training-free code generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 379–387, 2025.

Y. Lai, S. Poddar, S. Lee, G. Chen, M. Hu, B. Yu, P. Luo, and D. Z. Pan. AnalogCoder-Pro: Unifying analog circuit generation and optimization via multi-modal LLMs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2026. doi:10.1109/TCAD.2026.3673493. arXiv:2508.02518.

J. Li, H. Zhi, R. Lyu, Z. Chen, W. Li, Z. Bi, K. Zhu, Y. Zeng, W. Shan, C. Yan, F. Yang, Y. Li, and X. Zeng. AnalogGym: An open and practical testing suite for analog circuit synthesis. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2024. arXiv:2409.08534.

A. Mehradfar, X. Zhao, Y. Niu, S. Babakniya, M. Alesheikh, H. Aghasi, and S. Avestimehr. AICircuit: A multi-level dataset and benchmark for AI-driven analog integrated circuit design. In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2024. arXiv:2407.18272.

H. Pretl and G. Zachl. IIC-OSIC-TOOLS, version 2026.02. Institute for Integrated Circuits, Johannes Kepler University Linz. Zenodo. https://doi.org/10.5281/zenodo.14387234, 2026.

B. Razavi. *Design of Analog CMOS Integrated Circuits*. McGraw-Hill, second edition, 2017.

S. Schippers. Xschem: Schematic capture and netlisting EDA tool. https://xschem.sourceforge.io.

SkyWater Technology Foundry and Google. SkyWater open source PDK (SKY130). https://github.com/google/skywater-pdk.

Z. Tao, Y. Shi, Y. Huo, R. Ye, Z. Li, L. Huang, C. Wu, N. Bai, Z. Yu, T.-J. Lin, and L. He. AMSNet: Netlist dataset for AMS circuits. In *IEEE LLM Aided Design Workshop (LAD)*, pp. 1–5, 2024. See also Y. Shi et al., AMSnet 2.0, IEEE International Conference on LLM-Aided Design, 2025.

M. Venn et al. Tiny tapeout. https://tinytapeout.com.

H. Vogt et al. Ngspice: Open source mixed-signal circuit simulator. https://ngspice.sourceforge.io.

B. Xu, K. Zhu, M. Liu, Y. Lin, S. Li, X. Tang, N. Sun, and D. Z. Pan. MAGICAL: Toward fully automated analog IC layout leveraging human and machine intelligence. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019. Circuit release: https://github.com/magical-eda/MAGICAL-CIRCUITS.

A Difficulty-tier breakdown of Task 1

Figure 4 plots exact isomorphism per difficulty tier for the seven evaluated models, the per-tier view behind the tier claims in Section 4.

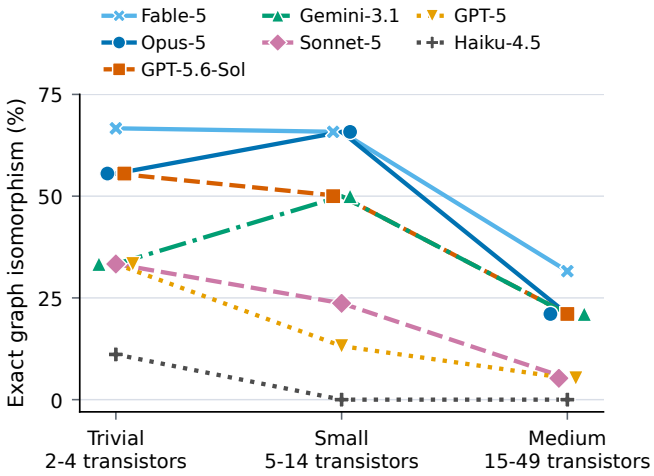


Figure 4: Exact isomorphism by difficulty tier ($n = 9, 38, 19$): six models drop from small to medium, while Haiku is already at floor; Opus and Gemini improve from trivial to small. Markers with identical scores are offset slightly within each tier for visibility; only their vertical position encodes accuracy.

B Full dataset inventory

Table 5 expands the spine of Table 2 into the complete harvest, licensing, composition, and redundancy record.

