QUARC: A BENCHMARK FOR LLM-GENERATED QUANTUM CIRCUITS AND AN AUDIT OF ITS ACCEPTANCE REGION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scoring an LLM's code against a reference can reject correct solutions. Tolerating equivalences can remove such false rejections, yet each tolerated transformation can widen what the verifier accepts. QuARC brings this tradeoff to quantum circuits, scoring 86 tasks by membership in an explicit acceptance region rather than distance to a reference. The region admits any circuit meeting the task's predicate up to global phase, gate decomposition, qubit relabeling, and ancillas that start and end in the all-zero state. Relabeling, however, applies only where a task fixes no layout. An audit on probes and baseline generations measures what each design choice prevents and admits. A check at the all-zero input alone, for instance, accepts 42 generations wrong under their prompts, none of which QuARC's verifier accepts. Relabeling decides 76 acceptances, 8 of them wrong under the prompt's labels, and no false rejection that it prevents is established. Restricting relabeling to prompts naming no qubit roles or order would remove these wrong acceptances. Separately, a disguised copy of the canonical output passes 16 of the 70 tasks without free parameters, all among the 17 judged at a single input. Across the evaluated models, pass rates span 55.8% to 91.5%. We release the corpus, verifier, audit scripts, and baseline generations.

1 INTRODUCTION

Large language models (LLMs) increasingly write code, and benchmarks judge it by running each candidate against a test suite (Chen et al., 2021; Austin et al., 2021). However, an evaluator can misjudge a solution in two ways. Tests can accept a solution that fails the intended task (Liu et al., 2023), and a comparison with a reference form can reject one that meets it (Zhong et al., 2020).

Quantum circuits raise the risk of false rejection, because many different circuits meet the same task (Guo et al., 2025; Slim et al., 2026). A Bell state arises from a Hadamard gate on either qubit followed by a controlled NOT (CNOT) to the other. Both circuits prepare the same state, so a state comparison accepts both, but a unitary comparison against one rejects the other (Fig. 2a). Benchmarks for LLM generated quantum code judge against a reference: a canonical solution with unit tests (Vishwakarma et al., 2024), a reference output distribution (Guo et al., 2025; Slim et al., 2026), or a ground truth result (Yang et al., 2025).

Admitting these equivalent forms can remove false rejections, but every acceptance it adds is a potential false acceptance. QuARC (Quantum Acceptance Region Corpus) holds 86 tasks in six categories and names, for each, the admitted set as an acceptance region $A(T)$, scored by membership. A circuit lies inside $A(T)$ when it meets the task's predicate up to four allowances. The allowances are global phase, exact rewriting into other gates, qubit relabeling where the task fixes no layout, and ancillas returned to their all-zero start. This region is closed under exact transpilation that neither routes nor removes SWAP gates (§3). The runner, the verifier we ship, tests membership and departs from it at points named in §3.3 and §7, such as a relabeling search capped at five qubits. Fig. 1 summarizes the region and the two ways we probe it.

Every task accepts its own canonical solution, 86 of 86, yet a region of transformations cannot be checked member by member. How much does the region admit? A first, partial answer is a per task bound on how often a Haar random unitary passes a model of the test, taken without relabeling. It is

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.

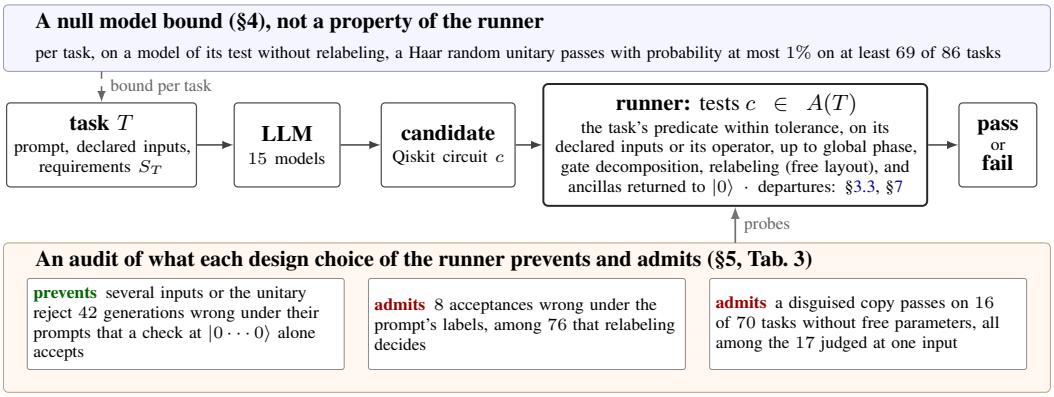


Figure 1: QuARC at a glance. The runner tests membership in the acceptance region $A(T)$ and departs from it at points named in §3.3 and §7. The upper band bounds how often a Haar random unitary lands in a modeled acceptance event, not the runner's own test (§4). The lower band audits what each design choice prevents and admits (§5, Tab. 3).

at most 1% for at least 69 of 86 tasks (§4). It concerns that model, not the rate at which the shipped runner accepts wrong submissions, which the audit tests.

The audit asks, per design choice, which errors it prevents and admits, on structured probes and on the $3{,}475$ generations gradable against an exact ground truth (§5, Tab. 3). A test at the all-zero input alone would admit 42 generations wrong under their prompts, which the runner rejects by checking several inputs or the unitary. Relabeling, the tolerance the Bell example illustrates, decides 76 of $3{,}158$ acceptances, and 8 are wrong under the prompt's labels. Of these 76, 52 lie on the two quantum Fourier transform (QFT) tasks and are the textbook transform in the opposite bit order, which only Qiskit's convention excludes. On this baseline, relabeling mostly settles a bit order convention and averts no rejection known to be false. Restricting relabeling to prompts naming no qubit roles or order would remove the 8 wrong and 2 ambiguous cases, keeping 66. A disguised copy, a circuit that prepares the canonical output of one input directly, passes 16 of the 70 tasks without free parameters. All 16 fall among the 17 tasks judged at a single input, so a check at one input, not a tolerance, lets the copy through. On 10 of them, a ban on prebuilt state preparation hangs on a gate name that offline transpilation evades, though none of the 420 extractable generations calls a banned name.

The baseline spans 15 models and $3{,}870$ generations, with pass rates from 55.8% to 91.5% (§6). A paired test separates 49 of 105 model pairs, so the ranking reads in bands. For builders, the audit suggests checking several inputs or the unitary, enforcing bans structurally, restricting relabeling to such prompts, and reading the permutation a transpiler records (§8).

This work makes four contributions.

1. QuARC, a corpus of quantum circuit tasks scored by membership in an acceptance region (§3, §4).
2. A null model bound per task, with proofs and calibration (§4, App. C, App. D).
3. An audit protocol that measures what each design choice prevents and admits (§5, App. E, App. G).
4. A baseline of model generations read in bands of a paired test (§6).

The release adds the scripts behind the bounds, calibrations, statistics, and the audit. All measurements on model generations come from one prompt template and one pass of three samples per task (§7).

2 Related Work

QuARC relates to three lines of prior work: code benchmarks graded by tests, benchmarks for LLM generated quantum code (Tab. 1), and quantum equivalence checking.

Table 1: What each benchmark judges against, the equivalences it declares tolerated, and how the threshold of its correctness comparison is set. A dash: none declared (individual tests may tolerate some). *Auxiliary metric only. ‡Free layout only. QHE: Qiskit HumanEval. KL: Kullback–Leibler.

Benchmark	Judged against	Tolerated equivalences	Threshold setting
QHE	handwritten unit tests	–	fixed per test
QuanBench	tests, KL to reference	global phase*, decomposition*	fixed, 0.05
QuanBench+	checks or KL	same output statistics	calibrated on canonical runs
QCircuitBench	end to end tests, fidelity	–	graded score
QCoder	state vector, constraints	global phase, set per task	–
QuARC	membership in $A(T)$	global phase, decomposition, relabeling‡, ancilla	fixed per task; Haar null bound (§4)

False acceptance and false rejection. A code benchmark that grades by tests (Chen et al., 2021; Austin et al., 2021) can accept a wrong solution, and a metric that matches code to a reference can reject a correct one. For false acceptance, tests "can be limited in both quantity and quality": extended EvalPlus tests lower pass rates by up to 19.3–28.9% (Liu et al., 2023). For false rejection, Zhong et al. (2020) report that the Spider metric leads to a "false negative" rate of about 2.5% on average and 8.1% at worst. They propose test suite accuracy to approximate the semantic accuracy of database queries generated from text. QuARC examines both directions: a null model bound (§4) and an audit (§5) probe acceptance, and a qubit relabeling measurement probes rejection (§6).

Benchmarks for LLM quantum code. Tab. 1 compares QHE (Vishwakarma et al., 2024), QuanBench (Guo et al., 2025), QuanBench+ (Slim et al., 2026), QCircuitBench (Yang et al., 2025), and QCoder (Mikuriya et al., 2025). QHE's handwritten tests include structural conditions such as a fixed circuit depth, which rejects an equivalent circuit of another depth. QuanBench and QuanBench+ document how a fidelity comparison can penalize correct circuits. QuanBench reports that differences such as qubit ordering and relative phase lower the process fidelity of some correct samples. Fidelity is only a complementary metric there, and tests and a KL threshold decide acceptance. QuanBench+ calibrates a single KL threshold on 1,000 canonical runs per probabilistic task, so that, pooled over tasks, the canonical solution rarely fails under finite shot variability. Whereas QCoder sets a global phase rule per task, QuARC declares one set of tolerated transformations for all tasks (relabeling only under a free layout). A per task bound limits the probability that a uniformly random (Haar) unitary falls into its modeled acceptance event (§4).

Equivalence checking. Quantum equivalence checking addresses whether two circuits implement the same functionality (Burgholzer & Wille, 2020; 2021). Random stimuli verification checks correctness by simulation, and high error detection rates can be achieved with only a few suitably generated random stimuli (Burgholzer et al., 2021). QuARC randomizes the candidate, not the input, and its bound says how much of a null model a modeled acceptance event admits, not how likely one verdict is wrong. A task checking an observable such as the argmax (most likely output) accepts more circuits than those equivalent to the canonical solution under the four transformations (§3). The audit (§5) tests whether this is too much.

3 Acceptance Regions

A task's acceptance region is the set of circuits it can count as correct. We define it, show when relabeling is an equivalence and where exact transpilation stays inside it, and describe how the runner checks candidates and where it departs from $A(T)$.

3.1 Formalization

A task T specifies a canonical solution c_T on n_T qubits, a verification type, and, where the type has one, a threshold θ_T. A candidate c may use $a \geq 0$ extra qubits, and U_c is its operator on all $n_T + a$ qubits, U_{c_T} that of c_T on its n_T qubits. Candidates relate to c_T through four declared equivalence transformations:

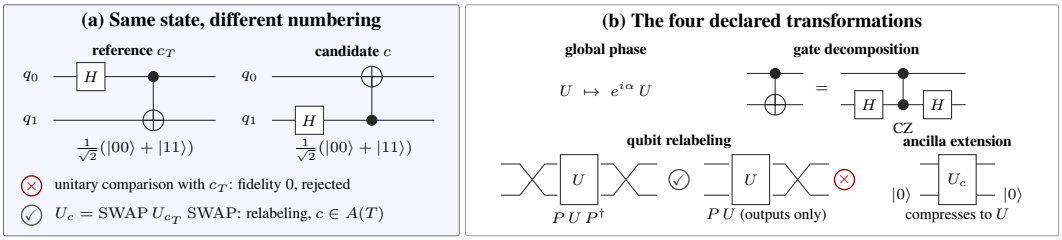


Figure 2: (a) Both circuits prepare the same Bell state from $|00\rangle$. A unitary comparison against c_T rejects the candidate, although $U_c = \text{SWAP}\, U_{c_T}\, \text{SWAP}$ puts it in $A(T)$ on a free layout. (b) The four declared transformations. SWAP routing permutes only the outputs, which is no relabeling.

- *Global phase:* $U \mapsto e^{i\alpha}U$ for any real α, which no measurement distinguishes from U.
- *Gate decomposition:* any rewriting that implements the same operator.
- *Qubit relabeling:* $U \mapsto PUP^\dagger$ for a qubit permutation P with adjoint $P^\dagger$, applied at both ends.
- *Ancilla extension:* a wider circuit counts as U if, with extra qubits in $|0\rangle$, it applies U and restores them.

The acceptance region of T is

$$A(T) = \{c : \exists P \in \mathcal{P}_T :\ V_T\big(\Pi_0(P^\dagger U_c P)\big) = 1\}. \tag{1}$$

Here $\mathcal{P}_T$ contains every permutation of the $n_T + a$ qubits, or only the identity if T fixes its layout. The type's predicate V_T reads $W = \Pi_0(P^\dagger U_c P)$, the compressed, relabeled candidate operator. Eq. (1) undoes the relabeling on the candidate's side ($P^\dagger U_c P$), and Π_0 compresses the a extra qubits onto $|0\rangle$ at both ends. Since Π_0 does not renormalize, amplitude left outside $|0\rangle$ on extra qubits is lost from the fidelity. For the fidelity types, $V_T = [F \geq \theta_T]$ with F the state or unitary fidelity of Tab. 2. The other types check output probabilities, stabilizer expectations, or, for the two custom verifiers, teleported states and Simon's output pattern. Accordingly, θ_T is a fidelity or a probability threshold, and for stabilizers sets a tolerance. Structural requirements S_T (for example prescribed or forbidden gates, or an exact qubit count) are checked separately. A candidate is accepted only if it lies in $A(T)$ and meets S_T, so S_T can reject a member of $A(T)$, such as a decomposition into forbidden gates.

Reference comparisons, such as the unitary comparison in Fig. 2a, are Eq. (1) with $\mathcal{P}_T = \{I\}$ and $a = 0$. Admitting extra qubits, a larger $\mathcal{P}_T$, or a predicate that accepts more can only add candidates. Every candidate in $A(T)$ that a reference comparison rejects is admitted by the region, which removes a false rejection when the candidate is correct. Every acceptance $A(T)$ adds is a potential false acceptance, and §5 probes this direction. Thus $A(T)$ is not the set of circuits equivalent to c_T: the threshold admits operators within a radius (§3.3), and types that read only output probabilities admit more (§5).

3.2 Relabeling and closure under exact transpilation

Relabeling is an equivalence only where a task gives its qubits no roles. A fixed layout ($\mathcal{P}_T = \{I\}$) is declared by 15 tasks. Some other tasks name qubit roles in their prompt without declaring one (the arithmetic tasks), and for those the region admits relabelings the prompt does not.

The region $A(T)$, not its conjunction with S_T, is closed under every composition of the transformations T admits. Exact transpilation that neither routes nor removes SWAP gates is one such composition: gate translation is a decomposition, and placement onto a larger register is a relabeling with ancillas. SWAP routing permutes only the outputs and is no relabeling (Fig. 2b). Routed circuits lie outside this guarantee. Approximate synthesis stays in $A(T)$ only as far as the type's threshold or predicate admits it. The runner carries closure for relabelings only where it knows or can search a layout, and seven of the nine tasks with more than five qubits leave their layout free.

Table 2: Nine verification types by family (overlap, eigenvalue, output predicate). W: compressed, relabeled candidate operator; $d = 2^{n_T}$; type per task: App. A. A few functional tasks check listed output probabilities or an ancilla marginal instead. TV: total variation distance.

Verification type	Tasks	Compared quantity	Threshold
state fidelity	10	$\lvert\langle 0\cdots 0 \rvert U_{c_T}^\dagger W \lvert 0\cdots 0\rangle\rvert^2$	0.99 (0.999 for 3)
parametric state fidelity	12	state overlap at each parameter point	0.99
unitary fidelity	16	$\lvert\mathrm{tr}(U_{c_T}^\dagger W)\rvert^2/d^2$	0.99 (0.999 for 2)
parametric unitary fidelity	4	unitary overlap at each parameter point	0.99
code space unitary fidelity	1	unitary overlap on a code subspace	0.99
stabilizer eigenvalue	1	expectation of each stabilizer	within $1-\theta_T$, $\theta_T = 0.99$
functional	35	argmax and TV ≤ 0.05 per declared input	none
probability	4	mass per marked outcome and in total	floors and total, per task
custom	3	teleported state; Simon's output pattern	0.95; Simon: 0.1 / 0.01

3.3 From definition to runner

The runner checks S_T, then the observable of one of the nine types. A single unwanted rotation $R_z(\varphi)$ on a unitary task, or on a qubit the target state holds in equal superposition, gives $F = \cos^2(\varphi/2)$. The acceptance radius is $\varphi_T = 2\arccos\sqrt{\theta_T}$, 0.2003 rad at $\theta_T = 0.99$ and 0.0633 rad at 0.999. T3.11 (controlled phase $\pi/3$) uses 0.999, since at 0.99 angle $\pi/3 + 0.1$ would pass ($F = 0.998$).

Most functional tasks read, at each declared input (2 to 32 per task), only the most likely output and whether the distribution lies within total variation distance 0.05 of the canonical one. On the state fidelity, probability, and stabilizer types and on the two Simon tasks, the verdict rests on the single input $|0\cdots 0\rangle$ (§5).

Given a layout from the candidate, by explicit mapping or transpiler metadata, the runner checks once under that relabeling. Without a layout it retries a failed candidate on every qubit relabeled image, as long as the circuit it compares has at most five qubits, and beyond that the failure stands. Tasks with a fixed layout are checked only in the given numbering. The runner departs from $A(T)$ both ways: on the parametric types it searches a relabeling per parameter point, admitting more than one shared relabeling would. On the state types it does not check that extra qubits return to $|0\rangle$. Conversely, on the unitary types it accepts a candidate wider than the task only with a layout and up to a width limit, and an exact qubit count in S_T excludes ancillas (§7).

4 The Benchmark and Its Null Model

With the regions defined, we turn to the tasks and the prompt behind them (§4.1) and to how far chance acceptance by a Haar random unitary can be bounded, and for which event (§4.2).

4.1 Corpus and prompt

The six task categories hold 9, 19, 10, 16, 14, and 18 of the 86 tasks, and an identifier T$x.y$ marks a task of category x. Difficulty is a separate field and does not follow the category. Each task, listed in App. A, carries a description, a canonical solution, a verification type (Tab. 2) and, where a task splits into functionally distinct sub components, a component count (App. F). All 15 models see one minimal prompt. A fixed system instruction asks for Qiskit code that builds a variable `qc`. The task description follows, and the template adds no hardware context, chain of thought scaffolding, few shot examples, or tool descriptions (App. H). Of the 86 descriptions, 10 forbid a prebuilt state preparation, and §5 tests whether this ban holds.

4.2 A bound on chance acceptance

An acceptance region raises a question a reference comparison does not: how often does an unrelated operator land in it? For each task, ε_T bounds the probability that a Haar random (uniformly distributed) unitary on the task's n_T qubits lies in its *modeled acceptance event* M_T. This is the event its first,

unsharpened bound covers (App. C), taken at the identity relabeling. In general it is neither the shipped test's event nor $A(T)$ (§3). The first bounds in Tab. 6 and Fig. 3 (App. B) range from 0.25 to 10^{-1533}. Each is closed form, from a lemma instrument or, for three tasks, a task specific bound, with proofs in App. C. No first bound is a Monte Carlo estimate.

We say a task reaches 99% rejection when a bound of at most 10^{-2} holds for the event in question. Of the 86 tasks, 52 do so for M_T on their first bound. The other 34 are escalated (first bound above 10^{-2}) and carry a sharpened bound. For at most 17 of them, a bound below 10^{-2} is shown only for the shipped test's event, not for M_T. Counting each task on the event its row names, at least 82 of the 86 reach 99% rejection. On M_T alone, at least 69 do. An empirical Haar calibration of all 86 rows (App. D) samples a predicate of its own, in general neither M_T nor the shipped test. It puts its 95% upper bound below 10^{-2} on 85, all but T6.6, whose sampler yields 2^{-4} by construction. On 22 of the 85 it does so trivially because that predicate has measure zero.

The modeled event and the tested one differ in both directions. On six functional rows (the other functional rows were not compared this way), the published ε_T exceeds a bound on the shipped test's event by a factor 10^3 to 2.1×10^{36}. On T1.17, where the shipped test also accepts relabeled copies of the target, the test's event has a probability no smaller than ε_T. No lemma inequality is violated, but reading ε_T as the false positive rate of the shipped runner would be wrong.

5 Auditing the Acceptance Region

The bound in §4 concerns random unitaries. The audit asks which errors each design choice of the runner prevents and which it admits, on structured probes and on the baseline generations of §6 (Tab. 3). A generation is correct if it matches the canonical target on every declared input up to 10^{-6}, and correct up to relabeling if one qubit permutation achieves this. Of the $3{,}870$ generations, $3{,}475$ can be graded this way, without custom, stabilizer, and mid-circuit measurement tasks and extraction or simulation failures.

Inputs and predicate. A check at $|0 \cdots 0\rangle$ alone accepts 82 generations that fail the exact ground truth even up to relabeling, and the runner accepts none of them. Of the 82, 42 are wrong under their prompts. Checking every declared input, or each basis state separately on unitary tasks, rejects 39 of the 42. Only the full unitary check rejects the other 3, which differ from the target in relative phases. Of the other 40, 30 lie on T4.12, whose prompt does not fix the target, and 1 is a textbook variant. The remaining 9 follow prompts that contradict themselves or use a factor-two angle convention that neither form matches. The disguised copy, built for one input, passes on 16 of the 70 tasks without free parameters, all among the 17 judged on a single input. On the 10 tasks that ban prebuilt state preparation by gate name, none of the 420 extractable generations calls a banned name.

Relabeling. All 52 quantum Fourier transform generations that pass only under relabeling contain the requested SWAP gates and equal the textbook transform in the opposite bit order. The prompts fix that order only by asking for a Qiskit circuit. After two coding passes by different LLMs and 3 tasks decided from the prompt text (T2.16 overriding both), 29 tasks remain whose prompts name no qubit roles or order. Restricting relabeling to them removes the 8 generations wrong under the prompt's labels and the 2 ambiguous ones. The restricted rule keeps the other 66 generations (only 31 if T2.16 fixes roles) and changes the rank of four models, but reverses no pair the paired test separates.

Threshold and tolerances. Tolerances can only add acceptances. On the baseline, global phase and ancilla extension decide no verdict and relabeling decides 76. The predicate accepts an R_y rotation appended to a canonical solution up to exactly the radius φ_T of §3.3 on 25 of 26 fidelity tasks and up to 0.0897 on T6.14. The radius is 0.2003 at $\theta_T = 0.99$ and 0.0633 at 0.999. Types that read only output probabilities accept any final phase. Among accepted functional generations with a unitary operator, a relative phase difference occurs only on T3.15, whose prompt allows it.

What the runner itself adds. The runner's own paths for wider candidates decide none of the $3{,}870$ verdicts. Transpiled at the default setting of Qiskit 2.1.1, which the release pins, 210 accepted generations fail after the transpiler moves SWAP gates into an output permutation. The runner does not read it. Each equals the original once the permutation is undone. On four tasks every canonical output is a tie, so 180 acceptances rest on floating point tie breaking.

Table 3: What each design choice prevents and admits. FA: false acceptance (§2). Generations are graded against the exact ground truth. Counts on the unit in column 2.

Design choice	Measured on	Prevents	Admits	Control
Several declared inputs	3,475 gradable generations	39 FA that a check at $\|0\cdots0\rangle$ alone admits	a copy passes on 16 of 70 tasks, all judged on one input	all 82 gradable canonical solutions pass
Unitary check	same	three relative phase errors that a basis-state check admits	–	same
Relabeling, free layout	76 generations it decides	no false rejection shown	8 FA under the prompt's labels, 2 ambiguous, 66 on a bit order set by Qiskit's convention	52 of 52 QFT cases are the transform in the opposite bit order
Relabeling, prompts naming no roles	same	–	0 FA, keeps the 66	two LLM coding passes, Cohen's $\kappa = 0.95$, three tasks decided from the prompt text
Global phase, ancilla	325 and 365 generations; 3 wider ones	no verdict decided	–	extra qubits return to $\|0\rangle$
Threshold θ_T	rotations on 26 canonical fidelity solutions; 325 fidelity graded mutants near θ_T	–	rotations pass up to exactly φ_T on 25 of 26 (T6.14: up to 0.0897), and 36 of 989 wrong mutants pass functional tolerances	mutant verdicts match $[F \geq \theta_T]$, 325 of 325
Preparation ban by name	10 tasks; 420 extractable generations	–	offline transpiled copy passes; none calls prebuilt preparation by name	canonical passes, filler alone fails
Reading only the initial layout	3,158 accepted, transpiled at Qiskit's default	–	210 rejected for an output permutation alone, besides 360 by S_T, 90 on ties, 1 other	SWAP circuit, F $0.25 \to 1$
Argmax on ties	180 accepted on four tasks	–	verdict rests on floating point tie breaking	tie tolerant rule changes no baseline verdict
Cross acceptance	1,532 ordered task pairs	–	11 accepted, 9 matching	own task, 86 of 86

6 RESULTS

The audit in §5 asked where the runner accepts too much. Fixing the reference's qubit numbering probes the opposite direction: what a reference comparison would reject. We report the baseline and what the corpus can still tell apart, then ask whether the scoring criterion itself decides which models receive credit.

6.1 THE BASELINE

Each of the 15 models received all 86 tasks, three samples per task at temperature 0.2, under one fixed prompt template (App. H) and pinned provider endpoints, giving 3,870 generations. Pass rates per model span 55.8% to 91.5% (Tab. 4), with 81.6% overall. Of the 712 failed generations, 189 yield no extractable circuit and the runner rejects the other 523.

An exact McNemar test (McNemar, 1947) on per task majority verdicts, Benjamini–Hochberg adjusted (Benjamini & Hochberg, 1995) over 105 comparisons, separates 49 of the 105 model pairs

Table 4: Baseline per model, 258 generations each. Confidence interval (CI) at 95%: bias-corrected and accelerated bootstrap (Efron, 1987) over the 86 tasks, 10^4 replicates. C1–C6: pass rate (%) per category (state preparation, algorithms, decomposition, variational, arithmetic, error correction). Extr., Rej.: counts without an extractable circuit and rejected by the runner. Without 16: pass rate on the tasks other than the 16 that accept a disguised copy (§5). Flips: accepted generations within the scope of §6.2 that a fixed qubit numbering would reject, 53 in total.

Model	Pass (%)	95% CI	C1	C2	C3	C4	C5	C6	Extr.	Rej.	Without 16 (%)	Flips
Gemini 3.8 Flash	91.5	[84.1, 96.1]	100.0	98.2	96.7	68.8	100.0	90.7	2	20	91.9	3
Grok 4.6	90.3	[82.9, 95.0]	100.0	93.0	100.0	64.6	100.0	92.6	0	25	91.4	4
Qwen3.8 Max	89.5	[81.4, 94.2]	100.0	94.7	100.0	64.6	100.0	87.0	7	20	88.6	1
GLM 5.3	89.1	[83.3, 93.8]	100.0	98.2	96.7	64.6	100.0	83.3	5	23	88.1	2
GPT 6 Astra	88.4	[79.1, 93.0]	100.0	100.0	90.0	62.5	100.0	83.3	0	30	87.1	0
Fable 5.1	87.6	[80.6, 92.6]	85.2	96.5	100.0	60.4	100.0	87.0	7	25	87.6	5
DeepSeek V4 Pro	85.7	[77.1, 91.1]	100.0	87.7	96.7	62.5	97.6	81.5	5	32	86.2	3
Opus 5	85.3	[76.4, 90.7]	88.9	91.2	96.7	68.8	92.9	79.6	3	35	86.7	6
Kimi K3	85.3	[77.5, 90.7]	96.3	94.7	96.7	56.2	97.6	79.6	9	29	87.1	3
DeepSeek V4.1 Flash	82.6	[74.0, 88.4]	96.3	89.5	100.0	45.8	100.0	77.8	13	32	83.3	4
Sonnet 4.6	78.3	[69.4, 85.7]	88.9	91.2	83.3	45.8	100.0	68.5	5	51	81.0	6
Gemini 3 Flash	78.3	[69.0, 85.3]	85.2	91.2	80.0	54.2	85.7	75.9	8	48	80.5	3
GPT OSS 120B	74.0	[64.7, 81.0]	81.5	75.4	80.0	56.2	90.5	68.5	18	49	77.1	6
R1 Distill 70B	62.4	[52.3, 70.9]	74.1	70.2	56.7	35.4	71.4	68.5	66	31	64.8	3
Qwen3 Coder	55.8	[45.7, 65.1]	63.0	63.2	66.7	39.6	61.9	48.1	41	73	60.5	4

(Tab. 11, App. I). The nine leading models are not separated from one another, and only one of the 14 pairs of rank neighbors is. The ranking is best read in bands, not as an order of neighbors. Removing the 16 tasks that accept a disguised copy (column Without 16 in Tab. 4) changes the rank of five models, with tied models sharing the better rank. Every pair whose order or tie changes is one the paired test does not separate on the original data. With only 49 of 105 pairs separated, this shows no reordering of what the baseline resolves, not robustness.

What the corpus can still tell apart. Counting a task as solved by a model when at least two of its three samples pass, all 15 models solve 38 tasks, none solves 3, and 45 remain. On these 45 tasks, pass rates spread over 65.2 points, against 35.7 points on all 86 tasks. App. K.5 traces the three unsolved tasks to an incomplete specification (T4.11, T4.12) and, for 9 of 44 failures on T4.15, to parameter names.

6.2 Effect of the scoring criterion on credit

In an ablation, fixing the reference's qubit numbering would withdraw credit from 14 of the 15 models, between one and six generations each (Flips in Tab. 4). The scope holds 876 accepted generations on state and unitary fidelity tasks with a free layout and at most five qubits, and 53 of them would lose credit. Withdrawing this credit changes the rank of five models, counted as in §6.1.

Of the 53, 52 lie on the two quantum Fourier transform tasks. Each equals the target once the numbering of all qubits is reversed, that is, it is the transform with the opposite bit order. Each contains the requested SWAP gates, and the prompts fix which end of the register carries the highest order bit only by asking for a Qiskit circuit. We count these 52 as accepted only under relabeling, not as correct. The 53rd, on T3.13, swaps q_0 with q_2 and q_1 with q_3 where the prompt asks to reverse the order of all four qubits. A relabeling maps one onto the other, yet the generation is wrong.

Outside this scope, the runner accepts 23 more generations only under relabeling. On T4.6, 14 again differ from the target only in the bit order of the prompt's kets, which Qiskit's convention fixes. Two on T4.8 and three on T5.7 violate qubit roles the prompt names. The two on T6.6, which a check by three software development kits (SDKs) rejects (App. G), fail all four named stabilizers without relabeling. The last two, on T4.16, pass because relabeling flips the sign of the generator. They match the prompt's fermionic definition, whose Jordan–Wigner form has the opposite sign from the Pauli form the prompt writes out. Over all 3,158 accepted generations, relabeling decides credit in 76. Of these, 66 follow a bit order the prompts fix only by asking for a Qiskit circuit. Another 8 are wrong under the prompt's labels, and 2 rest on a prompt whose two forms disagree. The count therefore establishes no false rejection that relabeling prevents. Whether the 66 are correct turns on that convention, and for the 2 it turns on which of the prompt's two forms of the generator holds.

7 LIMITATIONS

The measurements of §5 and §6 come from one runner, and those on model generations from one prompt template and one pass over 15 models. The paragraphs below name what this leaves unmeasured and where the runner departs from the region.

A random null, not model errors. The bound ε_T caps the chance that a uniformly random (Haar) unitary falls into the modeled acceptance event of §4. The disguised copies and small rotations of §5 are structured, not random, so ε_T says nothing about how often such candidates pass.

Where the runner departs from the region. A wider candidate passes a unitary type only with a recovered layout and up to 10 qubits, and an exact qubit count in S_T excludes ancillas. These limits can only reject. Without a fixed or a recovered layout, the state type traces out the highest indexed qubits of a wider candidate of up to 12 qubits. The candidate passes if the reduced state is pure and close to the target, whatever the traced qubits hold. With a recovered layout, both state types instead project the extra qubits onto $|0\rangle$ and renormalize. Both paths can accept a candidate that Eq. (1) rejects and no bound covers them, but on the baseline neither decides a verdict (§5). On the 16 variational tasks, the runner binds test values to parameters by name and every unmatched parameter to 0.0. If no name matches and the counts agree, it binds in sorted name order instead. The sorted fallback carries 55 of the 408 accepted generations there, on nine tasks, and renaming parameters turns 9 failures on T4.15 into passes. Whether the 55 would fail without it is not measured.

Not measured. Verdicts come from noiseless simulation, so device noise is not measured. Transpilation is measured only for Qiskit, without routing to a device (§5). Qubit roles were coded, not probed by swapping, so the eight generations that §6 finds wrong under the prompt's labels are only what the census caught. A catalog of 1,475 edge-case instances is enumerated but was never run, and appended rotations probe the threshold instead (§5).

One pass, correlated samples. The baseline is one pass of three samples per task, collected from 10 to 12 September 2026 against one pinned endpoint per model. Run-to-run, prompt, and provider variation is therefore unknown, and Tab. 4 compares served endpoints rather than model weights. Tab. 4 and the paired test take the task as unit, but the Wilson intervals of App. E, G and K do not. No appendix models how this clustering affects its intervals.

Flawed prompts. Six tasks have prompts that leave the target open, contradict themselves, or make parameter names decisive (T3.6, T4.4, T4.11, T4.12, T4.15, T4.16), and they stay in Tab. 4. Without them, five models change rank, the paired test separates 44 instead of 49 pairs, and no separated pair reverses.

Contamination. Of the 86 tasks, 83 follow a cited published construction, so a passing generation may recall a known circuit rather than derive it. Scoring by membership cannot tell the two apart. The split by provenance in App. I rests on only three tasks of our own design and is descriptive only.

8 CONCLUSION

A code verifier can misjudge a solution in two ways, and every acceptance a tolerated equivalence adds is a potential false acceptance. QuARC makes this tradeoff explicit for quantum circuits. It scores a circuit by membership in an explicit acceptance region, closed under the transformations each task admits, rather than by distance to one reference. A Haar null model bounds the chance that a random unitary falls into a modeled acceptance event, at most 1% for at least 69 of the 86 tasks. A check at the all-zero input alone accepts 42 baseline generations wrong under their prompts (§5), and the runner accepts none. On this baseline the benefit comes from the predicate, while the tolerances decide few verdicts. A disguised copy of the canonical output passes 16 of the 70 tasks without free parameters, all judged at a single input. Relabeling admits eight generations wrong under the prompt's labels, and no false rejection it prevents is established. Restricting it to prompts naming no qubit roles or order would remove them. Across 15 models, a paired test separates only 49 of 105 pairs, so the ranking is best read in bands. For builders, the audit suggests checking several inputs or the unitary, enforcing bans structurally, restricting relabeling to such prompts, and reading the permutation a transpiler records. Measurements on generations rest on one prompt template and one pass, and all verdicts on noiseless simulation (§7). Since tolerances can only add acceptances, every new tolerance needs its own audit, as do further prompts and devices.

AI USE STATEMENT

At several stages of this work, the authors used generative AI tools, namely LLM-based coding and writing agents, to aid and improve their work. The 3,870 generations analyzed in §6 are experimental data and are not covered here. The core idea, the conceptual framework and the hypotheses come from the authors, and the agents challenged them. The authors refined the methodology and experimental design through feedback from the agents. The authors selected the tasks, mostly from published constructions, and developed them, including canonical solutions and tests, together with the agents, which checked, revised and questioned the drafts. The agents consolidated the task registry. The authors formulated the lemmas and precision bounds and wrote the proofs in conjunction with the agents, which assisted with ideation, proof steps and red-teaming. The agents sped up the coding workflows, drafted code for the measurement, audit and reproduction scripts, and assembled them into the released artifact. In collaboration with the authors, the agents coded which prompts fix qubit roles, classified the 76 generations that pass only after relabeling (§6), and interpreted the results. They also translated working notes from another language into English. They drafted a structure for the paper, edited all sections for a consistent style, and searched and summarized the literature.

Generative AI tools were not used to synthesize data. Random unitaries and the audit inputs of §5 come from seeded or deterministic scripts.

The authors have reviewed all significant AI-assisted work. They read the paper text and the proofs and regularly checked the code. Reviewer agents additionally checked text and numbers against the released data and reviewed the proofs. The agents come from a provider whose models the authors evaluate.

ETHICS STATEMENT

This work involves no human subjects and no personal data. Its tasks and generations are quantum programs and model responses. We collected the 3,870 generations of §6 through a commercial application programming interface (API) service under its terms of use, with one pinned provider endpoint per model. Tab. 4 therefore compares served endpoints rather than model weights, as §7 notes. The agents that assisted this work come from a provider whose models are among those evaluated in Tab. 4. Beyond the tool use disclosed above, no model provider funded this work or took part in its design or analysis. The dataset carries 160 short quotations from the primary sources behind the tasks (44 distinct sentences), each at most 140 characters and each attributed to its source. Longer attributed excerpts are excluded from the data license and remain under the terms of their sources. The quotations have not been legally reviewed.

REPRODUCIBILITY STATEMENT

The supplementary material contains an anonymized archive of the benchmark release, version 1.2.0. It holds the 86 tasks with canonical solutions and tests, the runner, and the 3,870 raw generations of §6 with their re-scored verdicts. §3 defines the acceptance region and the runner, §4 the tasks and bounds, and §5 the audit probes. App. A lists the tasks, Apps. B and C give the bounds and proofs, and App. D reports the Haar calibration and its seed rule. Apps. E to L cover the mutation calibration, further validations, the prompt template, per-tier results, per-model error breakdowns, a model run analysis, and failure patterns.

The archive includes the scripts behind the bounds, calibrations, statistics, and the audit. It declares the software environment as version ranges and records the exact versions of one environment in which parts of the results were re-run. A full Haar sampling pass took about 339 processor hours on an M1 Pro, and new model generations need paid API access. Both therefore ship as scripts together with their recorded result rows.

REFERENCES

Dorit Aharonov, Andris Ambainis, Julia Kempe, and Umesh Vazirani. Quantum walks on graphs. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pp. 50–59, 2001.

Gian-Luca R Anselmetti, David Wierichs, Christian Gogolin, and Robert M Parrish. Local, expressive, quantum-number-preserving VQE ansätze for fermionic systems. *New Journal of Physics*, 23(11): 113010, 2021.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Hafiz Md. Hasan Babu, Lafifa Jamal, Sayanton Vhaduri Dibbo, and Ashis Kumer Biswas. Area and delay efficient design of a quantum bit string comparator. In *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 51–56, 2017. doi: 10.1109/ISVLSI.2017.130.

Anindita Banerjee and Anirban Pathak. An analysis of reversible multiplier circuits. *arXiv preprint arXiv:0907.3357*, 2009.

Adriano Barenco, Charles H Bennett, Richard Cleve, David P DiVincenzo, Norman Margolus, Peter Shor, Tycho Sleator, John A Smolin, and Harald Weinfurter. Elementary gates for quantum computation. *Physical Review A*, 52(5):3457, 1995.

Andreas Bärtschi and Stephan Eidenbenz. Deterministic preparation of Dicke states. In *International Symposium on Fundamentals of Computation Theory*, pp. 126–139. Springer, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Charles H Bennett and Stephen J Wiesner. Communication via one-and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69(20):2881, 1992.

Charles H Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70(13):1895, 1993.

Ethan Bernstein and Umesh Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26 (5):1411–1473, 1997. doi: 10.1137/S0097539796300921.

Michel Boyer, Gilles Brassard, Peter Høyer, and Alain Tapp. Tight bounds on quantum searching. *Fortschritte der Physik: Progress of Physics*, 46(4-5):493–505, 1998.

Sergey Bravyi and Alexei Kitaev. Universal quantum computation with ideal Clifford gates and noisy ancillas. *Physical Review A*, 71(2):022316, 2005.

Lukas Burgholzer and Robert Wille. Advanced equivalence checking for quantum circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(9):1810–1824, 2020.

Lukas Burgholzer and Robert Wille. QCEC: A JKQ tool for quantum circuit equivalence checking. *Software Impacts*, 7:100051, 2021.

Lukas Burgholzer, Richard Kueng, and Robert Wille. Random stimuli generation for the verification of quantum circuits. In *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, pp. 767–772, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chanez, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Benoît Collins and Piotr Śniady. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

David Deutsch and Richard Jozsa. Rapid solution of problems by quantum computation. *Proceedings: Mathematical and Physical Sciences*, pp. 553–558, 1992.

Thomas G Draper, Samuel A Kutin, Eric M Rains, and Krysta M Svore. A logarithmic-depth quantum carry-lookahead adder. *arXiv preprint quant-ph/0406142*, 2004.

Wolfgang Dür, Guifre Vidal, and J Ignacio Cirac. Three qubits can be entangled in two inequivalent ways. *Physical Review A*, 62(6):062314, 2000.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

Austin G Fowler, Simon J Devitt, and Lloyd CL Hollenberg. Implementation of Shor's algorithm on a linear nearest neighbour qubit array. *arXiv preprint quant-ph/0402196*, 2004.

Austin G Fowler, Matteo Mariantoni, John M Martinis, and Andrew N Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86(3):032324, 2012.

Daniel Gottesman. *Stabilizer Codes and Quantum Error Correction*. PhD thesis, California Institute of Technology, 1997.

Daniel M Greenberger, Michael A Horne, and Anton Zeilinger. Going beyond Bell's theorem. In *Bell's Theorem, Quantum Theory and Conceptions of the Universe*, pp. 69–72. Springer, 1989.

Xiaoyu Guo, Minggu Wang, and Jianjun Zhao. QuanBench: Benchmarking quantum code generation with large language models. In *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 2657–2669. IEEE, 2025.

Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical Review Letters*, 103(15):150502, 2009.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

HV Jayashree, Himanshu Thapliyal, Hamid R Arabnia, and Vinod Kumar Agrawal. Ancilla-input and garbage-output optimized design of a reversible quantum integer multiplier. *The Journal of Supercomputing*, 72(4):1477–1493, 2016.

Mingyoung Jeng, Alvir Nobel, Vinayak Jha, David Levy, Dylan Kneidel, Manu Chaudhary, Ishraq Islam, Muhammad Momin Rahman, and Esam El-Araby. Generalized quantum convolution for multidimensional data. *Entropy*, 25(11):1503, 2023.

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.

Samuel A Kutin, David Petrie Moulton, and Lawren M Smithline. Computation at a distance. *arXiv preprint quant-ph/0701194*, 2007.

Xiaoyu Li, Guowu Yang, Carlos Manuel Torres Jr, Desheng Zheng, and Kang L Wang. A class of efficient quantum incrementer gates for quantum circuit synthesis. *International Journal of Modern Physics B*, 28(01):1350191, 2014.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572, 2023.