Table 5: The full dataset inventory. Left: the harvest funnel with triage and licensing detail. Right: composition by role, transistor-count tier, era, and device–net graph hash.

Harvest funnel	n
Index projects (27 shuttles)	4,572
Entries with analog pins	237
Unique repositories	195
ship xschem (Schippers) schematics	133
ship SPICE netlists only	12
layout data only (unusable)	48
fail to clone	2
Usable repositories	145
license verified Apache-2.0	144
hold fetchable schematics	129
contribute classified circuits	125
Schematics fetched at pinned commits	1,448

Composition	n
Circuits	933
with a PDK transistor	809
top-level designs	425
sub-blocks	508
Testbenches	515
circuits driven by one	350
Top-level tiers: trivial (0–4 transistors)	164
small (5–14)	161
medium (15–49)	78
large (50–149)	20
very large (150+)	2
non-trivial (5 or more)	261
Eras: 2023–24 / 2025 / 2026	194 / 143 / 88
Graph-hashable top-level designs	357
unique device–net graph hashes	273
redundant designs	84
duplicate groups / cross-repository	43 / 32
Fixed tasks: transcription / sizing	66 / 17

C BLINDED-IMAGE COMPARISON

Table 6 reports both image conditions on the 66-task set for all seven evaluated models.

Table 6: Default against blinded images on the fixed 66-task set. Iso = exact graph isomorphism, Struct = structural F1, W/L = parameter accuracy. The exact-isomorphism drop grows with default-image accuracy while parameter accuracy holds or improves.

	Iso		Struct		W/L	
Model	default	blinded	default	blinded	default	blinded
claude-fable-5	56.1%	37.9%	0.853	0.821	0.921	0.920
claude-opus-5	51.5%	36.4%	0.711	0.723	0.929	0.984
gpt-5.6-sol	42.4%	30.3%	0.831	0.775	0.942	0.973
gemini-3.1-pro-preview	39.4%	30.3%	0.767	0.744	0.977	0.989
claude-sonnet-5	19.7%	12.1%	0.604	0.494	1.000	1.000
gpt-5	13.6%	10.6%	0.512	0.455	0.988	1.000
claude-haiku-4-5	1.5%	1.5%	0.307	0.253	1.000	1.000

D UNCERTAINTY, SENSITIVITY, AND DIAGNOSTICS

Everything here is computed from the released per-task tables by one script, with a fixed bootstrap seed.

Interval estimates and paired tests. Table 7 carries 95% Wilson intervals for exact isomorphism and task-bootstrap intervals for the sizing mean. Adjacent transcription intervals overlap throughout the frontier. Paired McNemar tests on the 66 shared tasks separate gemini-3.1-pro-preview from claude-sonnet-5 ($p = 0.001$) and gpt-5 from claude-haiku-4-5 ($p = 0.008$); no adjacent pair above sonnet separates (fable against opus $p = 0.45$, opus against sol $p = 0.18$, sol against gemini $p = 0.75$). The default-against-blinded drop is significant for claude-fable-5 ($p = 0.002$), claude-opus-5 ($p = 0.013$), and gpt-5.6-sol ($p = 0.039$), and a consistent trend for the remaining models ($p = 0.11$, 0.23, 0.69, 1.0 down the accuracy order).

Table 7: Point estimates with 95% intervals: Wilson for exact isomorphism over 66 tasks, bootstrap over tasks (10,000 resamples) for the 17-task sizing mean.

Model	Exact iso [CI]	Sizing [CI]
claude-fable-5	56.1 [44.1, 67.4]	31.8 [11.7, 54.5]
claude-opus-5	51.5 [39.7, 63.2]	74.3 [52.7, 92.2]
gpt-5.6-sol	42.4 [31.2, 54.4]	78.2 [58.3, 94.8]
gemini-3.1-pro-preview	39.4 [28.5, 51.5]	91.2 [77.9, 99.6]
claude-sonnet-5	19.7 [11.9, 30.8]	85.4 [68.3, 98.6]
gpt-5	13.6 [7.3, 23.9]	79.0 [58.7, 94.3]
claude-haiku-4-5	1.5 [0.3, 8.1]	78.3 [58.1, 94.1]

Repeat-run variance. The three leading models ran three independent times per task (Table 8). Transcription is tight: the largest range is 4.5 points, and claude-fable-5 scores above claude-opus-5 in every run. Sizing variance concentrates where the safety layer acts: claude-fable-5's three runs refuse 11, 12, and 11 prompts and stay within 3.8 points, while claude-opus-5's refusals fall from 4 to 4 to 1 and its mean rises 16.4 points; gpt-5.6-sol never refuses, and its 8.7-point range comes from convergence draws (14, 14, and 12 of 17 sized decks).