Seth Lloyd. Universal quantum simulators. *Science*, 273(5278):1073–1078, 1996.

Dmitri Maslov. Advantages of using relative-phase Toffoli gates with an application to multiple control Toffoli optimization. *Physical Review A*, 93(2):022311, 2016. doi: 10.1103/PhysRevA.93.022311.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Elizabeth S Meckes. *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press, 2019.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.

Taku Mikuriya, Tatsuya Ishigaki, Masayuki Kawarada, Shunya Minami, Tadashi Kadowaki, Yohichi Suzuki, Soshun Naito, Shunya Takata, Takumi Kato, Tamotsu Basseda, et al. QCoder benchmark: Bridging language generation and quantum hardware through simulator-based feedback. In *Proceedings of the 18th International Natural Language Generation Conference*, pp. 743–752, 2025.

Mikko Möttönen, Juha J. Vartiainen, Ville Bergholm, and Martti M. Salomaa. Transformation of quantum states using uniformly controlled rotations. *Quantum Information and Computation*, 5(6): 467–473, 2005. ISSN 1533-7146.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, Cambridge, 10th edition, 2010. ISBN 9781107002173.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020.

Qiskit Contributors. Qiskit: EfficientSU2 ansatz library, 2024a. URL https://github.com/Qiskit/qiskit/blob/main/qiskit/circuit/library/n_local/efficient_su2.py. Standard library implementation of the EfficientSU2 ansatz used for T4.14.

Qiskit Contributors. Qiskit: RealAmplitudes ansatz library, 2024b. URL https://github.com/Qiskit/qiskit/blob/main/qiskit/circuit/library/n_local/real_amplitudes.py. Standard library implementation of the RealAmplitudes ansatz used for T4.13.

Qiskit Contributors. Qiskit: U gate (UGate) standard library implementation, 2024c. URL https://github.com/Qiskit/qiskit/blob/main/qiskit/circuit/library/standard_gates/u.py. Reference implementation for the parametrized single qubit unitary, used as the canonical decomposition source for T3.6.

Robert Raussendorf and Hans J Briegel. A one-way quantum computer. *Physical Review Letters*, 86 (22):5188, 2001.

Norbert Schuch and Jens Siewert. Natural two-qubit gate for quantum computation using the XY interaction. *Physical Review A*, 67(3):032301, 2003.

Maria Schuld and Francesco Petruccione. *Supervised learning with quantum computers*, volume 17. Springer, 2018.

Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.

Dan Shepherd and Michael J Bremner. Temporally unstructured quantum computation. *Proceedings: Mathematical, Physical and Engineering Sciences*, pp. 1413–1439, 2009.

Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509, 1997. doi: 10.1137/S0097539795293172.

Daniel R. Simon. On the power of quantum computation. *SIAM Journal on Computing*, 26(5): 1474–1483, 1997.

Ali Slim, Haydar Hamieh, Jawad Kotaich, Yehya Ghosn, Mahdi Chehimi, Ammar Mohanna, Hasan Abed Al Kader Hammoud, and Bernard Ghanem. QuanBench+: A unified multi-framework benchmark for LLM-based quantum code generation. *arXiv preprint arXiv:2604.08570*, 2026.

John A Smolin and David P DiVincenzo. Five two-bit quantum gates are sufficient to implement the quantum Fredkin gate. *Physical Review A*, 53(4):2855, 1996.

Vlatko Vedral, Adriano Barenco, and Artur Ekert. Quantum networks for elementary arithmetic operations. *Physical Review A*, 54(1):147, 1996.

Sanjay Vishwakarma, Francis Harkins, Siddharth Golecha, Vishal Sharathchandra Bajpe, Nicolas Dupuis, Luca Buratti, David Kremer, Ismael Faro, Ruchir Puri, and Juan Cruz-Benito. Qiskit HumanEval: An evaluation benchmark for quantum code generative models. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 1, pp. 1169–1176. IEEE, 2024.

James D Whitfield, Jacob Biamonte, and Alán Aspuru-Guzik. Simulation of electronic structure Hamiltonians using quantum computers. *Molecular Physics*, 109(5):735–750, 2011.

Rui Yang, Ziruo Wang, Yuntian Gu, Yitao Liang, and Tongyang Li. QCircuitBench: A large-scale dataset for benchmarking quantum algorithm design. In *NeurIPS 2025 Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=NkiLldW2bi.

Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 396–411, 2020.

Karol Zyczkowski and Hans-Jürgen Sommers. Induced measures in the space of mixed quantum states. *Journal of Physics A: Mathematical and General*, 34(35):7111–7125, 2001.

APPENDIX

A TASK REGISTRY

Tab. 5 reproduces selected columns of the task registry together with a citation column. The metadata of the released dataset records a verification that predates the September sharpening of the tasks and names its method, "Independent mathematical verification with numpy (no Qiskit dependency)"; this appendix reports that entry and does not repeat the verification. The acceptance condition itself is not reproduced here. Its threshold is printed only where the released dataset and the registry agree on it and the shipped test reads it. The Test type column maps each task to one of the nine acceptance region constructions: state fidelity, unitary fidelity, functional, probability, parametric state fidelity, parametric unitary fidelity, custom, stabilizer eigenvalue, and code space unitary fidelity. The composite flag and k value identify tasks whose acceptance region decomposes into $k \geq 2$ functionally distinct sub components, validated separately in Tab. 9.

The task identifiers are not contiguous. They are the historical draft numbering, and an identifier is retired rather than reused once its task is dropped, so the 86 shipped tasks leave 22 gaps across the six tiers. The retired identifiers are T1.1, T1.3, T1.4, T1.6, T1.7, T1.9, T1.13, T1.16, T2.4, T2.5, T2.8, T2.13, T2.14, T2.15, T2.20, T3.1, T3.4, T3.7, T3.9, T3.10, T4.1 and T6.7. The delivered dataset records a removal reason for 12 of them, in its `metadata.v2_changes` block. Every recorded reason names a similarity to other tasks, either as a named pseudo replicate or as a similar gate structure. The remaining 10 identifiers carry no such entry. None of the retired identifiers appears among the delivered tasks. The registry below then lists every delivered task, matching the identifier set in `artifacts/task_registry/task_registry.csv`.

Table 5: Task registry, dataset v1.2.0. Tasks are organized in six tiers (T1: state preparation, T2: standard algorithms, T3: gate decomposition, T4: variational circuits, T5: quantum arithmetic, T6: error correction primitives). Composite tasks ($k \geq 2$ functionally distinct sub components) are validated separately by the composite task pipeline (Tab. 9); a check mark in the comp. column flags such a task, and the k column is filled for those tasks only. Abbreviations in the Test type column, one per verification type: `SF` state fidelity, `UF` unitary fidelity, `PSF` parametric state fidelity, `PUF` parametric unitary fidelity, `Func` functional, `Prob` probability, `CSU` code space unitary fidelity, `StabE` stabilizer eigenvalue, `Cust` custom. The n_q column is the qubit count of the task. The K column reproduces the value that the released registry records for each task. App. C defines the symbol as the number of conditions the shipped specification of a task prescribes, its test cases, expected peaks, or listed outputs, and for each of the 38 tasks that carry a value the recorded value matches that definition. Where the registry records no value, the cell carries a placeholder mark ($--$), which describes that record and is not a statement about the task's acceptance test. Each body column, from Task through k, is reproduced from `artifacts/task_registry/task_registry.csv`, which ships with the benchmark. The Thr. column prints the acceptance threshold only where the released dataset and the registry agree on it and the shipped test reads it; what the value bounds depends on the verification type. For 46 tasks the registry's `threshold` column and the dataset field `test.threshold` both record a value and the two values are equal; on 2 of them, T2.22 and T6.19, the deciding branch of the shipped test does not read that field, so those cells are left empty and 44 values are shown. For 25 tasks only the registry's `threshold` column records one, while the released dataset carries no `test.threshold` field, so the cell is left empty; for 15 tasks neither records one. The Cite column is a hand curated literature attribution and is not derived from the task data. It ships as `artifacts/task_registry/task_citations.csv`, whose 86 rows reproduce the cells printed here, and as `artifacts/task_registry/citation_keys.csv`, which carries the bibliographic record of all 42 keys, 26 of them marked as compiled and not verified, so that each one resolves without the bibliography of this paper; the attribution record behind each of the 98 task key pairs ships with them in `artifacts/task_registry/attribution/`, where 89 rest on a located attestation and 9 on a documented rationale. It is a different field from `primary_sources` in the released dataset, which records what a task was constructed from: of the 83 cited rows, 60 name an author that `primary_sources` also names, 52 of them with matching years throughout. That agreement carries less weight than the count suggests, because every task lists the same textbook among its sources: 29 of the 60 agree only through that textbook, and 31 through some other work. The remaining 23 cite a source chosen by editorial decision, each recorded as such with its reason in the released file. An empty cell marks a task whose statement and acceptance test are constructions specific to QuARC; it is not a claim that the underlying algorithm is new. The three such tasks build on published work, recorded in the `primary_sources` field of the released dataset: T2.10 is a six state teleportation conjunction over the protocol of Bennett et al. (1993), and T2.17 and T2.18 are Simon variants with fixed hidden strings (Simon, 1997). Fifteen tasks fix the qubit layout by construction and are therefore exempt from the qubit relabeling invariance of §3.1: T1.12, T2.1, T2.3, T2.10, T2.11, T2.12, T2.17, T2.18, T2.19, T2.22, T2.24, T3.5, T3.12, T6.15 and T6.16. The harness runs these identity only, without the permutation retry it applies elsewhere. The harness caps its relabeling search at five qubits, so for the 7 tasks above that size that do not fix their layout the invariance is not guaranteed either.

Task	Name	n_q	Test type	K	comp.	k	Thr.	Cite
T1.2	Bell State Phi-	2	`SF`	$--$			0.99	Nielsen & Chuang (2010)
T1.5	GHZ State (3 qubits)	3	`SF`	$--$			0.99	Nielsen & Chuang (2010); Greenberger et al. (1989)
T1.8	W State (3 qubits)	3	`SF`	$--$			0.99	Dür et al. (2000)
T1.10	Phase Kickback	2	`SF`	$--$			0.99	Nielsen & Chuang (2010)
T1.11	SWAP Gate from CNOTs	2	`UF`	$--$			0.99	Nielsen & Chuang (2010)
T1.12	Toffoli Gate Decomposition	3	`UF`	$--$			0.99	Barenco et al. (1995); Nielsen & Chuang (2010)
T1.14	GHZ State (5 qubits)	5	`SF`	$--$			0.99	Nielsen & Chuang (2010); Greenberger et al. (1989)
T1.15	Dicke State D(4,2)	4	`SF`	$--$			0.99	Bärtschi & Eidenbenz (2019)

(continued on next page)

(continued)

Task	Name	n_q	Test type	K	comp.	k	Thr.	Cite
T1.17	1D Cluster State (5 qubits)	5	SF	– –			0.99	Raussendorf & Briegel (2001)
T2.1	Deutsch-Jozsa Balanced (3 qubits)	4	Func	16				Deutsch & Jozsa (1992); Nielsen & Chuang (2010)
T2.2	Deutsch-Jozsa Constant (3 qubits)	4	Func	16				Deutsch & Jozsa (1992); Nielsen & Chuang (2010)
T2.3	Bernstein-Vazirani (s=101)	4	Func	16				Bernstein & Vazirani (1997); Nielsen & Chuang (2010)
T2.6	Complete Grover (2q, $\|11\rangle$)	2	Prob	– –			0.99	Nielsen & Chuang (2010)
T2.7	QFT (3 qubits)	3	UF	– –			0.99	Nielsen & Chuang (2010)
T2.9	Simple Phase Estimation	3	Func	3				Nielsen & Chuang (2010)
T2.10	Quantum Teleportation	3	Cust	– –				
T2.11	Superdense Coding (10)	2	Func	4				Bennett & Wiesner (1992); Nielsen & Chuang (2010)
T2.12	Grover (3q, |101>)	3	Prob	8				Nielsen & Chuang (2010)
T2.16	QFT (4 qubits)	4	UF	– –			0.99	Nielsen & Chuang (2010)
T2.17	Simon's Algorithm (2-bit)	4	Cust	– –				
T2.18	Simon's Algorithm (3-bit)	6	Cust	– –				
T2.19	Bernstein-Vazirani (s=1011)	5	Func	32				Bernstein & Vazirani (1997)
T2.21	Quantum Walk (Coined, 4 Positions)	3	Func	8				Aharonov et al. (2001)
T2.22	Grover Search with 2 Solutions (3 qubits)	3	Prob	2				Boyer et al. (1998)
T2.23	Heisenberg Hamiltonian Simulation (2 qubits)	2	UF	– –			0.99	Nielsen & Chuang (2010)
T2.24	Quantum Walk on 4-Position Cycle	3	Func	8				Aharonov et al. (2001)
T2.25	Heisenberg Chain 3q (Trotter)	3	UF	– –			0.99	Nielsen & Chuang (2010)
T2.26	Phase Estimation 4q (S-Gate)	4	Func	3				Nielsen & Chuang (2010)
T3.2	CRy(pi/4) Decomposition	2	UF	– –			0.99	Barenco et al. (1995)
T3.3	iSWAP Gate	2	UF	– –			0.99	Schuch & Siewert (2003)
T3.5	Fredkin (CSWAP) Gate	3	Func	8				Smolin & DiVincenzo (1996); Nielsen & Chuang (2010)
T3.6	Controlled-U Gate	2	UF	– –			0.99	Qiskit Contributors (2024c)
T3.8	XX Gate (theta=pi/4)	2	UF	– –			0.99	Lloyd (1996)
T3.11	Controlled Phase CP(pi/3)	2	UF	– –			0.999	Nielsen & Chuang (2010)
T3.12	Toffoli from CZ	3	UF	– –			0.99	Nielsen & Chuang (2010)
T3.13	Linear SWAP Network (4 qubits)	4	UF	– –			0.99	Kutin et al. (2007); Fowler et al. (2004)
T3.14	Multi-Controlled Ry Gate (CCRy)	3	UF	– –			0.99	Barenco et al. (1995)
T3.15	Relative-Phase Toffoli	3	Func	8				Maslov (2016)
T4.2	Hardware Efficient Ansatz (4q, 2L)	4	PSF	– –				Kandala et al. (2017)
T4.3	QAOA Mixer (3 qubits)	3	PSF	– –			0.99	Farhi et al. (2014)
T4.4	QAOA Cost Layer (MaxCut Triangle)	3	PUF	– –			0.99	Farhi et al. (2014)
T4.5	Complete QAOA (3q, 1L)	3	PSF	– –			0.99	Farhi et al. (2014)
T4.6	Single Excitation Ansatz	2	PUF	– –			0.99	Anselmetti et al. (2021)
T4.7	Data Encoding (Angle)	3	PSF	– –			0.99	Schuld & Petruccione (2018)
T4.8	Data Encoding (Amplitude)	2	PSF	– –			0.99	Möttönen et al. (2005)

(continued on next page)

(continued)

Task	Name	n_q	Test type	K	comp.	k	Thr.	Cite
T4.9	Variational Classifier Layer	4	PSF	−−			0.99	Schuld et al. (2020)
T4.10	IQP Circuit	4	PSF	−−			0.99	Shepherd & Bremner (2009)
T4.11	Quantum Convolutional Layer	4	PSF	−−			0.99	Cong et al. (2019)
T4.12	Quantum Pooling Layer	4	PUF	−−			0.99	Cong et al. (2019)
T4.13	RealAmplitudes Ansatz (4 qubits, 2 layers)	4	PSF	−−			0.99	Qiskit Contributors (2024b)
T4.14	EfficientSU2 Ansatz (4 qubits, 1 layer)	4	PSF	−−			0.99	Qiskit Contributors (2024a)
T4.15	Data Reuploading Layer	2	PSF	−−			0.99	Pérez-Salinas et al. (2020)
T4.16	UCCSD Single Excitation Operator	4	PUF	−−			0.99	Whitfield et al. (2011)
T4.17	QAOA 4q 2-Layer (MaxCut)	4	PSF	−−			0.99	Farhi et al. (2014)
T5.1	Quantum Half Adder	3	Func	4	✓	2		Vedral et al. (1996); Nielsen & Chuang (2010)
T5.2	Quantum Full Adder	4	Func	8	✓	3		Vedral et al. (1996)
T5.3	Quantum Incrementer (3-bit)	3	Func	8				Li et al. (2014)
T5.4	Quantum Decrementer (3-bit)	3	Func	8				Li et al. (2014)
T5.5	Ripple Carry Adder (2-bit)	6	Func	16				Vedral et al. (1996)
T5.6	Quantum Comparator (2-bit)	5	Func	16				Babu et al. (2017)
T5.7	Modular Addition (mod 3)	4	Func	9				Vedral et al. (1996)
T5.8	Controlled 2-bit Addition	5	Func	32				Vedral et al. (1996)
T5.9	Quantum Multiplier (1-bit x 1-bit)	3	Func	4				Nielsen & Chuang (2010)
T5.10	Quantum Multiplier (2-bit x 1-bit)	5	Func	8	✓	2		Banerjee & Pathak (2009); Nielsen & Chuang (2010)
T5.11	Controlled Incrementer (2-bit)	3	Func	8				Li et al. (2014); Jeng et al. (2023)
T5.12	Quantum Subtractor (2-bit)	5	Func	16				Vedral et al. (1996)
T5.13	Quantum Multiplier (2-bit x 2-bit)	8	Func	16	✓	3		Jayashree et al. (2016); Banerjee & Pathak (2009)
T5.14	Carry Lookahead Adder (2-bit)	6	Func	16				Draper et al. (2004)
T6.1	Bit-Flip Code Encoding	3	Func	2				Nielsen & Chuang (2010)
T6.2	Bit-Flip Code Syndrome	5	Func	4				Nielsen & Chuang (2010)
T6.3	Phase-Flip Code Encoding	3	UF	−−			0.99	Nielsen & Chuang (2010)
T6.4	Shor Code Encoding	9	SF	−−			0.999	Nielsen & Chuang (2010)
T6.5	Steane Code Encoding	7	SF	−−			0.999	Nielsen & Chuang (2010)
T6.6	5-Qubit Code Space Preparation	5	StabE	−−			0.99	Gottesman (1997)
T6.8	Logical Z on Bit-Flip Code	3	CSU	−−			0.99	Nielsen & Chuang (2010)
T6.9	Surface Code Z-Plaquette	5	Func	5	✓	4		Fowler et al. (2012)
T6.10	Surface Code X-Star	5	Func	5				Fowler et al. (2012)
T6.11	Repetition Code (5 qubits)	5	Func	2	✓	4		Nielsen & Chuang (2010)
T6.12	Transversal CNOT (Bit-Flip)	6	Func	4				Nielsen & Chuang (2010)
T6.13	Error Detection Circuit	4	Func	8				Nielsen & Chuang (2010)

(continued on next page)

(continued)

Task	Name	n_q	Test type	K	comp.	k	Thr.	Cite
T6.14	Magic State Preparation (T)	1	`SF`	−−			0.999	Nielsen & Chuang (2010); Bravyi & Kitaev (2005)
T6.15	Shor's Algorithm (N=15)	7	`Prob`	16				Shor (1997); Nielsen & Chuang (2010)
T6.16	HHL Algorithm (2x2 System)	4	`Func`	2				Harrow et al. (2009)
T6.17	Transversal Hadamard (Steane Code)	7	`UF`	−−			0.999	Nielsen & Chuang (2010)
T6.18	CSS Code Stabilizer Measurement	3	`Func`	4				Nielsen & Chuang (2010)
T6.19	Full QEC Cycle: Bit-Flip Code	5	`Func`	2	✓	4		Nielsen & Chuang (2010)

B Closed Form Lemma Bounds

Each acceptance region $A(T)$ in the benchmark carries an analytically derived bound on a model acceptance event under a Haar random null adversary (App. D). The bound is built from one of the Lemma instruments of App. C or, for three tasks, from a task specific custom bound. Of the nine instruments L1 to L9, L7 is a Monte Carlo instrument and the others are closed form. All three L7 compound rows of Tab. 6 (T2.12, T2.22, T6.15) carry closed form Dirichlet tail bounds (remark in §C.9), so none of their bounds rests on a Monte Carlo estimate; on these three rows the estimates of the L7 sampler are superseded by those closed form bounds. L1 covers state fidelity and parametric state fidelity acceptance (Fubini–Study volume, Zyczkowski & Sommers, 2001). L2 (argmax uniformity, Mezzadri, 2006) combines with L3 (Dirichlet marginal, Zyczkowski & Sommers, 2001) into the L2,L3 compound for functional acceptance in the $K < d$ regime (on the seven rows with $K = d$ at the condition count $K_{\mathrm{c}} < d$ of their weakest test case, §C.1), and with L4 (joint argmax) into the L2,L4 compound for K fold joint functional acceptance. L5 covers unitary fidelity and parametric unitary fidelity acceptance via a Weingarten Markov bound. L6 covers multi output argmax and stabilizer eigenvalue acceptance. L7 covers probability acceptance by Monte Carlo estimation under the Dirichlet representation (Zyczkowski & Sommers, 2001), with a Hoeffding shot noise correction (Hoeffding, 1963). L8 covers code space unitary fidelity acceptance under a Haar adversary on the full space projected onto the code subspace. L9 covers *conjunctive* functional acceptance, where the harness requires every test case of a task to pass: the chain bound of §C.11 evaluates that conjunction in closed form. It sharpens the 14 `L2,L3` rows that the tabulated stand in form leaves above the escalation threshold, and it changes no released classification. The three custom bounds (T2.10, T2.17, T2.18) are documented in the proof bundle. Tab. 6 reports the per task assignment.