Table 8: Three independent runs of the leading models. Transcription is exact isomorphism on the 66-task set; sizing is the 17-task mean with refusal counts in parentheses.

	Transcription		Sizing	
Model	runs	mean (range)	runs (refusals)	mean (range)
claude-fable-5	56.1 / 53.0 / 57.6	55.6 (4.5)	31.8 (11) / 28.1 (12) / 31.8 (11)	30.6 (3.8)
claude-opus-5	51.5 / 48.5 / 48.5	49.5 (3.0)	74.3 (4) / 74.3 (4) / 90.7 (1)	79.8 (16.4)
gpt-5.6-sol	42.4 / 43.9 / 42.4	42.9 (1.5)	78.2 (0) / 78.2 (0) / 69.5 (0)	75.3 (8.7)

Internal duplicates in the transcription set. The fixed 66 tasks span 53 distinct device–net graphs; ten duplicate groups, inherited from collection redundancy, hold the other 13 members, the largest holding four copies of one graph. Rescoring every model on the 53 group representatives moves no exact-isomorphism score by more than 3.6 points (claude-fable-5 56.1 to 52.8, claude-opus-5 51.5 to 49.1, gpt-5.6-sol 42.4 to 39.6, gemini 39.4 to 35.8, claude-sonnet-5 19.7 to 20.8, gpt-5 13.6 to 15.1, claude-haiku-4-5 1.5 to 1.9) and changes no ordering.

Sizing decomposition. Table 9 splits the 17-task mean into its parts. The metric-bearing mean reorders the frontier: claude-opus-5 (86.4) edges gemini (85.1) once survival credit is out, and claude-sonnet-5 drops from 85.4 to 75.2. Copy-robust scoring, which zeroes proposals the regurgitation diagnostic flags as full reference reproductions, moves gemini from 91.2 to 73.6 and every other model by less than twelve points.

Table 9: The 17-task sizing mean decomposed: metric-bearing tasks ($n = 10$), survival-only convergence ($n = 7$), answered-only mean, full-copy proposals flagged by the regurgitation diagnostic, and the mean with flagged proposals scored zero.

Model	All 17	Metric 10	Surv. 7	Answered	Copies	Zeroed
claude-fable-5	31.8	34.1	2/7	90.2	2/6	20.2
claude-opus-5	74.3	86.4	4/7	97.2	1/13	68.6
gpt-5.6-sol	78.2	82.9	5/7	78.2	1/17	72.6
gemini-3.1-pro-preview	91.2	85.1	7/7	91.2	3/17	73.6
claude-sonnet-5	85.4	75.2	7/7	85.4	2/17	74.0
gpt-5	79.0	84.3	5/7	79.0	2/17	67.2
claude-haiku-4-5	78.3	73.1	6/7	83.2	1/16	72.8

Stripped-name emission in the blinded condition. Fifty-three of the 66 tasks carry at least one idiosyncratic author label removed by blinding. Counting tasks where a model's output contains any such hidden name: claude-fable-5 emits one on 5 tasks, claude-opus-5 and gpt-5.6-sol on 6, gemini on 4, claude-sonnet-5 and gpt-5 on 5, claude-haiku-4-5 on 1. The recurring names are mixer-port idioms (LO_N, LO_P, IFOUT_N, IFOUT_P) that the topology itself suggests. In the default condition, where the labels are visible, the same models repeat them on 21 to 29 of the 53 tasks, so the counter is sensitive; near-zero emission under blinding argues against retrieval of the labeled artifact.