For task T2.22 (Grover with K=2) the joint Dirichlet tail bound 10^{-7} (remark in §C.9) is the tabulated value; the gap to the empirical estimate was an event mismatch of the calibration proxy sampler, resolved in App. D, and the bound lies far below the 0.01 threshold.

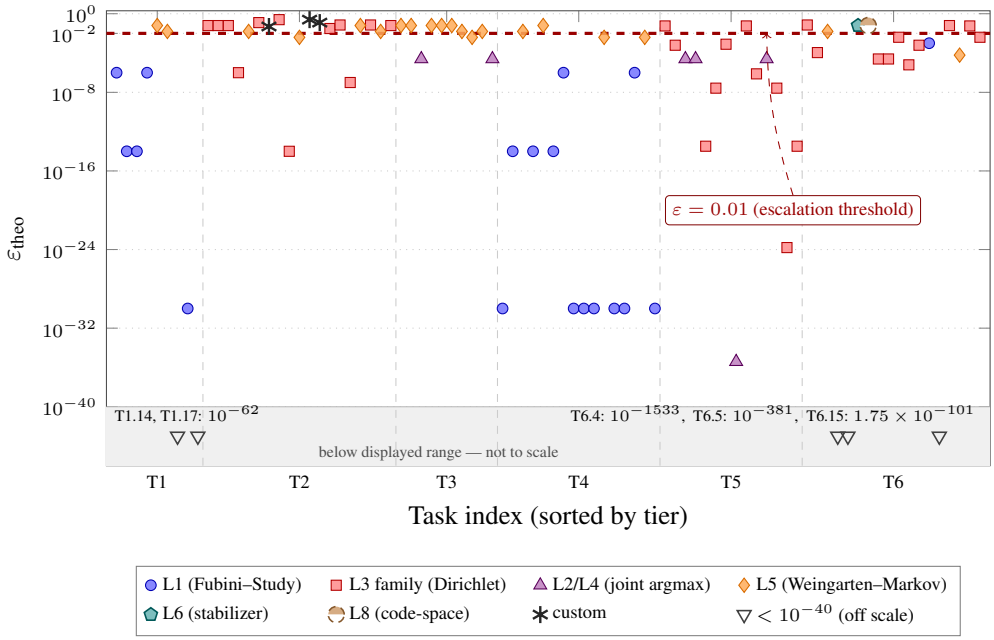


Figure 3: Closed form bound landscape across the 86 tasks, sorted by tier (T1 to T6). Each marker is a single task and the y axis shows the closed form $\varepsilon_{\text{theo}}$ from the per task Lemma assignment of Tab. 6. The dashed horizontal line marks the escalation threshold $\varepsilon = 0.01$. A task above this line is flagged for escalation, is sampled in the calibration of App. D at $N \geq 10^9$, and carries a sharpened bound in Tab. 6; Tab. 7 reports the calibration of all 86 tasks at mixed sample sizes, 68 of them at $N \geq 10^9$ and 18 at 10^6 to 10^7. That calibration evaluates a sampled predicate under a Haar null model and is not a measurement of the shipped test function (App. D); for T6.6 the sampler draws one fair bit for each of the n_{stab} stabilizer generators of the target code instead of a Haar state and so reproduces $2^{-n_{\text{stab}}}$ by construction. Thirty four tasks lie above the threshold: 15 L5 rows, 14 L2/L3 rows, the three custom bounds (T2.10 Teleportation, T2.17 and T2.18 Simon) and the two rows T6.6 (L6) and T6.8 (L8). Five of the L2/L3 rows (T2.9, T2.21, T2.24, T2.26, T6.16) joined that set with dataset v1.1.0, which changed their verification type. The L1 fidelity tasks are the opposite extreme: none of them escalates, and all but one sit at least four orders of magnitude below the threshold. Five tasks fall below the displayed range and are drawn as open downward triangles in the shaded band at the bottom. That band is not to scale and the marker height inside it carries no meaning; the true values are $\varepsilon_{\text{theo}} = 10^{-62}$ for T1.14 and T1.17, 1.75×10^{-101} for T6.15, 10^{-381} for T6.5 and 10^{-1533} for T6.4. The last two are $(1 - F^{\star})^{d-1}$ at $d = 128$ and $d = 512$ with $F^{\star} = 0.999$; both underflow double precision arithmetic and are evaluated exactly. The L3 family marker covers the `L2,L3`, `L3` and `L3,L7` rows; the sharpened bounds of Tab. 6 are not drawn.

Table 6: Lemma bound assignment per task, dataset v1.2.0. $\varepsilon_{\text{theo}}$ is the per task analytically derived bound on a model acceptance event under the Haar null (App. D), not a false positive rate of the shipped test: closed form for the L1–L6 and L8 instruments, the three Custom Bounds (T2.10, T2.17, T2.18), and the three L3,L7 rows T2.12, T2.22, and T6.15 (Dirichlet tail bounds on the events of dataset v1.2.0, remark in §C.9). The same symbol names the same bound in Tab. 7 and Fig. 3, except that for T2.12 and T6.15 Tab. 7 carries the bound of the earlier event their calibration sampled. The K column reproduces the registry value of Tab. 5, the number of conditions the shipped specification prescribes. Tasks with $\varepsilon_{\text{theo}} > 10^{-2}$ are flagged for escalation (34 of the 86) and sampled in the calibration of Tab. 7. For the 29 `L2,L3` and the 6 `L2,L4` rows the F^* column reproduces the field of the same name in the released lemma mapping: 1.00 on 31 rows, 0.9900 on T2.2, T6.2 and T6.11, and 0.9500 on T6.18. It agrees with the registry's `threshold` column on 19 of the 20 rows whose registry entry declares one and disagrees on T6.19, which carries 1.00 here against 0.99 in the registry; the remaining 15 declare none. Neither bound formula uses it ($K/\binom{d}{K}$ and $1/d!$ respectively), and the deployed tests of these rows compare output distributions case by case (§C.11) rather than a fidelity against a threshold. The value is not inert elsewhere: the sampled predicate of App. D is defined through it, and the rows that were sampled at $F^* = 1.00$ with a subset size below d (App. D) are the ones whose predicate has Haar measure exactly zero. The column and the sampler do not always agree: T5.1, T5.9 and T6.19 carry 1.00 here and were sampled at 0.99. The m column is the relabeling orbit of the declared target state, filled for the 8 single case state fidelity rows whose qubit count stays inside the harness permutation retry ($n \leq 5$, §C.3); where it runs the bound in force is $m\,\varepsilon_{\text{theo}}$. An empty cell means that no orbit is tabulated in this column for that row, not that the orbit is trivial; for the 20 L5 rows the orbit is priced in the $\log_{10}\varepsilon_{\text{s}}$ column instead. The five `L2,L3` rows T2.9, T2.21, T2.24, T2.26 and T6.16 carry no m because dataset v1.1.0 turned them from single case state fidelity into multi case functional tasks. On 10 of the 29 `L2,L3` rows, T2.1, T2.2, T2.3, T2.9, T2.11, T2.19, T2.21, T2.24, T2.26 and T6.16, $\varepsilon_{\text{theo}}$ is the single case form evaluated at the condition count K_{c} of the weakest case rather than at K (App. C); it does not price the conjunction over the remaining cases, which the L9 value in the $\log_{10}\varepsilon_{\text{s}}$ column does. The $\log_{10}\varepsilon_{\text{s}}$ column carries the base ten logarithm of the sharpened bound computed from the instruments of App. C wherever one has been computed: L9 (§C.11) for the 15 `L2,L3` rows, where for T2.9, T2.21, T2.24 and T2.26 the joint Dirichlet box probability of §C.11 lowers the first chain link, the moment bound of §C.7 for the 20 L5 rows, taken at order $k = 3$ for T3.2, T3.6 and T4.6, whose union over the relabeling orbit stays above the threshold at $k = 2$, and at $k = 2$ for the other 17 (T6.17 closes at $k = 1$ already; the smallest order rule of §C.7 is not applied to this row), the Beta tail of §C.8 for T6.6, the truncated unitary density of §C.10 for T6.8, and the task specific Beta tails of §C.12, §C.13 and §C.14 for the three Custom Bounds. Each cell is rounded up to one decimal, so the printed value is an upper bound on $\log_{10}\varepsilon_{\text{s}}$. Only escalation rows, L5 family rows, and T6.19 are filled; the remaining 46 stay empty. Six rows carry a sharpened bound yet do not escalate, for different reasons. For T6.19, the tabulated value sits below the threshold, but the union bound over the widened event that the runner effectively tests does not, so L9 is computed for it as well. The L5 family carries its moment bound regardless of escalation, so T2.16, T3.13, T4.12, T4.16 and T6.17 fill the column as well and the tabulated L5 rows are 20. The column changes no classification: $\varepsilon_{\text{theo}}$, the `Esc.` column and the released mapping are unchanged, and the escalation set stays at 34 of the 86. What it reports is how many of the 40 rows with a sharpened bound already lie below 10^{-2} — all 40 of them, T2.18 at $\log_{10}\varepsilon_{\text{s}} = -27.3$ included, with the two Simon rows priced against the rule the shipped test enforces. For the 20 L5 rows the column bounds the union over the qubit relabelings the harness admits, for a candidate on exactly the task's n qubits, and remains a bound on a model event (App. D); the orbit size m is computed per row from the declared target and is 1 wherever the task fixes the qubit layout or exceeds the harness permutation limit. For T2.17 and T2.18 the column prices the event the shipped test decides on the violating strings, where that test is stricter than the model; for T6.6 it prices the stabilizer condition of the shipped runner at the identity qubit layout, not the union over the qubit relabelings the harness also tries, while $\varepsilon_{\text{theo}} = 2^{-n_{\text{stab}}}$ prices the sign coincidence (§C.8); for the L9 rows it prices the conjunction over the test cases under the tolerances of the shipped runner; for the remaining instruments it prices the model event, and the scope caveat recorded in App. D applies to those rows unchanged. On the three Custom rows the F^* column carries the registry field of that name; it enters the bound of T2.10 ($\varepsilon_{\text{theo}} = 1 - F^*$) and not those of T2.17 and T2.18, whose tests require marginal mass at least 0.1 on each valid string and at most 0.01 on each orthogonality violating string, so for T2.17 the column is not its acceptance threshold.

Task	Lemma	d	K	F^*	$\varepsilon_{\text{theo}}$	$\log_{10}\varepsilon_{\text{s}}$	m	Esc.
T1.2	`L1`	4		0.9900	$1.00e-6$		1	
T1.5	`L1`	8		0.9900	$1.00e-14$		1	
T1.8	`L1`	8		0.9900	$1.00e-14$		1	
T1.10	`L1`	4		0.9900	$1.00e-6$		1	
T1.11	`L5`	4		0.9900	0.06313	-2.0		✓
T1.12	`L5`	8		0.9900	0.01578	-3.3		✓
T1.14	`L1`	32		0.9900	$1.00e-62$		1	
T1.15	`L1`	16		0.9900	$1.00e-30$		1	

(continued on next page)

(continued)

Task	Lemma	d	K	F^*	$\varepsilon_{\text{theo}}$	$\log_{10}\varepsilon_{\text{s}}$	m	Esc.
T1.17	L1	32		0.9900	$1.00e-62$		60	
T2.1	L2,L3	16	16	1.00	0.0625	-125.0		✓
T2.2	L2,L3	16	16	0.9900	0.0625	-123.6		✓
T2.3	L2,L3	16	16	1.00	0.0625	-125.0		✓
T2.6	L3	4		0.9900	$1.00e-6$			
T2.7	L5	8		0.9900	0.01578	-2.5		✓
T2.9	L2,L3	8	3	1.00	0.125	-13.7		✓
T2.10	Cust T2.10	8		0.9500	0.05	-3.7		✓
T2.11	L2,L3	4	4	1.00	0.25	-7.8		✓
T2.12	L3,L7	8	8	0.9900	$1.00e-14$			
T2.16	L5	16		0.9900	0.003946	-3.1		
T2.17	Cust T2.17	16		0.9900	0.25	-9.8		✓
T2.18	Cust T2.18	64		0.1000	0.125	-27.3		✓
T2.19	L2,L3	32	32	1.00	0.03125	-571.9		✓
T2.21	L2,L3	8	8	1.00	0.07143	-18.7		✓
T2.22	L3,L7	8	2	0.9000	$1.00e-7$			
T2.23	L5	4		0.9900	0.06313	-2.0		✓
T2.24	L2,L3	8	8	1.00	0.07143	-19.5		✓
T2.25	L5	8		0.9900	0.01578	-2.5		✓
T2.26	L2,L3	16	3	1.00	0.0625	-48.7		✓
T3.2	L5	4		0.9900	0.06313	-2.5		✓
T3.3	L5	4		0.9900	0.06313	-2.0		✓
T3.5	L2,L4	8	8	1.00	$2.48e-5$			
T3.6	L5	4		0.9900	0.06313	-2.5		✓
T3.8	L5	4		0.9900	0.06313	-2.0		✓
T3.11	L5	4		0.9990	0.06256	-2.1		✓
T3.12	L5	8		0.9900	0.01578	-3.3		✓
T3.13	L5	16		0.9900	0.003946	-4.0		
T3.14	L5	8		0.9900	0.01578	-2.8		✓
T3.15	L2,L4	8	8	1.00	$2.48e-5$			
T4.2	L1	16		0.9900	$1.00e-30$			
T4.3	L1	8		0.9900	$1.00e-14$			
T4.4	L5	8		0.9900	0.01578	-3.3		✓
T4.5	L1	8		0.9900	$1.00e-14$			
T4.6	L5	4		0.9900	0.06313	-2.5		✓
T4.7	L1	8		0.9900	$1.00e-14$			
T4.8	L1	4		0.9900	$1.00e-6$			
T4.9	L1	16		0.9900	$1.00e-30$			
T4.10	L1	16		0.9900	$1.00e-30$			
T4.11	L1	16		0.9900	$1.00e-30$			
T4.12	L5	16		0.9900	0.003946	-3.4		
T4.13	L1	16		0.9900	$1.00e-30$			
T4.14	L1	16		0.9900	$1.00e-30$			
T4.15	L1	4		0.9900	$1.00e-6$			
T4.16	L5	16		0.9900	0.003946	-3.4		
T4.17	L1	16		0.9900	$1.00e-30$			
T5.1	L2,L3	8	4	1.00	0.05714	-27.8		✓
T5.2	L2,L3	16	8	1.00	$6.22e-4$			
T5.3	L2,L4	8	8	1.00	$2.48e-5$			
T5.4	L2,L4	8	8	1.00	$2.48e-5$			
T5.5	L2,L3	64	16	1.00	$3.28e-14$			
T5.6	L2,L3	32	16	1.00	$2.66e-8$			
T5.7	L2,L3	16	9	1.00	$7.87e-4$			
T5.8	L2,L4	32	32	1.00	$3.80e-36$			
T5.9	L2,L3	8	4	1.00	0.05714	-27.8		✓

(continued on next page)

(continued)

Task	Lemma	d	K	F^*	$\varepsilon_{\text{theo}}$	$\log_{10}\varepsilon_{\text{s}}$	m	Esc.
T5.10	L2, L3	32	8	1.00	$7.61e-7$			
T5.11	L2, L4	8	8	1.00	$2.48e-5$			
T5.12	L2, L3	32	16	1.00	$2.66e-8$			
T5.13	L2, L3	256	16	1.00	$1.59e-24$			
T5.14	L2, L3	64	16	1.00	$3.28e-14$			
T6.1	L2, L3	8	2	1.00	0.07143	-16.1		✓
T6.2	L2, L3	32	4	0.9900	$1.11e-4$			
T6.3	L5	8		0.9900	0.01578	-2.8		✓
T6.4	L1	512		0.9990	$1.00e-1533$			
T6.5	L1	128		0.9990	$1.00e-381$			
T6.6	L6	32		0.9900	0.0625	-28.3		✓
T6.8	L8	8		0.9900	0.06313	-22.5		✓
T6.9	L2, L3	32	5	1.00	$2.48e-5$			
T6.10	L2, L3	32	5	1.00	$2.48e-5$			
T6.11	L2, L3	32	2	0.9900	0.004032			
T6.12	L2, L3	64	4	1.00	$6.30e-6$			
T6.13	L2, L3	16	8	1.00	$6.22e-4$			
T6.14	L1	2		0.9990	0.001		1	
T6.15	L3, L7	128	16	0.9000	$1.75e-101$			
T6.16	L2, L3	16	2	1.00	0.0625	-35.3		✓
T6.17	L5	128		0.9990	$6.11e-5$	-8.1		
T6.18	L2, L3	8	4	0.9500	0.05714	-27.8		✓
T6.19	L2, L3	32	2	1.00	0.004032	-77.2		

C Formal Proof Bundle

This appendix states and proves the nine Lemma instruments L1 to L9 of App. B and the three task specific Custom Bounds (T2.10, T2.17, T2.18). L7 differs from the other eight: it is a Monte Carlo estimator with a standard error, not a bound, and the remark that follows it gives closed forms for the three rows it covers. Two preparatory sub sections (*Notation*, *L5 Correction Justification*) come first. Each Lemma then has one sub section with its statement, its proof, the verification types it applies to, and a numeric instance with parameters from the task registry (App. A); L9 is built from four parts, L9a to L9d. The three Custom Bound sub sections close the appendix. A deterministic reproduction script released alongside the benchmark evaluates the formula column of the released mapping for all 86 tasks and finds a maximum relative deviation of 2.65×10^{-6} from the values stored in the released mapping; Tab. 6 prints them rounded. For T6.4 and T6.5 that comparison is vacuous: the closed form $(1-F^\star)^{d-1} = (10^{-3})^{d-1}$ underflows double precision, so the stored and the recomputed value are both zero and the script records agreement without testing anything. The exact entries 10^{-1533} and 10^{-381} printed in Tab. 6 follow from that closed form evaluated in exact arithmetic, not from the reproduction script. No $\varepsilon_{\text{theo}}$ entry of Tab. 6 is a Monte Carlo estimate: the three L7 rows T2.12, T2.22 and T6.15 carry the closed form values of the remark in §C.9. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below or an inequality sign is rounded in its safe direction instead, and an equality sign joins only exact values.

C.1 Notation

The following symbols are used uniformly throughout this bundle and across the cross referenced QuARC artifacts.

- d – full Hilbert-space dimension. For a task with n_{qubits} qubits, $d = 2^{n_{\text{qubits}}}$.
- d_{code} – code-space dimension. Whenever the test acts on a stabilizer code subspace $\mathcal{C} \subseteq \mathbb{C}^d$, $d_{\text{code}} = \dim \mathcal{C} \leq d$.

- $F^\star$ – threshold fidelity (acceptance threshold of a fidelity test). The per task value that enters the bounds is the F^* column of Tab. 6.
- K – the number of conditions the shipped specification of a task prescribes: its test cases, expected peaks, or listed outputs. It is the value recorded in the released registry (Tab. 5); where the registry records none, K in this appendix denotes the corresponding count of the task's specification, such as the parameter cases of T4.6 or the stabilizer generators of T6.6. K_{c} – for a functional `L2,L3` task with more than one test case, the number of argmax or distribution conditions of its weakest single case. The tabulated single case bound of Tab. 6 is evaluated at K_{c} on ten rows, T2.1, T2.2, T2.3, T2.9, T2.11, T2.19, T2.21, T2.24, T2.26 and T6.16, and at K on the other 19 multi case `L2,L3` rows. On 18 of those 19 the cases prescribe pairwise distinct output indices (for T6.18 through distribution targets of mass at least 0.95, whose event lies inside the argmax event), so the tabulated value is at least the bound of Eq. (13) and therefore bounds the joint event; the cases of T6.10 prescribe ancilla bands and no output index, and its tabulated value is a conservative placeholder (App. D).
- n_{stab} – number of independent stabilizer generators of the target code space (only defined for stabilizer eigenvalue tests).
- μ_{Haar} – the unique left-/right-invariant probability measure on $U(d)$, and (by abuse of notation) the induced uniform measure on $\mathbb{CP}^{d-1}$.
- $|\psi\rangle = \sum_{i=0}^{d-1} \psi_i |i\rangle \in \mathbb{C}^d$ – a pure state with computational-basis amplitudes ψ_i.
- $V \in U(d)$ – the target unitary of a fidelity test, and $U \sim \mu_{\mathrm{Haar}}$ – a Haar random adversarial unitary.
- $\|p - p_{\mathrm{target}}\|_{\mathrm{TV}} = \frac{1}{2}\sum_i |p_i - p_{\mathrm{target},i}|$ – total-variation distance between probability vectors.
- n_{input} – number of input register qubits in a Simon's algorithm style test, where the marginal output space on the input register has dimension $2^{n_{\mathrm{input}}}$.