Commit-date split of the sizing tasks. Applying each model's cutoff to the 17 sizing tasks: claude-fable-5 scores 33.3 post against 31.1 pre ($n = 6/11$), gpt-5.6-sol 82.6 against 75.8 ($6/11$), gemini 99.9 against 81.5 ($9/8$), claude-sonnet-5 98.5 against 78.2 ($6/11$), gpt-5 89.5 against 63.9 ($10/7$), claude-haiku-4-5 86.3 against 71.1 ($8/9$); claude-opus-5 has no post-cutoff sizing task. No model scores worse on provably-unseen sizing tasks.

E SIMULATION BROWSER

The release includes a local browser interface over the verified circuit–testbench pairs (Figure 5). For each pair it shows the circuit and testbench schematics, both extracted ngspice netlists, the upstream repository at its pinned commit, and any specifications the project's own documentation publishes. A run started from the page executes through the same pinned-container entry point as the pipeline, so its result is bit-identical to the batch simulatability check, and the waveforms are parsed from the run's raw files. The sizing task set of Section 3 is drawn from these verified pairs. A static companion page with precomputed runs of ten representative pairs, one per circuit class, is at `https://admirable-sunburst-858a7d.netlify.app/`; the live browser ships with the anonymous repository and requires only Docker.

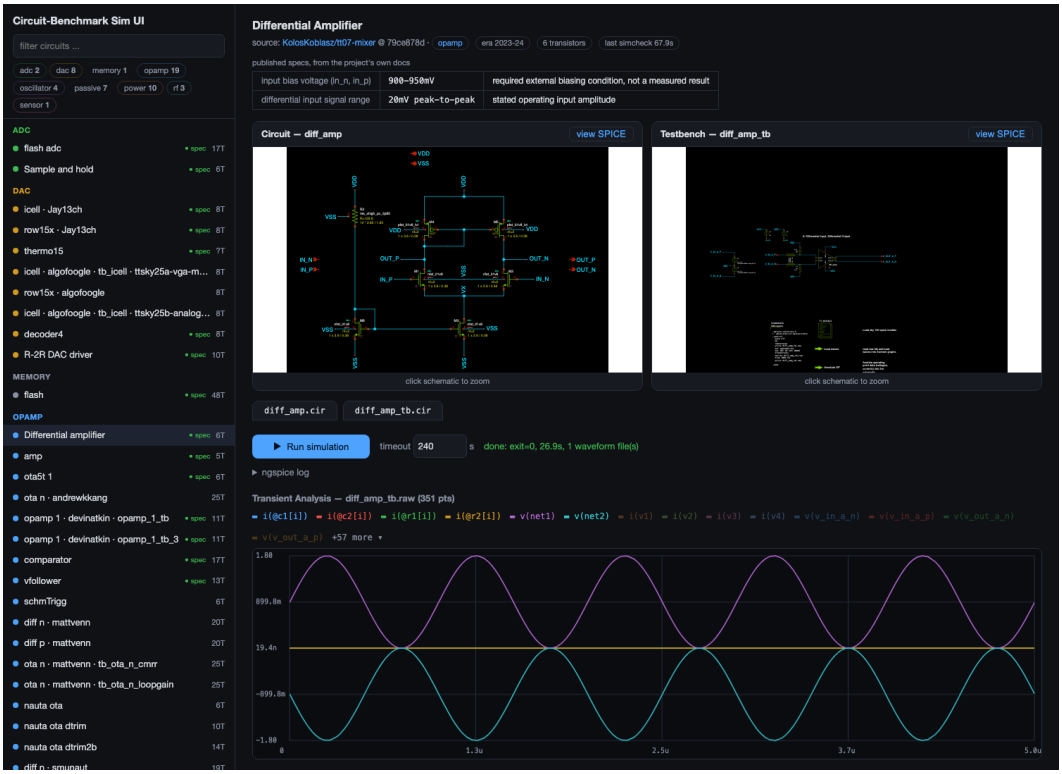


Figure 5: The simulation browser on one verified pair, a differential amplifier from the collection: published specifications, circuit and testbench schematics, and the transient waveforms produced by a run inside the pinned container (exit 0 in 26.9 seconds; the two complementary output sines around the input common mode).