Ancilla extension symmetry. The invariance that the introduction lists alongside qubit relabeling is defined here, since the Lemma bounds are stated on the data register alone. A candidate circuit may act on $n_{\mathrm{qubits}} + m$ qubits with $m \geq 0$, that is it may carry m ancillas beyond the n_{qubits} the task specifies. Let L be the injective map from the task's data positions into the candidate's wires that the runner recovers from the candidate itself, from a layout dictionary in its code, an output permutation comment, or the transpiler layout of the circuit; on a task that fixes its layout the runner discards any such map. With a recovered map, the evaluated state is the ancilla conditioned marginal

$$|\psi_L\rangle \;=\; \frac{\left(I_L \otimes \langle 0|^{\otimes m}\right)|\psi\rangle}{\left\|\left(I_L \otimes \langle 0|^{\otimes m}\right)|\psi\rangle\right\|}, \tag{2}$$

and the state fidelity and parametric state fidelity tests apply the acceptance region $A(T) \subseteq \mathbb{C}^d$ to $|\psi_L\rangle$ rather than to $|\psi\rangle$ when the runner has recovered a map. For these two types, candidates that differ only by unentangled ancillas returned to $|0\rangle$ therefore receive the same verdict, which is the symmetry the benchmark claims. Without a recovered map the state fidelity test reduces a wider candidate of at most twelve qubits by a partial trace over its highest indexed qubits and fails it if the reduced state is not pure; on a task that fixes its layout, or beyond twelve qubits, a wider candidate fails. The unitary fidelity test, given a recovered map and a candidate of at most ten qubits, evaluates the block of the candidate unitary with every ancilla prepared and read out in $|0\rangle$, without renormalization; otherwise a wider candidate fails. Other verification types are not described here. The definition is a projection followed by a renormalization, not a partial trace: it is exact only when the ancillas are disentangled from the data register at the end of the circuit, and a candidate that leaves ancilla entanglement behind is scored on the postselected branch. The Lemma bounds of this appendix are stated for the d dimensional data register and a candidate on exactly the task's qubits; the postselection step itself lies outside these bounds and is not priced in this appendix.

C.2 L5 Correction Justification

A bound of the form

$$\Pr_{U \sim \mu_{\mathrm{Haar}}}\left[|\mathrm{tr}(U^\dagger V)/d|^2 \geq F^\star\right] \leq (1 - F^\star)^{d^2-1}, \tag{3}$$

justified by analogy with L1 through the embedding $U(d) \hookrightarrow \mathbb{C}^{d^2}$, is not used in this appendix, for two reasons:

1. $U(d)$ has real dimension d^2, while the Euclidean sphere S^{2d^2-1} in $\mathbb{C}^{d^2} \cong \mathbb{R}^{2d^2}$ has real dimension $2d^2-1$. Every unitary has Frobenius norm $\sqrt{d}$, so the sphere contains the rescaled group $d^{-1/2}U(d)$, and this is a measure-zero submanifold of S^{2d^2-1}. Consequently, the uniform surface measure on S^{2d^2-1} does *not* restrict to the Haar measure on $U(d)$. The L1-style Fubini–Study volume integral cannot be transported to $U(d)$ without further work.
2. A clean derivation of a volume-style bound on $U(d)$ would require concentration of measure on $U(d)$ (Meckes, 2019). Such a bound could be tighter than the Markov bound, but its derivation is outside the scope of QuARC. In the relevant regime ($n_{\text{qubits}} \leq 10$, $d \leq 1024$) the Markov bound already satisfies $1/(d^2 F^\star) \leq 0.01$ for $d \geq 11$ (with $F^\star = 0.99$).

The bound used in QuARC for unitary-fidelity tests is therefore the Markov bound $1/(d^2 F^\star)$ stated and proved in Lemma L5 below, together with its higher moment form, and its code-subspace analogue $1/(d_{\text{code}}\, d\, F^\star)$ (Haar adversary on the full $U(d)$, projected onto the code subspace) stated as Lemma L8.

C.3 Lemma L1 (Fubini–Study Volume)

For the Haar measure on pure states of $\mathbb{C}^d$, equivalently the Fubini–Study volume on $\mathbb{CP}^{d-1}$, and any fixed target state $|\psi_T\rangle \in \mathbb{CP}^{d-1}$, the volume of the fidelity ball at threshold $F^\star$ admits the closed form

$$\text{Vol}_{\mu_{\text{Haar}}}\big(\{|\psi\rangle \in \mathbb{CP}^{d-1} : |\langle\psi|\psi_T\rangle|^2 \geq F^\star\}\big) \;=\; (1-F^\star)^{d-1}. \tag{4}$$

Proof. Throughout, $d \geq 2$ and $0 \leq F^\star \leq 1$, as the statement requires: the former keeps $\mathbb{CP}^{d-1}$ nondegenerate and the $\text{Beta}(1, d-1)$ density used below well defined, the latter places $F^\star$ inside the range $[0,1]$ of $|\langle\psi|\psi_T\rangle|^2$. Write μ_{Haar} for the uniform probability measure on the pure states of $\mathbb{C}^d$. Because it is a probability measure, the volume in (4) equals the probability that a Haar random $|\psi\rangle$ lies in the fidelity ball,

$$\text{Vol}_{\mu_{\text{Haar}}}\big(\{|\psi\rangle : |\langle\psi|\psi_T\rangle|^2 \geq F^\star\}\big) \;=\; \Pr\big[\,|\langle\psi|\psi_T\rangle|^2 \geq F^\star\,\big].$$

Extend $|\psi_T\rangle$ to an orthonormal basis of $\mathbb{C}^d$ that has $|\psi_T\rangle$ as its first vector, which is possible because any unit vector extends to an orthonormal basis, and read this basis as the image of the computational basis under a fixed unitary. Because μ_{Haar} on pure states is invariant under every fixed unitary (Zyczkowski & Sommers, 2001), the law of $|\langle\psi|\psi_T\rangle|^2$ does not depend on which unit vector $|\psi_T\rangle$ is, so it coincides with the law of a single computational basis probability p_0 of a Haar random state. By §C.5 (Eq. (9)), $p_0 \sim \text{Beta}(1, d-1)$, with density $(d-1)(1-u)^{d-2}$ on $u \in [0,1]$. A tail probability is the integral of the density over the tail, so

$$\Pr\big[\,|\langle\psi|\psi_T\rangle|^2 \geq F^\star\,\big] \;=\; \int_{F^\star}^{1} (d-1)(1-u)^{d-2}\, du \;=\; (1-F^\star)^{d-1}, \tag{5}$$

where the second equality is the substitution $w = 1-u$, under which the integral becomes $\int_0^{1-F^\star} (d-1)\, w^{d-2}\, dw = (1-F^\star)^{d-1}$. Together with the identification of volume and probability above, this is exactly (4).

The bound applies to the state fidelity and parametric state fidelity verification types. As an instance, task T1.10 is a state fidelity check on $d = 4$ at threshold $F^\star = 0.99$, so

$$\varepsilon_{\text{T1.10}} \;=\; (1-F^\star)^{d-1} \;=\; (0.01)^3 \;=\; 10^{-6}, \tag{6}$$

well below the 0.01 escalation threshold, and T1.10 is not an escalation candidate.

Remark (parametric state fidelity: per case rate and relabel retry). For the parametric state fidelity tasks the published $\varepsilon_T = (1-F^\star)^{d-1}$ is the bound of Eq. (4) for one parameter case against its declared target, and two effects of the deployed runner separate it from the rate at which the runner accepts. First, the runner does not compare against the target alone: for each parameter case it

maximizes the fidelity over the qubit relabelings admitted by the layout search, so the accepting set for one case is a union of fidelity caps, one per distinct relabeled target, rather than the single cap of Eq. (4). For $F^\star \geq 1/2$, two such caps are disjoint whenever the fidelity between their centers stays below $(2F^\star - 1)^2$, because the Fubini–Study angle obeys the triangle inequality, and disjoint caps add. Of the ten single case state fidelity targets three are not invariant under relabeling, T1.10 being one of the seven that are. A moving target makes the printed bound too small only where the permutation retry runs, which leaves one of the three, T1.17, with an orbit of 60: T6.4 and T6.5 sit above the retry's ceiling of $n \leq 5$. That one bound stays far below the threshold even so, at 6×10^{-61}. For T4.15 ($d = 4$, $F^\star = 0.99$, two distinct relabeled targets per case, center fidelity well below $(2F^\star - 1)^2 = 0.9604$) the per case rate is exactly $2(1 - F^\star)^{d-1} = 2 \times 10^{-6}$, twice the published number. In general the per case rate lies between $(1 - F^\star)^{d-1}$ and $m(1 - F^\star)^{d-1}$, with m the number of distinct relabeled targets of the case, which reaches 24 in this family. Second, and in the opposite direction, the runner requires *all* K parameter cases to pass. Each case binds its own parameter values and is evaluated on its own output state, so the conjunction is not empty (the canonical solution passes every case); it is contained in each single case, and under a null model in which each case is Haar distributed its probability is at most the per case rate of any one case. The published value is therefore a bound for one case at the declared target, and the deployed per task rate is at most m times it, with m taken for any one of the task's cases; the empirical calibration of App. D measures a single case without the relabel union and so estimates neither quantity.

C.4 Lemma L2 (argmax Basis)

For a Haar-random pure state $|\psi\rangle = \sum_{i=0}^{d-1} \psi_i |i\rangle$ on $\mathbb{C}^d$, the probability that the argmax of the computational-basis probabilities lands on any fixed index $j \in \{0, \ldots, d-1\}$ is uniform,

$$\Pr_{|\psi\rangle \sim \mu_{\text{Haar}}} \Big[\arg\max_i |\psi_i|^2 = j \Big] = \frac{1}{d}. \tag{7}$$

Proof. The argument follows Mezzadri (2006). The Haar measure μ_{Haar} on pure states is invariant under the basis permutation unitaries $\{U_\sigma : \sigma \in S_d\}$ with $U_\sigma |i\rangle = |\sigma(i)\rangle$. For any two basis indices j and j', the transposition $\sigma = (j\, j')$ maps the event $\{|\psi\rangle : \arg\max_i |\psi_i|^2 = j\}$ bijectively onto $\{|\psi\rangle : \arg\max_i |\psi_i|^2 = j'\}$ and preserves μ_{Haar}. The events $\{\arg\max = j\}$ for $j = 0, \ldots, d-1$ therefore all carry the same probability. Up to a measure-zero tie set they partition the unit sphere, so each event has probability $1/d$.

The bound applies to the functional verification type in the C1 ($K = d$) regime, and supplies the per position equiprobability ingredient that feeds L4 (joint argmax) and the symmetry input that feeds L6 (multi output argmax). As an instance, task T5.8 is a functional test on $d = 32$ with $K = 32$ output classes, so the per position random guess probability is

$$\Pr = 1/d = 1/32 = 3.125 \times 10^{-2}, \tag{8}$$

in agreement with the uniform predictor on a 32 dimensional basis. The full task bound for T5.8 uses L4, and L2 supplies the building block.

C.5 Lemma L3 (Dirichlet Marginal)

For a Haar-random pure state $|\psi\rangle \sim \mu_{\text{Haar}}$ on $\mathbb{C}^d$, the computational-basis probabilities $p_i = |\psi_i|^2$ form a Dirichlet vector $(p_0, \ldots, p_{d-1}) \sim \text{Dirichlet}(1, \ldots, 1)$, and the marginal of any single component is

$$p_i \sim \text{Beta}(1,\, d-1). \tag{9}$$

Proof. The argument follows Zyczkowski & Sommers (2001). Write $\psi_i = (a_i + i b_i)/\|(a, b)\|$ with $a_i, b_i \sim \mathcal{N}(0, 1)$ i.i.d. The resulting pure state is Haar distributed because the standard complex Gaussian on $\mathbb{C}^d$ is unitarily invariant. Then

$$|\psi_i|^2 = \frac{a_i^2 + b_i^2}{\sum_j (a_j^2 + b_j^2)} \tag{10}$$

is the ratio of a χ^2_2 distributed numerator (an exponential law with mean 2) to the sum of d such i.i.d. terms. Only ratios enter, so the scale is irrelevant and the vector has the law of d i.i.d. $\mathrm{Exp}(1)$ variables divided by their sum. This is exactly the Dirichlet$(1,\dots,1)$ representation, uniform on the probability simplex Δ^{d-1}. Marginalizing out $d-1$ components of a Dirichlet$(1,\dots,1)$ leaves Beta$(1,\ d-1)$, which is the stated claim.

The bound applies to the functional verification type in the C2 ($K < d$) regime via the tabulated $\varepsilon = K\binom{d}{K}^{-1}$ (combined with L2) and to the probability verification type via the Beta$(1, d-1)$ marginal, whose survival function $\Pr[p_i \geq F^\star] = (1-F^\star)^{d-1}$ feeds the per component check. As an instance, task T2.6 is a single component probability test on $d = 4$ checking the marked basis state $|11\rangle$ at threshold $F^\star = 0.99$, so

$$\varepsilon_{\text{T2.6}} = (1-F^\star)^{d-1} = (0.01)^3 = 10^{-6}, \tag{11}$$

well below the 0.01 escalation threshold, and T2.6 is not an escalation candidate.

Corollary (the C2 bound, derived, and the factor K by which the tabulated form overstates it). The C2 form quoted above is carried by the mapping but is nowhere derived; the derivation is short, and it sharpens the constant. By L3 the basis probabilities $(p_0,\dots,p_{d-1})$ are Dirichlet$(1,\dots,1)$ and therefore exchangeable: their joint law is invariant under every permutation of the d components. The prescribed top K index set is a function of that vector alone, so exchangeability carries over to it and each of the $\binom{d}{K}$ subsets of size K is equally likely to be the realized one, up to the measure zero tie set. A fixed prescribed subset is hit with probability

$$\varepsilon_{\text{C2}} \;=\; \binom{d}{K}^{-1}, \tag{12}$$

with no prefactor, and L2 is the case $K = 1$ of the same statement. The tabulated $K\binom{d}{K}^{-1}$ is thus conservative by exactly the factor at which it is evaluated (K_{c} on the ten rows named in §C.1, K on the others) and agrees with the exact value only where that factor is 1. Sharpening it costs nothing: across the 29 rows of this family the same 14 escalate under either form, no row changes side, and the escalation set stays at 34 of the 86 tasks. The 4 rows whose factor exceeds 1 and that carry a current Monte Carlo run confirm the exact form. That run is the released Dirichlet sample `lemma_L7_mc.csv`, which draws $(p_0,\dots,p_{d-1})$ and asks whether the realized top K set is the prescribed one, so it prices the model event of the display above and not the predicate the deployed test accepts. On the model event the exact value lies inside the 95% interval in each case: at $d = 8$, $K = 4$ (T5.1, T5.9 and T6.18) the sample brackets $1/70$ and excludes the tabulated $4/70$, and T6.1 at $d = 8$, $K = 2$ behaves the same way. We record the tightening here and leave the tabulated column as it stands.

Remark (the relabel orbit of the `L2, L3` *rows).* Where the runner retries a failed candidate on every qubit relabeled image (§3.3), it relabels the prepared input index and the prescribed output index by the same bit permutation, so the argmax event it effectively tests is a union of copies of the tabulated one. The retry does not run on a task that declares `fixed_qubit_layout`, and six `L2, L3` rows declare it, T2.1, T2.3, T2.11, T2.19, T2.24 and T6.16; for them the factor is 1. As a model computation, for a single case with one prescribed output index of Hamming weight w on n qubits, a prepared input that every relabeling fixes (such as the all zero input), and no fixed layout the copies are disjoint and there are $\binom{n}{w}$ of them; on the multi case rows the orbit acts on the whole tuple and the factor is at most $n!$. The multi case row where this matters is T6.19, on $n = 5$ qubits with $d = 32$ and $K = 2$, without a fixed layout. The prescribed output 24 of its first case has $\binom{5}{2} = 10$ relabeled images, the prepared input 1 of its second case has 5, and only the 3 disjoint from a given image of 24 can accompany it, so the 120 permutations yield $m = 30$ distinct conditions; unlike the single case model above, these copies overlap. Eq. (13) caps each of them at $(d-K)!/d! = 1/992$, so their union is at most $30/992 < 0.031$, which is above the 0.01 escalation threshold. T6.19 is the only row of this family that lies below the 10^{-2} threshold on its tabulated value and above it on the widened union; elsewhere a larger K or a smaller orbit keeps the union below it. For this reason the chain bound of §C.11 is evaluated for this row too, even though it does not escalate. It prices the conjunction of both cases under the tolerances of the shipped runner, including the total variation condition of the comparison against the canonical output, and the union over all 120 relabelings through its factor N; the result stays below 10^{-77} in Tab. 6. The escalation set is untouched, and the

tabulated $\varepsilon_{\text{theo}} = 2\binom{32}{2}^{-1} \approx 4.03 \times 10^{-3}$ bounds the model event of one prescribed argmax tuple, not the widened one.

C.6 LEMMA L4 (JOINT BASIS)

For a Haar distributed unitary U on $U(d)$ and $K \le d$ arbitrary input vectors $|x_1\rangle, \ldots, |x_K\rangle$ fed one at a time to U, consider a basis position test that prescribes for the kth input a target index j_k, that is a required value of the position $\arg\max_i |\langle i|U|x_k\rangle|^2$ of the largest computational basis probability. Assume the prescribed indices $j_1, \ldots, j_K$ are pairwise distinct. Then the joint probability that every input attains its prescribed position satisfies

$$\Pr[\text{all } K \text{ prescribed positions attained}] \le \frac{(d-K)!}{d!}. \tag{13}$$

In particular, for $K = d$ this is $1/d!$. The distinctness of the j_k is a hypothesis on the specification, not a property of any test family; rows that prescribe a repeated output position (T2.1, T2.2, T2.3, T2.19) do not meet it and are priced by the chain bound of §C.11 instead. Equality holds exactly when the realized argmax positions are almost surely pairwise distinct: for $K = 1$ at every d, where the condition is vacuous, and for $d = 2$, $K = 2$ with orthogonal inputs, where it gives $1/2$. For $d \ge 3$ with $K \ge 2$ and orthonormal inputs the bound is strict.

Proof. Set

$$A_k \;=\; \arg\max_i \big|\langle i|U|x_k\rangle\big|^2, \qquad k = 1, \ldots, K, \tag{14}$$

for the realized position of the kth output, defined up to the measure zero tie set. For a basis permutation $\sigma \in S_d$ with permutation unitary P_σ, the map $U \mapsto P_\sigma U$ preserves the Haar measure by its left invariance (Mezzadri, 2006), as used in §C.4, and relabels the output basis of every input at once, so the vector $(A_1, \ldots, A_K)$ has a law invariant under the diagonal action $\sigma \cdot (a_1, \ldots, a_K) = (\sigma(a_1), \ldots, \sigma(a_K))$ of S_d. That action is transitive on the ordered K tuples with pairwise distinct entries, of which there are $d!/(d-K)!$, so all such tuples carry equal probability. Writing D for the event that $A_1, \ldots, A_K$ are pairwise distinct, these distinct tuples partition D into $d!/(d-K)!$ equiprobable outcomes, each of probability $\Pr[D] \cdot (d-K)!/d!$. Since $(j_1, \ldots, j_K)$ has distinct entries by hypothesis it is one of them, so

$$\Pr\big[(A_1, \ldots, A_K) = (j_1, \ldots, j_K)\big] \;=\; \Pr[D] \cdot \frac{(d-K)!}{d!} \;\le\; \frac{(d-K)!}{d!}, \tag{15}$$

the equality by equiprobability and the inequality because $\Pr[D] \le 1$. This is (13). Equality holds precisely when $\Pr[D] = 1$. For $K = 1$ a single index is vacuously distinct, so D is sure at every d and the value $1/d$ is attained. For $d = 2$, $K = 2$ with orthogonal inputs the two outputs are orthogonal columns of U, so their basis probability vectors are $(|a|^2, |b|^2)$ and $(|b|^2, |a|^2)$ and their argmax positions never coincide, giving $\Pr[D] = 1$, so the joint probability equals $1/2 = (d-K)!/d!$. For $d \ge 3$ with $K \ge 2$ and orthonormal inputs two orthogonal columns share an argmax position with positive probability, so $\Pr[D] < 1$ and the bound is strict.

The step this proof avoids is worth naming. It is *not* legitimate to condition on the event $\{A_1 = j_1\}$ and then treat the second output as a fresh Haar state on the orthogonal complement of $|j_1\rangle$: conditioning on a fixed first *column* does leave a Haar state on the complement, but conditioning on the *event* does not fix the column, and for $d \ge 3$ the conditional probability $\Pr[A_2 = j_1 \mid A_1 = j_1]$ is strictly positive. The symmetry argument above needs no such conditioning.

The bound applies to the functional verification type, in the C1 ($K = d$, bound $1/d!$) regime, and in the C3 (distribution) regime only where every prescribed distribution keeps more than half of its mass on one index after the tolerance, since only then is the acceptance event contained in the argmax conjunction. As an instance, task T5.8 is a functional test on $d = 32$ with $K = 32$, so

$$\varepsilon_{\text{T5.8}} \le \frac{1}{d!} = \frac{1}{32!} < 3.81 \times 10^{-36}, \tag{16}$$

strictly, since $d \ge 3$ and $K \ge 2$. T5.8 does not fix its qubit layout and has $n = 5$ qubits, so the runner also tries the 120 qubit relabelings; the union over them is at most $120/32! \approx 4.6 \times 10^{-34}$, still far below the 0.01 escalation threshold.

C.7 Lemma L5 (Haar Unitary Fidelity, Markov Bound)

For any fixed $V \in U(d)$ and a Haar-random unitary $U \sim \mu_{\text{Haar}}$ on $U(d)$, the probability that the normalized trace fidelity exceeds threshold $F^\star$ admits the Markov bound

$$\Pr_{U\sim\mu_{\text{Haar}}}\left[|\mathrm{tr}(U^\dagger V)/d|^2 \geq F^\star\right] \leq \frac{1}{d^2 F^\star}. \tag{17}$$

Proof. The argument follows Collins & Śniady (2006). The Weingarten formula for second moments on $U(d)$ gives

$$\mathbb{E}_{U\sim\mu_{\text{Haar}}}\left[U_{ij}\,\overline{U_{kl}}\right] = \frac{1}{d}\,\delta_{ik}\delta_{jl}. \tag{18}$$

For any fixed $V \in U(d)$, this implies

$$\begin{aligned}\mathbb{E}\left[|\mathrm{tr}(U^\dagger V)|^2\right] &= \sum_{i,j,k,l} V_{ij}\,\overline{V_{kl}}\,\mathbb{E}\left[\overline{U_{ij}}\,U_{kl}\right] \\ &= \frac{1}{d}\sum_{i,j}|V_{ij}|^2 = \frac{\mathrm{tr}(V^\dagger V)}{d} = 1,\end{aligned} \tag{19}$$

using $\mathrm{tr}(V^\dagger V) = d$ for $V \in U(d)$. Hence $\mathbb{E}[|\mathrm{tr}(U^\dagger V)/d|^2] = 1/d^2$. Markov's inequality applied to the non negative random variable $X = |\mathrm{tr}(U^\dagger V)/d|^2$ with threshold $F^\star$ then yields

$$\Pr[X \geq F^\star] \leq \frac{\mathbb{E}[X]}{F^\star} = \frac{1}{d^2 F^\star}. \tag{20}$$

The bound applies to the unitary fidelity and parametric unitary fidelity verification types. The code subspace variant for code space unitary fidelity is given separately as Lemma L8 with the same Weingarten plus Markov proof scheme, applied to a Haar unitary on the full $U(d)$ projected onto the code subspace. As an instance, task T2.16 is a unitary fidelity check on $d = 16$ at threshold $F^\star = 0.99$, so

$$\varepsilon_{\text{T2.16}} = \frac{1}{d^2 F^\star} = \frac{1}{256 \cdot 0.99} \approx 3.9457 \times 10^{-3}, \tag{21}$$

below the 0.01 escalation threshold, and T2.16 is not an escalation candidate. By contrast, the smaller dimension task T1.11 with $d = 4$ and $F^\star = 0.99$ yields $1/(16 \cdot 0.99) \approx 6.31 \times 10^{-2} > 0.01$, and is listed as an escalation candidate.

Remark (the relabel orbit of the target). The instance above states the bound for a single fixed target V, whereas the runner accepts a candidate whose trace fidelity clears $F^\star$ against *any* qubit relabeled image $PVP^\dagger$ of the target. The event it tests for a candidate on exactly the task's n qubits is therefore a union over the orbit of V under qubit permutation conjugation, and Eq. (17) bounds only one of its members. The orbit size is a property of the target. For T3.13 the target is the bit reversal permutation, whose orbit has three distinct members; Markov applied to the maximum over the orbit gives 0.00712 (Monte Carlo, $N = 2 \times 10^6$, standard error 3×10^{-6}), under the 0.01 escalation threshold. For T2.16 the target is the four qubit QFT, whose orbit has the full 24 members; the same estimate gives 0.01372 (same protocol, standard error 4×10^{-6}), which is *above* 0.01: Markov's inequality applied to the orbit maximum does not keep T2.16 below the threshold. The higher moments below close this in closed form, and the two Monte Carlo figures are not used for any tabulated value.

The moment route. Eq. (17) applies Markov's inequality to the second moment. Higher moments are available in closed form: the traces of a Haar unitary satisfy

$$\mathbb{E}\left|\mathrm{tr}\,U\right|^{2k} = k! \qquad \text{for } 1 \leq k \leq d, \tag{22}$$

an identity from the theory of random unitary matrices that we state without proof. For fixed V the product $U^\dagger V$ is again Haar distributed, so Eq. (22) holds for $\mathrm{tr}(U^\dagger V)$ as well. Writing $X = |\,\mathrm{tr}(U^\dagger V)|^2$ and applying Markov to X^k rather than to X,

$$\Pr\left[X \geq d^2 F^\star\right] = \Pr\left[X^k \geq (d^2 F^\star)^k\right] \leq \frac{k!}{(d^2 F^\star)^k}. \tag{23}$$

The case $k = 1$ is Eq. (17). The identity is stated for $k \le d$ only, every order used below satisfies $k \le d$, and a task with a small dimension gains little from a larger order. For $d = 2$ and $F^\star = 0.99$ the order $k = 2$ still yields 0.128, an order of magnitude above the escalation threshold; the bound sharpens where the dimension is large enough to pay for it, and nowhere else.

The orbit, priced. The union over the m members of the relabeling orbit costs a factor m, so the union the harness tests for a candidate on exactly the task's n qubits is bounded by $m\,k!/(d^2F^\star)^k$ for every $k \le d$. Every order gives a valid bound. The released values take $k = 2$ on the rows with $m = 1$, and on the rows with $m > 1$ the smallest order at which the product stays below the escalation threshold; T6.17 ($m = 1$) closes at $k = 1$ already and carries its $k = 2$ value. The orbit size is computed per task from the declared target, with a global phase equalisation and with the harness convention respected: a task that fixes its qubit layout is tested at the identity alone and has $m = 1$ whatever the symmetry of its target, and above the permutation cap of five qubits the harness does not search at all. Of the 20 L5 rows, 11 have $m > 1$.

For T3.13 ($m = 3$) and T2.16 ($m = 24$) the order $k = 2$ already suffices, giving 9.34×10^{-5} and 7.47×10^{-4}. Three rows need one order more: T3.2, T3.6 and T4.6 carry $m = 2$ and close at $k = 3$ with 3.02×10^{-3}. Every escalating L5 row closes under the union, and the $\log_{10} \varepsilon_s$ column of Tab. 6 reports the union and not the single member. The union bounds an event of the Haar null model for a candidate on exactly the task's n qubits; a wider candidate is scored by a different path of the runner, which this bound does not cover.

Remark (parametric unitary fidelity: the conjunction is not measured). For the parametric unitary fidelity tasks the runner requires all K parameter cases to clear $F^\star$ simultaneously, each case again maximized over the admitted qubit relabelings, whereas both Eq. (17) and the calibration sampler of App. D price a single case. Ignoring the conjunction is conservative; ignoring the maximization over relabelings is not, and the union bound below accounts for it. For T4.6 ($d = 4$, $K = 3$, $F^\star = 0.99$) the published bound is $1/(d^2F^\star) \approx 6.31 \times 10^{-2}$, while the single case calibration returns zero hits at $N = 10^9$; a union bound over the two admitted relabelings applied to the Wilson 95% upper bound of that single case run puts the deployed conjunction event at or below 7.683×10^{-9}. That figure is an empirical confidence bound, not a closed form one, and it is quoted here only to show the size of the reserve. The closed form bound for T4.6 is the order $k = 3$ union of Eq. (23), 3.02×10^{-3}; the conjunction over its K parameter cases is still not priced, and that remains conservative.

C.8 Lemma L6 (Multi-output argmax)

For a Haar random pure state $|\psi\rangle \sim \mu_{\text{Haar}}$ on a joint Hilbert space of dimension d carrying K disjoint register subsystems of local dimensions $d_1, \dots, d_K$, let A_k be the argmax position of the computational basis marginal on register k. For every prescribed tuple $(j_1, \dots, j_K)$ of argmax positions the joint probability obeys the product bound

$$\Pr\big[A_k = j_k \text{ for all } k\big] \;\le\; \prod_{k=1}^{K} d_k^{-1}, \tag{24}$$

which the proof sharpens to an equality. When every register is binary, so $d_k = 2$ for all k, the right side reads 2^{-K}.

Proof. Write the joint space as $\mathbb{C}^d = \big(\bigotimes_{k=1}^{K} \mathbb{C}^{d_k}\big) \otimes \mathbb{C}^{d_0}$, where $\mathbb{C}^{d_0}$ collects the registers carrying no constraint ($d_0 = 1$ if there are none), and let

$$p_a^{(k)} \;=\; \sum_{\text{all other register labels}} \big|\psi_{\dots\, a\, \dots}\big|^2, \qquad a \in \{0, \dots, d_k - 1\}, \tag{25}$$

be the computational basis marginal on register k, with $A_k = \arg\max_a p_a^{(k)}$ its argmax position, defined up to the measure zero tie set. Consider the product group $G = S_{d_1} \times \cdots \times S_{d_K}$ acting on $\mathbb{C}^d$ by the local basis permutation unitaries $P_\tau = P_{\tau_1} \otimes \cdots \otimes P_{\tau_K} \otimes I_{d_0}$. Each P_τ is unitary, so it preserves μ_{Haar} on pure states, and it acts on the marginals of (25) by relabeling register k with τ_k; the argmax position moves with that relabeling, hence

$$(A_1, \dots, A_K) \;\longmapsto\; \big(\tau_1(A_1), \dots, \tau_K(A_K)\big). \tag{26}$$

Because P_τ preserves the state distribution and transforms the argmax vector by (26), the law of $(A_1, \ldots, A_K)$ is invariant under G. The action (26) is transitive on the full product set $\prod_{k=1}^{K} \{0, \ldots, d_k - 1\}$, since each factor S_{d_k} is transitive on its own register and the factors act independently. A G invariant law on a set on which G acts transitively is the uniform law, so

$$\Pr\big[(A_1, \ldots, A_K) = (j_1, \ldots, j_K)\big] \;=\; \prod_{k=1}^{K} d_k^{-1} \tag{27}$$

for every tuple $(j_1, \ldots, j_K)$ of prescribed argmax positions, with equality rather than inequality. Eq. (27) evaluates the event of (24) exactly, so the displayed bound holds and is tight up to the measure zero tie set. In the binary case $d_k = 2$ the product reads 2^{-K}.

Corollary (stabilizer sign pattern). Let $g_1, \ldots, g_K$ be independent commuting Pauli operators on n qubits, with $K = n_{\text{stab}}$, and let $|\psi\rangle \sim \mu_{\text{Haar}}$ be a Haar random pure state on $\mathbb{C}^{2^n}$. Then for every sign pattern $s = (s_1, \ldots, s_K) \in \{+1, -1\}^K$,

$$\Pr\big[\operatorname{sign}\langle\psi|g_k|\psi\rangle = s_k \text{ for all } k\big] \;=\; 2^{-K}, \tag{28}$$

and in particular $\Pr\big[\langle\psi|g_k|\psi\rangle > 0 \text{ for all } k\big] = 2^{-n_{\text{stab}}}$.

Proof. The proof mirrors the previous one, with Pauli conjugations in place of register permutations. For a Hermitian Pauli string E on the n qubits, a tensor product of I, X, Y, Z, the operator E and each generator g_k either commute or anticommute, and $E^2 = I$, so

$$E\, g_k\, E \;=\; (-1)^{c_k(E)}\, g_k, \qquad c_k(E) \in \{0, 1\},$$

with $c_k(E) = 1$ exactly when E and g_k anticommute. In the symplectic representation of the Pauli group over $\mathrm{GF}(2)$ the bit $c_k(E)$ is the symplectic form of E with g_k, which is bilinear, so the map $E \mapsto \big(c_1(E), \ldots, c_K(E)\big) \in \mathrm{GF}(2)^K$ is linear (Nielsen & Chuang, 2010, Ch. 10). This map is surjective: the generators are independent, so their symplectic vectors are linearly independent, and the symplectic form is nondegenerate, so the K commutation functionals $E \mapsto c_k(E)$ are themselves linearly independent and their joint image is all of $\mathrm{GF}(2)^K$. Hence for any two sign patterns s, s' there is a Hermitian Pauli string E with $(-1)^{c_k(E)} = s_k s'_k$ for every k.

Fix such an E and send $|\psi\rangle$ to $E|\psi\rangle$. The operator E is unitary, so this map preserves μ_{Haar} on pure states, and, using $E^\dagger = E$,

$$\langle E\psi|\, g_k\, |E\psi\rangle \;=\; \langle\psi|\, E g_k E\, |\psi\rangle \;=\; (-1)^{c_k(E)}\, \langle\psi|\, g_k\, |\psi\rangle,$$

the last step by the conjugation relation above. The map therefore multiplies each expectation $\langle\psi|g_k|\psi\rangle$ by $(-1)^{c_k(E)}$ and, by the choice of E, carries the event of sign pattern s bijectively onto the event of sign pattern s'. A measure preserving bijection makes the two events equiprobable, and since s, s' were arbitrary, all 2^K sign pattern events share one probability.

The events meet only on the tie set where some $\langle\psi|g_k|\psi\rangle = 0$. Each g_k is a nontrivial Pauli string, hence traceless with eigenvalues ± 1, so $|\psi\rangle \mapsto \langle\psi|g_k|\psi\rangle$ is a real quadratic form that is not identically zero; its zero set is a proper real algebraic subset of the state space and carries μ_{Haar} measure zero, and the union over the K generators is still null. Off this null set the 2^K events partition the state space, and being equiprobable each carries probability 2^{-K}, which establishes (28). Because $K = n_{\text{stab}}$, the all positive pattern $s = (+1, \ldots, +1)$ has probability $2^{-n_{\text{stab}}}$.

This sign coincidence is the model event that the T6.6 row prices as $2^{-n_{\text{stab}}}$, at the four independent generators $XZZXI$, $IXZZX$, $XIXZZ$, $ZXIXZ$ of the $[\![5, 1, 3]\!]$ code, so $K = n_{\text{stab}} = 4$; the runner's acceptance event, a fidelity style test and not a bare sign coincidence, is taken up in the remark after the instance below.

Remark. Transitivity of the product group is what carries the proof of L6, and it is worth being explicit about why: pairwise independence of the K argmax events would *not* suffice. For $K \geq 3$, pairwise independent events can have an intersection strictly larger than the product of their probabilities, so a product bound cannot be inferred from pairwise statements alone. The transitivity argument bypasses the question by fixing the joint law of the whole tuple in one step, and the corollary on stabilizer

signs rests on the same kind of argument with the Pauli group in place of the local permutations. Regarding sources: the Mezzadri 2007 citation used elsewhere in this bundle stands for the left and right invariance of the Haar measure, whose specialization to the basis permutation unitaries is what §C.4 and §C.6 invoke; the register marginal of (25) is the standard reduced description of Nielsen & Chuang (2010, §2.4).

The bound applies to the stabilizer eigenvalue verification type through the corollary, where it gives $2^{-n_{\text{stab}}}$. As an instance, task T6.6 is a stabilizer eigenvalue check on the five qubit code $[\![5, 1, 3]\!]$ with $d = 32$ and $n_{\text{stab}} = 4$ at threshold $F^\star = 0.99$, so

$$\varepsilon_{\text{T6.6}} = 2^{-n_{\text{stab}}} = 2^{-4} = 0.0625, \tag{29}$$

above the 0.01 escalation threshold, and T6.6 is an escalation candidate. The bound $2^{-n_{\text{stab}}}$ is the probability that all n_{stab} expectation signs come out positive under the Haar measure and does not depend on the threshold $F^\star$; the test design uses an $F^\star = 0.99$ margin against noisy canonical solutions rather than against the Haar baseline. This asymmetry against L1, L5, L8 (which carry $F^\star$ in the bound) is structural: L1 bounds a fidelity ball volume, L5 and L8 bound a fidelity tail above the threshold, while L2, L4, L6 bound argmax or sign events whose acceptance criterion is a discrete coincidence, not a fidelity tail.

Remark (the deployed stabilizer test does not price the event of Eq. (29)). The statement just made, that the bound does not depend on $F^\star$, is correct for the event the lemma models, which is a coincidence of the n_{stab} expectation signs, and it is precisely what makes the tabulated figure the wrong price for the test the runner actually executes. The runner compares no signs. For each generator P it evaluates the expectation $\langle\psi|P|\psi\rangle$ and accepts that generator when $|\langle P\rangle - 1| < 1 - F^\star$, a continuous margin on the expectation value, and it accepts the task only when all n_{stab} generators clear the margin simultaneously. A non identity Pauli string on n qubits is a traceless involution on $\mathbb{C}^d$ with $d = 2^n$, so its $+1$ and -1 eigenspaces each have dimension $d/2$; for a Haar random pure state the mass in the $+1$ eigenspace is $p_+ \sim \text{Beta}(d/2, d/2)$ and $\langle P\rangle = 2p_+ - 1$. The acceptance condition $\langle P\rangle \geq F^\star$ thus reads $p_+ \geq (1 + F^\star)/2$, which for T6.6 at $d = 32$ is $p_+ \geq 0.995$, and

$$\Pr\big[\text{Beta}(16, 16) \geq 0.995\big] \approx 4.27 \times 10^{-29}, \tag{30}$$

evaluated exactly from $\Pr[\text{Beta}(a, b) \geq x] = \sum_{k=0}^{a-1} \binom{a+b-1}{k} x^k (1-x)^{a+b-1-k}$ (released as `t66_bound.json`). One generator already closes the case, because the conjunction over the four is contained in the event for any single one; the runner demands all four. T6.6 does not fix its qubit layout and has $n = 5$ qubits, so the runner also tries the 120 qubit relabelings, and the union over them is at most $120 \times 4.273 \times 10^{-29} < 5.2 \times 10^{-27}$. The acceptance event of the deployed test therefore sits far below the 0.01 escalation threshold instead of a factor of six above it, and its bound does depend on $F^\star$.

Two consequences follow. First, the tabulated $2^{-n_{\text{stab}}} = 0.0625$ prices the sign coincidence, and the standing escalation status of T6.6 rests on that reading. The $\varepsilon_{\text{theo}}$ column keeps that value and the escalation classification is unchanged; the sharpened column carries the single generator bound of Eq. (30) at the identity qubit layout, not the union over the relabelings. Second, the empirical Haar calibration of this row is no independent check of the figure. Its sampler draws no random state for T6.6: it tosses n_{stab} fair coins and counts a hit when all of them come up zero, so its expectation is $2^{-n_{\text{stab}}}$ by construction and it can only confirm the formula it already assumes. T6.6 is the only one of the 34 escalation rows for which this holds. Of the remaining rows, 19 draw the quantity the test decides on, though not always under the test's own rule: T2.17 and T2.18 at a marginal threshold of 0.1 instead of 0.01, T4.4 and T4.6 on a single target instead of the conjunction of their parameter cases, and the L5 rows without a fixed layout on one target instead of the union over its relabeled copies. The other 14 draw a Haar object but score a different predicate. Among those 14 are 5 tasks — T2.9, T2.21, T2.24, T2.26 and T6.16 — that dataset v1.1.0 turned from single case state fidelity into functional tasks; their calibration rows were re-run with the functional sampler after that change.

C.9 Lemma L7 (TV Distribution Match, Monte-Carlo)

For probability and functional C3 (distribution) tasks, L7 is a Monte Carlo estimate of the probability of the event $\{\|p - p_{\text{target}}\|_{\text{TV}} \leq \delta\}$ under the Haar null; closed form upper bounds on events of this kind are given by rule 3 of L9c (§C.11), and the three L7 rows have the closed forms of the remark below. The Monte Carlo runs released in `lemma_L7_mc.csv` do not estimate this total variation

event: they estimate argmax, top K and minimum peak mass events of the rows they cover. For $p \sim \mathrm{Dirichlet}(1, \ldots, 1)$ on Δ^{d-1},

$$\varepsilon_T = \Pr_{p \sim \mathrm{Dir}(1,\ldots,1)}[\|p - p_{\text{target}}\|_{\text{TV}} \le \delta] \approx \hat{\varepsilon}_{\text{MC}} \pm \sigma_{\text{MC}}, \qquad \sigma_{\text{MC}} \le \frac{1}{2\sqrt{N}}. \tag{31}$$

For $N = 10^9$ samples this gives $\sigma_{\text{MC}} \le 1.59 \times 10^{-5}$. A complementary Hoeffding shot noise correction governs finite N_{shots} estimates of each entry of p,

$$\Pr[|\hat{p}_i - p_i| \ge \eta] \le 2\exp\left(-2N_{\text{shots}}\,\eta^2\right). \tag{32}$$

Proof. The argument combines the Dirichlet representation of the Haar pure state measure from Zyczkowski & Sommers (2001) with the concentration of bounded sums from Hoeffding (1963). The Monte Carlo estimator

$$\hat{\varepsilon}_{\text{MC}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\|p^{(n)} - p_{\text{target}}\|_{\text{TV}} \le \delta\}, \tag{33}$$

with $p^{(n)} \overset{\text{i.i.d.}}{\sim} \mathrm{Dir}(1, \ldots, 1)$, is an unbiased estimator of ε_T. Each summand is Bernoulli(ε_T), so $\mathrm{Var}(\hat{\varepsilon}_{\text{MC}}) = \varepsilon_T(1-\varepsilon_T)/N \le 1/(4N)$, which gives the standard error $\sigma_{\text{MC}} \le 1/(2\sqrt{N})$. The bound $\sigma_{\text{MC}} \le 1/(2\sqrt{N})$ is the worst case at $\varepsilon_T = 1/2$ (maximum Bernoulli variance). In the low rate regime $\varepsilon_T \ll 0.01$ relevant to most QuARC tasks, $\sigma_{\text{MC}} \approx \sqrt{\varepsilon_T/N}$ is substantially tighter, which harmonizes with the per task Wilson 95% intervals reported in Tab. 7. The Hoeffding bound applied to each empirical probability $\hat{p}_i$, a sum of N_{shots} Bernoulli outcomes, yields the stated tail probability.

The bound applies to the probability and functional C3 (distribution) verification types. The released file `lemma_L7_mc.csv` records, for the thirteen rows it covers, an estimate $\hat{\varepsilon}_{\text{MC}}$ and a 95% interval from 10^8 draws; it records no σ_{MC}, and none of its values enters Tab. 6. As an instance, at $N = 10^9$ Monte Carlo samples the standard error is

$$\sigma_{\text{MC}} \le \frac{0.5}{\sqrt{10^9}} < 1.59 \times 10^{-5}, \tag{34}$$

three orders of magnitude below the 0.01 escalation threshold. The seed convention is seed $=$ (tier $\cdot\, 10^4 +$ subindex) $\cdot\, 10^3 + 42$, consistent with the Haar sampling seed convention of App. D.

Remark (closed forms for the three L7 rows under dataset v1.2.0). All three rows of Tab. 6 carried by L7 admit an exact closed form under the flat Dirichlet null, so no $\varepsilon_{\text{theo}}$ entry of these rows is a Monte Carlo estimate. Two facts are stated first and then applied to the three tasks. By the Dirichlet representation of the Haar pure state measure (Zyczkowski & Sommers, 2001), the computational basis probabilities p of a Haar random state on $d = 2^n$ coordinates follow $\mathrm{Dirichlet}(1, \ldots, 1)$, whose density is the constant $(d-1)!$ on the simplex Δ^{d-1}. Fix an index set S with $|S| = s$ and minimum masses $a_i \ge 0$ for $i \in S$, and write $T = 1 - \sum_{i\in S} a_i$ for the slack. Assuming $T > 0$,

$$\Pr[\,p_i \ge a_i \ \text{ for all } i \in S\,] \;=\; T^{\,d-1}. \tag{35}$$

The shift $y_i = p_i - a_i$ for $i \in S$ and $y_j = p_j$ otherwise carries this event onto $\{\, y_k \ge 0,\ \sum_k y_k = T \,\}$, a copy of the simplex scaled by T, whose volume is $T^{\,d-1}$ times that of Δ^{d-1}; multiplying by the constant density $(d-1)!$ yields (35). For $s = 2$ and two indices $i \ne j$ this is the joint tail

$$\Pr[\,p_i \ge a_i \ \wedge\ p_j \ge a_j\,] \;=\; (1 - a_i - a_j)^{\,d-1}. \tag{36}$$

Conditioned on the event of (35), the shifted vector y is uniform on the simplex scaled by T, so y/T is again flat $\mathrm{Dirichlet}(1, \ldots, 1)$ on d coordinates. By the aggregation property of the flat Dirichlet law, the sum of any r of its coordinates follows $\mathrm{Beta}(r, d-r)$; hence $\sum_{i\in S} y_i/T = Y$ with $Y \sim \mathrm{Beta}(s, d-s)$, and, since $y_i = p_i - a_i$, $\sum_{i\in S} p_i = \sum_{i\in S} a_i + T\,Y$. Therefore, for any threshold τ with $\sum_{i\in S} a_i < \tau \le 1$,

$$\Pr\big[\,p_i \ge a_i\ (i \in S)\ \wedge\ \textstyle\sum_{i\in S} p_i \ge \tau\,\big] \;=\; T^{\,d-1}\,\Pr\Big[\,\mathrm{Beta}(s, d-s) \ge \tfrac{\tau - \sum_{i\in S} a_i}{T}\,\Big], \tag{37}$$

and the second factor is 1 when $\tau \le \sum_{i\in S} a_i$, and also when $s = d$ (then $\sum_{i\in S} p_i = 1$ identically). Both (35) and (37) are exact for this model event. The shipped probability test (runner) accepts a

run when every declared peak index i carries marginal mass at least its `min_probability` and the declared peaks together carry at least the `total_peak_probability` (default 0.5); before forming the marginals it drops every basis probability below 10^{-10}. Dropping mass can only lower a marginal, so the runner's acceptance event is contained in the model event of (35) and (37): the closed forms are exact for the model event and upper bounds for the runner's event. On the three tasks below each declared peak is a single coordinate and every declared minimum exceeds 10^{-10}, so a peak coordinate that meets its minimum is never dropped, and there the two events coincide. All three tasks fix the qubit layout, so the harness applies no relabeling and no orbit factor enters, and under dataset v1.2.0 each peak marginal runs over all n qubits, so every declared peak is a single coordinate p_i. For T2.22 (Grover, $d = 8$) two peaks carry a minimum of 0.45 each and the total peak threshold is $0.9 = 0.45 + 0.45$, so the total mass condition is already implied by the two per index conditions ($\tau = \sum_{i\in S} a_i$, second factor 1) and $T = 1 - 0.9 = 0.1$; (35) gives

$$\varepsilon_{\text{T2.22}} \;=\; (1 - 0.45 - 0.45)^7 \;=\; (0.1)^7 \;=\; 10^{-7}. \tag{38}$$

For T2.12 ($d = 8$) all eight indices are peaks, each with a minimum of 99% of its canonical probability, so $\sum_{i\in S} a_i = 0.99$, $s = d = 8$, and $T = 1 - 0.99 = 0.01$; the total mass condition holds automatically because $s = d$, and (35) gives

$$\varepsilon_{\text{T2.12}} \;=\; (1 - 0.99)^7 \;=\; (0.01)^7 \;=\; 10^{-14}. \tag{39}$$

For T6.15 (Shor, $d = 128$) sixteen peaks carry a minimum of 0.0375 each, so $\sum_{i\in S} a_i = 16 \times 0.0375 = 0.6$ and $T = 0.4$, with a total peak threshold of 0.9; here $s = 16 < d$, the reduced threshold is $(0.9 - 0.6)/0.4 = 0.75$, and (37) gives

$$\varepsilon_{\text{T6.15}} \;=\; 0.4^{\,127}\,\Pr[\,\mathrm{Beta}(16, 112) \geq 0.75\,] \;\approx\; 1.75 \times 10^{-101}. \tag{40}$$

These three values are the $\varepsilon_{\text{theo}}$ entries of Tab. 6 for the L7 rows; no Monte Carlo estimate enters them. The minima enter these closed forms as the decimal values of the task specifications; the released floating point minima of T2.12 and T6.15 fall short of them, in sum by less than 10^{-14}, which raises the model event of T2.12 by a relative amount below 3×10^{-12} and that of T6.15 by one below 3×10^{-14}, and no printed digit changes.

C.10 Lemma L8 (L5 Analogue on Code Subspace, Projected Haar Null)

For a code subspace $\mathcal{C} \subseteq \mathbb{C}^d$ with projector $\Pi_{\mathcal{C}}$ and $d_{\text{code}} = \dim \mathcal{C} \leq d$, a Haar-random adversarial unitary $U \sim \mu_{\text{Haar}}$ on the *full* group $U(d)$ with code space restriction $U|_{\mathcal{C}} = \Pi_{\mathcal{C}} U \Pi_{\mathcal{C}}$, and the logical block $c|_{\mathcal{C}}$ (a $d_{\text{code}} \times d_{\text{code}}$ unitary) of a fixed code preserving target circuit $c \in U(d)$, the projected code subspace Markov bound reads

$$\Pr_{U\sim\mu_{\text{Haar}}(U(d))}\big[\,|\mathrm{tr}_{\mathcal{C}}((U|_{\mathcal{C}})^\dagger\, c|_{\mathcal{C}})/d_{\text{code}}|^2 \geq F^\star\big] \leq \frac{1}{d_{\text{code}}\, d\, F^\star}. \tag{41}$$

Proof. The argument follows Collins & Śniady (2006) on $U(d)$, exactly as in L5, with the trace restricted to the code subspace. The Weingarten second moment formula on $U(d)$ gives

$$\mathbb{E}_{U\sim\mu_{\text{Haar}}}\big[U_{ij}\,\overline{U_{kl}}\big] = \delta_{ik}\delta_{jl}/d. \tag{42}$$

Expanding the restricted trace over code space indices,

$$\begin{aligned}\mathbb{E}\big[|\mathrm{tr}_{\mathcal{C}}((U|_{\mathcal{C}})^\dagger\, c|_{\mathcal{C}})|^2\big] &= \sum_{i,j,k,l\in\mathcal{C}} (c|_{\mathcal{C}})_{ij}\,\overline{(c|_{\mathcal{C}})_{kl}}\;\mathbb{E}\big[\overline{U_{ij}}\,U_{kl}\big] \\ &= \frac{1}{d}\sum_{i,j\in\mathcal{C}} |(c|_{\mathcal{C}})_{ij}|^2 = \frac{\mathrm{tr}_{\mathcal{C}}\big((c|_{\mathcal{C}})^\dagger\,(c|_{\mathcal{C}})\big)}{d} = \frac{d_{\text{code}}}{d},\end{aligned} \tag{43}$$

using that $c|_{\mathcal{C}}$ is unitary on $\mathcal{C}$ for a code preserving c, so $\mathrm{tr}_{\mathcal{C}}((c|_{\mathcal{C}})^\dagger c|_{\mathcal{C}}) = d_{\text{code}}$. Hence the normalized fidelity variable $X = |\mathrm{tr}_{\mathcal{C}}((U|_{\mathcal{C}})^\dagger c|_{\mathcal{C}})/d_{\text{code}}|^2$ has $\mathbb{E}[X] = (d_{\text{code}}/d)/d_{\text{code}}^2 = 1/(d_{\text{code}}\, d)$, and Markov's inequality yields $\Pr[X \geq F^\star] \leq 1/(d_{\text{code}}\, d\, F^\star)$.

Remark. Drawing a Haar random *logical* unitary directly on $U(d_{\text{code}})$ would give the weaker constant $1/(d_{\text{code}}^2 F^\star)$. That null does not match the threat model or the empirical calibration of App. D, both

of which let the adversarial circuit act on the full d dimensional space and then restrict it to a d_{code} dimensional block (the runner to $\mathcal{C}$ against the logical target; the calibration sampler draws $U \sim \mu_{\text{Haar}}$ on $U(d)$ and takes a fixed $d_{\text{code}} \times d_{\text{code}}$ block against the identity, which has the same law under the Haar measure). The projected null stated above matches the calibration and is sharper by the factor d/d_{code}.

The bound applies to the code space unitary fidelity verification type. As an instance, task T6.8 is a code space unitary fidelity check on the logical Z gate of the $[\![3, 1, 1]\!]$ bit flip repetition code with $|0_L\rangle = |000\rangle$ and $|1_L\rangle = |111\rangle$, so $d = 8$, $d_{\text{code}} = 2$, and $F^\star = 0.99$. Substituting,

$$\varepsilon_{\text{T6.8}} = \frac{1}{d_{\text{code}}\, d\, F^\star} = \frac{1}{2 \cdot 8 \cdot 0.99} \approx 0.0631, \tag{44}$$

still above the 0.01 escalation threshold, and T6.8 remains an escalation candidate under the Markov form. The runner's relabeling retry does not widen this event: the code words $|000\rangle$ and $|111\rangle$ are fixed by every qubit permutation, so the projected block is the same for every relabeling. Empirical Haar sampling on $U(d)$ followed by code space projection gives $\hat{\varepsilon} = 0$ at $N = 10^9$, with Wilson 95% upper bound 3.84×10^{-9} (Tab. 7).

A closed form for the truncation. The moment route of §C.7 does not transfer to this lemma. There $A = \Pi U \Pi$ is the $d_{\text{code}} \times d_{\text{code}}$ truncation of a Haar $U(d)$, and the identity $\mathbb{E}|\operatorname{tr} A|^{2k} = k!$ of Eq. (22) fails for truncations: the moments of the truncated ensemble are not those of the full group. A second route does apply, and it is sharper than Eq. (44) by more than twenty orders of magnitude.

First the acceptance event is embedded in a spectral one. Since $c|_{\mathcal{C}}$ is unitary on $\mathcal{C}$, Cauchy–Schwarz gives $|\operatorname{tr}_{\mathcal{C}}(A^\dagger c)|^2 \le \operatorname{tr}(A^\dagger A)\, \operatorname{tr}(c^\dagger c) = d_{\text{code}} \operatorname{tr}(A^\dagger A)$, so the acceptance requirement $|\operatorname{tr}_{\mathcal{C}}(A^\dagger c)/d_{\text{code}}|^2 \ge F^\star$ implies

$$\operatorname{tr}(A^\dagger A) \;\ge\; d_{\text{code}} F^\star, \tag{45}$$

which for T6.8 reads $\operatorname{tr}(A^\dagger A) \ge 1.98$. The spectral event is a strict superset of the acceptance event, so its probability is a valid upper bound.

Second, that probability is available in closed form. For the $M \times M$ truncation of a Haar $U(N)$ with $N \ge 2M$, the eigenvalues $\lambda_1, \ldots, \lambda_M$ of $A^\dagger A$ follow the Jacobi ensemble of truncated Haar unitaries, with joint density

$$P(\lambda) \;\propto\; \prod_{i<j} (\lambda_i - \lambda_j)^2 \prod_i (1 - \lambda_i)^{N-2M} \qquad \text{on } [0,1]^M, \tag{46}$$

normalized by its integral over $[0, 1]^M$; we use this density without proof here, and the first control below checks its mean against direct sampling. For T6.8, $M = d_{\text{code}} = 2$ and $N = d = 8$, so the exponent is $N - 2M = 4$. The substitution $u_i = 1 - \lambda_i$ turns the event $\lambda_1 + \lambda_2 \ge d_{\text{code}} F^\star$ into $u_1 + u_2 \le 2 - d_{\text{code}} F^\star = 0.02$, a corner of the simplex on which Eq. (46) is a polynomial; the integral is a finite Dirichlet integral and is evaluated in exact rational arithmetic. The result is

$$\varepsilon_{\text{s,T6.8}} = \frac{1}{32\,226\,562\,500\,000\,000\,000\,000} \approx 3.10 \times 10^{-23}, \tag{47}$$

which is the $\log_{10} \varepsilon_{\text{s}}$ entry -22.5 of Tab. 6.

Two controls are worth naming, because a closed form this small invites the suspicion that something has been dropped. The density of Eq. (46) is checked at its mean: integrating it gives $\mathbb{E}[\operatorname{tr}(A^\dagger A)] = M^2/N = 0.5$ for T6.8, the known size of the truncation, and the same quantity sampled from the ensemble returns 0.5004 over 2×10^5 draws — the mean falls out of the density rather than being put in. And the new bound is compared against the published one: 3.10×10^{-23} must be smaller than the 0.0631 of Eq. (44), or it would not be a sharpening at all.

C.11 Lemma L9 (Conjunctive Functional Acceptance, Haar Chain Bound)

The `functional` verification type accepts a submission only if *every* one of its test cases passes; the released harness decides `passed = passed_count == total`. Instruments L1 to L8 price a single acceptance condition or, as L4 and the corollary of L6, a conjunction of argmax or sign events; none of them prices the conjunction of the shipped per case rules, which add total variation and band tolerances. The mapping table assigns the `L2, L3` rows the form $K\binom{d}{K}^{-1}$, that is K times

Eq. (12), evaluated at K_c on the ten rows named in §C.1, as a stand in. L9 closes the conjunction in closed form. It is not a sampling instrument: nothing is drawn, the chain is evaluated.

Let $d = 2^n$ be the data register dimension, $U \sim \mu_{\text{Haar}}$ on $U(d)$, and for integers $1 \le r < d'$ write

$$B(r, d', t) \;=\; \Pr[\text{Beta}(r,\, d' - r) \ge t] \;=\; \sum_{j=0}^{r-1} \binom{d'-1}{j} t^j (1-t)^{d'-1-j}, \tag{48}$$

with $B = 1$ for $t \le 0$ or $r \ge d'$, and $B = 0$ for $t > 1$.

L9a (mass bound in a subspace). Let $W \subseteq \mathbb{C}^d$ be a subspace of dimension d', let w be Haar distributed on the unit sphere $S(W)$, and let S be an index set with $|S| = s$. Then, with $r = \min(s, d')$,

$$\Pr\big[\textstyle\sum_{i\in S} |\langle i|w\rangle|^2 \ge t\big] \;\le\; B(r, d', t), \tag{49}$$

independently of W, with equality exactly when $\dim\big(\text{span}\{|i\rangle : i \in S\} \cap W\big) = r$.

Proof. Write $M = \text{span}\{|i\rangle : i \in S\}$ for the coordinate subspace indexed by S, and let $P = \sum_{i\in S} |i\rangle\langle i|$ be the orthogonal projection onto M. Since the vectors $|i\rangle$, $i \in S$, are orthonormal, Parseval's identity gives

$$\sum_{i\in S} |\langle i|w\rangle|^2 = \|Pw\|^2 = \langle w|G|w\rangle, \tag{50}$$

where G is the compression of P to W, that is, the operator on W fixed by $\langle u|G|v\rangle = \langle u|P|v\rangle$ for all $u, v \in W$. The middle equality holds because $P = P^\dagger = P^2$, and the last equality holds for $w \in W$ by the definition of G.

The operator G is Hermitian and positive semidefinite with every eigenvalue at most one, because $\langle u|G|u\rangle = \|Pu\|^2$ lies between zero and $\|u\|^2$ for each $u \in W$, an orthogonal projection being a contraction. Moreover G is the Gram operator of the map $u \mapsto Pu$ on W, so its rank equals $\dim P(W) \le \min(\dim W, \text{rank}\, P) = \min(d', s) = r$; hence at most r of its eigenvalues are nonzero.

By the spectral theorem for the Hermitian operator G on the finite dimensional space W, choose an orthonormal eigenbasis of W with real eigenvalues λ_j attached to unit eigenvectors e_j, ordered so that λ_j vanishes when j exceeds r; each λ_j is nonnegative and at most one. Setting $c_j = |\langle e_j|w\rangle|^2$, Eq. (50) becomes $\sum_{i\in S} |\langle i|w\rangle|^2 = \sum_j \lambda_j c_j$.

The law of w is invariant under the unitary group of W, which is what it means for w to be Haar distributed on $S(W)$. Hence the squared amplitudes carry the same joint law in every orthonormal basis of W. In the eigenbasis $\{e_j\}$ that law is, by §C.5 applied in dimension d' in place of d, the flat Dirichlet law on the simplex in dimension d'. Because λ_j vanishes for j exceeding r, while each remaining λ_j is at most one and each c_j is nonnegative, the weighted sum is dominated almost surely,

$$\sum_j \lambda_j c_j \;\le\; \sum_{j\le r} c_j =: Y. \tag{51}$$

The variable Y, being the sum of r coordinates of a flat Dirichlet law in dimension d', follows $\text{Beta}(r,\, d' - r)$ by the Dirichlet aggregation property of §C.9. Monotonicity of probability under the domination (51) therefore gives, for every threshold t,

$$\Pr\Big[\textstyle\sum_{i\in S} |\langle i|w\rangle|^2 \ge t\Big] \;\le\; \Pr[\,Y \ge t\,] \;=\; B(r, d', t), \tag{52}$$

where the last equality is the definition of B. The right side involves W only through d', since $r = \min(s, d')$, so the bound is the same for every W of dimension d'; this is the asserted independence of W.

It remains to identify when (52) is an identity for every t, equivalently when the mass carries exactly the law $\text{Beta}(r, d' - r)$. Writing $X = \sum_j \lambda_j c_j$, the gap $Y - X = \sum_{j\le r}(c_j - \lambda_j c_j)$ is nonnegative

almost surely, so the two laws agree only if $\mathbb{E}[Y - X]$ vanishes, which for a nonnegative variable forces $Y = X$ almost surely. The coordinates of the flat Dirichlet law are almost surely positive, so $Y = X$ almost surely holds if and only if every λ_j with $j \le r$ equals one; conversely, if all of these eigenvalues equal one then $X = Y$ almost surely and the bound is an identity.

Finally, the eigenspace of G for the eigenvalue one is exactly $M \cap W$. For a unit vector $u \in W$, the relation $Gu = u$ forces $\|Pu\|^2 = \|u\|^2$, hence $u \in \operatorname{range} P = M$; and every $u \in M \cap W$ satisfies $Pu = u$, hence $Gu = u$. Since G has at most r nonzero eigenvalues and $\dim(M \cap W) \le \min(s, d') = r$, every λ_j with $j \le r$ equals one if and only if this eigenspace has dimension r, that is $\dim(M \cap W) = r$. Combined with the preceding paragraph this proves the equality condition. When $\dim(M \cap W) < r$, some λ_j with $j \le r$ is below one, so $Y - X$ is positive on a set of positive probability, the two laws differ, and (52) is strict at some t.

L9b (chain bound). Let $|x_1\rangle, \dots, |x_K\rangle$ be pairwise orthonormal, let U be Haar distributed on $\mathrm{U}(d)$, and set $v_k = U|x_k\rangle$. For each case k let E_k be a measurable set of unit vectors of $\mathbb{C}^d$, and let $A_k = \{v_k \in E_k\}$ be the acceptance event of case k. Suppose that

$$\beta_k \;\ge\; \sup_{\dim W = d-k+1} \; \Pr_{w \sim \mathrm{Haar}(S(W))}[w \in E_k]\,. \tag{53}$$

Then

$$\Pr[A_1 \wedge \cdots \wedge A_K] \;\le\; \prod_{k=1}^{K} \beta_k. \tag{54}$$

Proof. Fix the numbering of the cases and condition on the earlier images $v_1, \dots, v_{k-1}$. By the QR construction of the Haar measure (Mezzadri, 2006), the left invariance of U makes v_k Haar distributed on the unit sphere of the orthogonal complement of $\operatorname{span}\{v_1, \dots, v_{k-1}\}$. That complement is a subspace W of dimension $d - k + 1$, because $v_1, \dots, v_{k-1}$ are the images of orthonormal vectors under a unitary and are therefore linearly independent. For this W the conditional acceptance probability equals $\Pr_{w \sim \mathrm{Haar}(S(W))}[w \in E_k]$, which the supremum in (53) dominates, so

$$\Pr[A_k \mid v_1, \dots, v_{k-1}] \;\le\; \beta_k \qquad \text{almost surely.}$$

Write $C_k = A_1 \wedge \cdots \wedge A_k$. Because C_{k-1} depends on $v_1, \dots, v_{k-1}$ alone, its indicator $\mathbf{1}_{C_{k-1}}$ is a function of those vectors, and the tower property gives

$$\Pr[C_k] = \mathbb{E}\big[\mathbf{1}_{C_{k-1}} \Pr[A_k \mid v_1, \dots, v_{k-1}]\big] \;\le\; \beta_k \Pr[C_{k-1}],$$

where the inequality inserts the preceding display under the nonnegative weight $\mathbf{1}_{C_{k-1}}$ and uses $\mathbb{E}[\mathbf{1}_{C_{k-1}}] = \Pr[C_{k-1}]$. The conditioning runs on the vectors $v_1, \dots, v_{k-1}$ throughout and never on the earlier events $A_1, \dots, A_{k-1}$; conditioning on the events is the step that §C.6 rejects. Starting from $\Pr[C_1] = \Pr[A_1] \le \beta_1$ and iterating the recursion up to $k = K$ yields $\Pr[C_K] \le \prod_{k=1}^{K} \beta_k$, which is (54). The derivation uses no property of the numbering beyond the labels $1, \dots, K$, so the same bound holds for every ordering of the cases. The bound is not asserted to be tight: β_k only dominates the supremum in (53), so each conditional factor may fall below β_k, and whether the factors are attained simultaneously is not determined here.

Remark (a product of the marginals would not be a bound). Replacing the conditional factors of (54) by the unconditional single case probabilities is not conservative in general: the cases can be positively associated, and then the naive product falls *below* the true conjunction probability. The dimension lowering chain $d \to d - 1 \to d - 2 \to \dots$ is what makes the product valid.

L9c (link bounds for the shipped acceptance rules). Let a chain link have nesting dimension $d' \le D$, where $D = d$ is the ambient dimension of the functional test and the first link has $d' = D$. Every case of the shipped dataset is decided by one of four paths of the released runner (a fifth, an X basis case that declares no ancilla mass and is accepted unconditionally, occurs in no shipped task), and for each the value stated below is a valid choice of β_k in (53). Rule 1, an ancilla marginal within η of a declared target mass $p^\star$ on the ancilla zero half of the basis, takes

$$\beta \;=\; \min\big\{\, B(\min(D/2,\, d'),\, d',\, p^\star - \eta),\; B(\min(D/2,\, d'),\, d',\, 1 - p^\star - \eta) \,\big\}.$$

Rule 2, a per index comparison against declared targets q_i within tolerance τ_2 on a declared index set S, takes, with $a_i = \max(0,\ q_i - \tau_2)$,

$$\beta \;=\; \min_{\emptyset \neq J \subseteq S} B\big(\min(|J|,\ d'),\ d',\ \textstyle\sum_{i\in J} a_i\big),$$

and in the first link, where $d' = D$, the probability (55) at the band with a_i and $b_i = \min(1,\ q_i + \tau_2)$ is also a valid β. Rule 3, argmax agreement together with total variation distance at most τ_3 against the canonical output q on its support S, takes

$$\beta \;=\; \min_{\emptyset \neq J \subseteq S} B\big(\min(|J|,\ d'),\ d',\ \textstyle\sum_{i\in J} q_i - \tau_3\big),$$

and in the first link the probability (55) at the band with $a_i = \max(0,\ q_i - \tau_3)$ and $b_i = \min(1,\ q_i + \tau_3)$ is again a valid β. The fallback path, a case that carries no canonical distribution and constrains the argmax alone, takes $\beta = 1/D$ in the first link and $\beta = 1$ in every later link. The tolerances η, τ_2 and τ_3 are read from the released runner rather than restated here, so rerunning the bound script after a change to the harness moves the bound; the released values are those of the shipped runner.

Proof. Fix a link. By the reduction of L9b, its image $v_k = U|x_k\rangle$, conditioned on the earlier images, is Haar distributed on the unit sphere of a subspace W of dimension d', so any bound on $\Pr[v_k \in E_k]$ that holds uniformly over such W is a valid β_k through (53). Each path bounds this probability.

Rule 1. Acceptance requires the mass of v_k on the ancilla zero half of the basis to exceed $p^\star - \eta$ and its mass on the complementary half to exceed $1 - p^\star - \eta$; each half is an index set of $D/2$ elements. Applying L9a (49) to the first half at threshold $p^\star - \eta$ and to the second at threshold $1 - p^\star - \eta$, each with $r = \min(D/2, d')$, bounds the probability of the corresponding half condition by the matching B value. Acceptance implies both half conditions at once, so by monotonicity of probability its probability is at most the smaller of the two B values, which is the stated β. Each L9a step is an equality exactly under the equality condition of (49), and the minimum is attained only when the binding half meets it.

Rule 2. With $a_i = \max(0,\ q_i - \tau_2)$, the per index test $|p_i - q_i| \le \tau_2$ for $i \in S$ gives $p_i \ge q_i - \tau_2$ and, with $p_i \ge 0$, therefore $p_i \ge a_i$; summing the terms of any nonempty $J \subseteq S$ yields $\sum_{i\in J} p_i \ge \sum_{i\in J} a_i$. For each such J, L9a with index set J and $r = \min(|J|, d')$ bounds $\Pr[\sum_{i\in J} p_i \ge \sum_{i\in J} a_i]$ by $B(\min(|J|, d'), d', \sum_{i\in J} a_i)$. Acceptance implies every one of these events, so by monotonicity its probability is at most their minimum over nonempty J, the stated β. In the first link $d' = D$ and W is the full space, so by §C.5 the coordinates of v_k carry the flat Dirichlet law on all D coordinates; acceptance then also forces $a_i \le p_i \le b_i$ for all $i \in S$ with $b_i = \min(1,\ q_i + \tau_2)$, whose probability is (55) exactly, a second valid β for that link.

Rule 3. Total variation distance equals the largest mass difference over all index sets, $\mathrm{TV}(p, q) = \max_A |\sum_{i\in A}(p_i - q_i)|$, so in particular $|\sum_{i\in J}(p_i - q_i)| \le \mathrm{TV}(p, q)$ for every $J \subseteq S$. The test $\mathrm{TV}(p, q) \le \tau_3$ therefore forces $\sum_{i\in J}(q_i - p_i) \le \tau_3$, that is $\sum_{i\in J} p_i \ge \sum_{i\in J} q_i - \tau_3$, for every $J \subseteq S$; the tolerance enters once per subset and not once per index, and the singleton sets give $|p_i - q_i| \le \tau_3$ for each i. For every nonempty J, L9a with index set J bounds $\Pr[\sum_{i\in J} p_i \ge \sum_{i\in J} q_i - \tau_3]$ by $B(\min(|J|, d'), d', \sum_{i\in J} q_i - \tau_3)$, and acceptance implies all of these events, so by monotonicity its probability is at most their minimum, the stated β. The argmax agreement that the case also demands is discarded here; it only removes accepted vectors, so the true acceptance probability lies below the stated bound and the bound stays valid but is not tight. In the first link the band event $a_i \le p_i \le b_i$ with $a_i = \max(0,\ q_i - \tau_3)$ and $b_i = \min(1,\ q_i + \tau_3)$ again holds under acceptance, and (55) prices it exactly.

Fallback. When the case fixes only the argmax, the first link asks that a Haar vector on the full space attain its largest coordinate at one declared index; the D coordinate probabilities are exchangeable under §C.4 and ties carry probability zero, so this event has probability $1/D$, the stated first link value. On a later link v_k is confined to a subspace of dimension $d' < D$, where the coordinates are no longer exchangeable and the argmax probability can exceed $1/D$: on $\mathrm{span}\{|0\rangle, |1\rangle\}$ inside $\mathbb{C}^4$ it is $1/2$. The later link therefore takes $\beta = 1$, which is valid for any event.

Box formula. For $p \sim \mathrm{Dirichlet}(1, \dots, 1)$ on D coordinates, an index set S, and reals $0 \le a_i \le b_i \le 1$,

$$\Pr[\, a_i \le p_i \le b_i \text{ for all } i \in S \,] \;=\; \sum_{J \subseteq S} (-1)^{|J|} \Big(1 - \textstyle\sum_{i \in S \setminus J} a_i - \sum_{j \in J} b_j\Big)_+^{D-1}. \tag{55}$$

Proof. The band event $\{a_i \le p_i \le b_i \text{ for all } i \in S\}$ is the event $\{p_i \ge a_i \text{ for all } i \in S\}$ with the union $\bigcup_{j \in S} \{p_j > b_j\}$ of upper end violations removed. Inclusion and exclusion over the set J of violated upper ends writes its probability as

$$\sum_{J \subseteq S} (-1)^{|J|} \Pr[\, p_i \ge a_i \ (i \in S \setminus J),\ p_j \ge b_j \ (j \in J) \,],$$

where replacing each strict inequality $p_j > b_j$ by $p_j \ge b_j$ moves the event by a null set only, because the flat Dirichlet law is atomless, and so leaves each probability unchanged. Every term is a joint lower bound on a set of coordinates, so by the minimum mass identity of the remark to §C.9, (35), it equals $(1 - \sum_{i \in S \setminus J} a_i - \sum_{j \in J} b_j)_+^{D-1}$, the positive part covering the case where the lower bounds already sum above one and the event is empty. Collecting the terms gives (55). The identity uses the flat Dirichlet law on all D coordinates, so it is available in the first link alone; on a later link v_k is confined to a subspace and its coordinate probabilities are not flat Dirichlet, so (55) does not apply there.

Across the 15 `L2,L3` rows that L9 prices, only two of the four paths are exercised: the per index comparison accounts for 24 of the chain links, and argmax agreement with total variation for the remaining 100, of which 98 lie in the 14 escalating rows. Neither the ancilla rule nor the fallback fires on any of these links; a different task set could exercise either path. These four paths are the ones the bound script recognizes, and the block claims completeness in no stronger sense. The tolerances are read from the released runner rather than restated, so rerunning the bound script after a change to the harness moves the bound; the released values are those of the shipped runner.

L9d (task bound). Let N be an upper bound on the number of qubit relabelings the harness tries on a task: $N = 1$ when the task declares `fixed_qubit_layout`, $N = n!$ when $n \le 5$, the permutation limit of the harness, and $N = 2$ when $n > 5$, the identity together with at most one layout retry. Then the Haar probability that the task is accepted is at most N times the chain product of L9b for any one ordering of its K cases. The released artifact takes, for a task of at most 7 cases, the smallest such product over all orderings, and for more cases a single fixed ordering, the cases sorted by ascending β_k at the full dimension d; both orderings are valid by L9b.

Proof. Let $\tau = 1, \dots, N$ index a set of relabelings that contains every relabeling the harness tries, and let $A_\tau \subseteq \mathrm{U}(d)$ be the set of unitaries it accepts under relabeling τ. Write $A_{\mathrm{id}} = \bigcap_{k=1}^{K} A_k$ for the accepted set of the untransformed specification, the object L9b bounds; the identity is one of the tried relabelings. The retry accepts a candidate once any tried relabeling does, so the task's accepted set is the union $A = \bigcup_{\tau=1}^{N} A_\tau$.

Relabeling the qubits of the data register permutes its tensor factors, so relabeling τ acts on a candidate unitary U by conjugation $U \mapsto P_\tau U P_\tau^\dagger$ with the permutation matrix P_τ of that relabeling, a bijection of $\mathrm{U}(d)$ that carries A_{id} onto A_τ. The Haar measure on $\mathrm{U}(d)$ is invariant under both left and right multiplication (Mezzadri, 2006), hence under conjugation by the unitary P_τ, so every copy carries the mass of the original, $\mu_{\mathrm{Haar}}(A_\tau) = \mu_{\mathrm{Haar}}(A_{\mathrm{id}})$. By L9b each ordering of the K cases gives $\mu_{\mathrm{Haar}}(A_{\mathrm{id}}) \le \prod_{k=1}^{K} \beta_k$, with β_k the bound of the chain link at position k in that ordering, (53). Finite subadditivity of μ_{Haar}, the equal masses of the N copies, and this bound combine to

$$\mu_{\mathrm{Haar}}(A) \le \sum_{\tau=1}^{N} \mu_{\mathrm{Haar}}(A_\tau) = N\, \mu_{\mathrm{Haar}}(A_{\mathrm{id}}) \le N \prod_{k=1}^{K} \beta_k. \tag{56}$$

The last inequality holds for the β_k of any one ordering, so it holds for the smallest product over all orderings and for any single fixed ordering alike; the released artifact takes the former when the task has at most 7 cases and the latter otherwise, both valid by L9b. The first inequality is an equality only

when the A_τ are pairwise disjoint up to a μ_{Haar} null set. The copies need not be disjoint: for T6.19 they overlap, so (56) is loose there, the overlap costing sharpness while leaving subadditivity, and with it the bound, intact. The tightness of the factor $\prod_{k=1}^{K} \beta_k$ is not addressed here.

Instance. L9 prices the 14 `L2,L3` rows of Tab. 6 that the stand in form leaves above the escalation threshold, over 122 test cases in total; 6 of these rows fix their qubit layout and have $N = 1$. The weakest of the fourteen is T2.11 at $\log_{10} \varepsilon \approx -7.81$, more than five orders of magnitude below 10^{-2}; the strongest is below 10^{-500}, which is why the released artifact carries base ten logarithms and not floating point values. The conclusion does not rest on the layout flag: recomputing every row with the full orbit $n!$ leaves the weakest at $\log_{10} \varepsilon \approx -7.51$. T6.19 is priced in addition (two cases, $N = 120$), which makes 15 rows and 124 test cases. In every one of these rows the cases prepare distinct computational basis inputs, so the orthonormality hypothesis of L9b holds.

What L9 does not say. The bound prices the Haar null. It says that a uniformly random unitary almost never satisfies the conjunction; it says nothing about structured trivial submissions, which are a different threat model; the copy attack of §5 measures one such submission separately. A small ε here means "a random matrix does not pass", not "the task is hard".

C.12 Custom-Bound: T2.10 (Teleportation, 6-State Tomographic)

For task T2.10, a quantum teleportation test (Bennett et al., 1993) on $n_{\text{qubits}} = 3$ and $d = 8$ with the six-state tomographic test set $S = \{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |+i\rangle, |-i\rangle\}$ and per-state reduced density matrix fidelity threshold $F^\star = 0.95$, the conservative custom bound reads

$$\varepsilon_{\text{T2.10}} \;\le\; 1 - F^\star \;=\; 0.05. \tag{57}$$

Proof. The acceptance test is a conjunction of six per state reduced density matrix fidelity checks

$$F_{\text{state}}(\rho_{q_2}^{(|\psi_k\rangle)}, |\psi_k\rangle\langle\psi_k|) \ge F^\star \quad \text{for } |\psi_k\rangle \in S, \tag{58}$$

evaluated for a single adversarial circuit U across all six prepared inputs. Under the Haar null, $U \sim \mu_{\text{Haar}}$ on $U(8)$ is applied to the fixed prepared input $|\phi_k\rangle = |\psi_k\rangle_{q_0} \otimes |0\rangle_{q_1} \otimes |0\rangle_{q_2}$, so by left invariance of the Haar measure the output $|\Psi_k\rangle = U|\phi_k\rangle$ is a Haar random pure state on $\mathbb{C}^8$. Decompose $\mathbb{C}^8 = \mathbb{C}^4_{q_0 q_1} \otimes \mathbb{C}^2_{q_2}$ and let $\{|e\rangle\}_{e=1}^4$ be an orthonormal basis of the traced out factor $\mathbb{C}^4_{q_0 q_1}$. The per state fidelity is

$$F_k \;=\; \langle\psi_k|\rho_{q_2}|\psi_k\rangle \;=\; \sum_{e=1}^{4} \left|(\langle e| \otimes \langle\psi_k|)\,|\Psi_k\rangle\right|^2, \tag{59}$$

the aggregated squared overlap of a Haar random state with the four orthonormal vectors $|e\rangle \otimes |\psi_k\rangle$. A unitary change of basis that maps these four vectors to four computational basis vectors preserves the Haar measure, so by the Dirichlet representation of Haar basis probabilities (L3, flat $\text{Dirichlet}(1, \ldots, 1)$ on Δ^7) and the aggregation property of flat Dirichlet vectors (remark of §C.9), the mass on a fixed 4 dimensional subspace of $\mathbb{C}^8$ follows $F_k \sim \text{Beta}(4,\, 4)$. Hence the per state tail is exactly

$$\Pr[F_k \ge F^\star] \;=\; I_{1-F^\star}(4,4) \;=\; 140 \int_0^{0.05} u^3(1-u)^3\, du \;\approx\; 1.94 \times 10^{-4} \;\le\; 1 - F^\star = 0.05, \tag{60}$$

where $I_x(a, b)$ is the regularized incomplete Beta function. The acceptance event is the intersection of the six per state events for the same U, so $\varepsilon_{\text{T2.10}} \le \min_k \Pr[F_k \ge F^\star] = I_{0.05}(4,4) \approx 1.94{\times}10^{-4} \le 0.05$. The published conservative bound (57) therefore holds a fortiori.

The bound applies to the custom verification type, T2.10 row in the mapping table, with $d = 8$ and $F^\star = 0.95$. The tabulated $\varepsilon_{\text{theo}}$ is the conservative 0.05, so the a priori escalation classification of T2.10 is unchanged; the sharper value $I_{0.05}(4,4) \approx 1.94 \times 10^{-4}$ bounds the event for one prepared state and hence the six state conjunction, which is contained in it. The task fixes its qubit layout, so no relabeling union enters. Empirical Haar sampling delivers $\hat{\varepsilon} = 0$ on $N = 10^9$ samples, with Wilson 95% upper bound 3.84×10^{-9}, consistent with that bound.

C.13 CUSTOM-BOUND: T2.17 (SIMON 2-BIT, SIGNATURE-SPACE COLLISION)

For task T2.17, Simon's algorithm (Simon, 1997) with hidden generator $s = (1, 1)$ on $n_{\text{input}} = 2$ input qubits plus two oracle-output qubits ($d = 16$), whose shipped test requires marginal mass at least $\theta_T = 0.1$ on each valid string and at most 0.01 on each orthogonality violating string (θ_T denotes the acceptance threshold of the task at hand; it is a fidelity for the fidelity types, at the per task value of the $F^\star$ column of Tab. 6, and a marginal probability for the two Simon tasks), the signature space collision bound reads

$$\varepsilon_{\text{T2.17}} \;\le\; 2^{-n_{\text{input}}} \;=\; 2^{-2} \;=\; 0.25. \tag{61}$$

Proof. The runner accepts task T2.17 only if the input register marginal of each orthogonality violating string carries mass at most 0.01; the orthogonality violating strings are $\{01, 10\}$. Under the Haar null the output state is Haar random on $\mathbb{C}^{16}$, so its 16 basis probabilities are flat $\text{Dirichlet}(1, \ldots, 1)$ (§C.5). Each input register marginal q_y aggregates the 4 output coordinates that share its input label, so the violating mass $Q = q_{01} + q_{10}$ aggregates 8 of the 16 coordinates and $Q \sim \text{Beta}(8, 8)$ by the Dirichlet aggregation property (remark of §C.9). The task fixes its qubit layout, so no relabeling union enters.

Acceptance forces each violating runner marginal to at most 0.01, and the runner omits basis probabilities below 10^{-10}, at most 4 per string, so the true marginal of each violating string is at most $0.01 + 4 \times 10^{-10}$ and the true violating mass obeys $Q \le 2\,(0.01 + 4 \times 10^{-10}) = 0.02 + 8 \times 10^{-10}$. This bound uses only the violating strings and drops the constraint $q_y \ge \theta_T$ on the valid strings $\{00, 11\}$, so it is not tight. Since acceptance places Q in $[0,\, 0.02 + 8 \times 10^{-10}]$ and the drop shift 8×10^{-10} does not change the printed value of the Beta tail, the published bound (61) follows a fortiori from

$$\begin{aligned} \varepsilon_{\text{T2.17}} \;&\le\; \Pr\big[\text{Beta}(8, 8) \le 0.02 + 8 \times 10^{-10}\big] \\ &=\; I_{0.02+8\times10^{-10}}(8, 8) \;<\; 1.46 \times 10^{-10} \;\le\; 2^{-2} = 0.25. \end{aligned} \tag{62}$$

The bound applies to the custom verification type, T2.17 row in the mapping table, with $d = 16$ and $n_{\text{input}} = 2$. The tabulated $\varepsilon_{\text{theo}} = 2^{-2} = 0.25$, far above the 0.01 escalation threshold, is the conservative value, so T2.17 remains an a priori escalation candidate; the value $I_{0.02+8\times10^{-10}}(8, 8) \approx 1.45 \times 10^{-10}$ of the proof is the entry of the $\log_{10} \varepsilon_s$ column of Tab. 6.

C.14 CUSTOM-BOUND: T2.18 (SIMON 3-BIT, SIGNATURE-SPACE COLLISION)

For task T2.18, Simon's algorithm (Simon, 1997) with hidden generator $s = (1, 1, 0)$ on $n_{\text{input}} = 3$ input qubits ($d = 64$), with the same two thresholds ($\theta_T = 0.1$ on the valid strings, 0.01 on the violating ones), the signature space collision bound reads

$$\varepsilon_{\text{T2.18}} \;\le\; 2^{-n_{\text{input}}} \;=\; 2^{-3} \;=\; 0.125. \tag{63}$$

Proof. For task T2.18 the runner accepts only if the input register marginal of each string in $\{010, 011, 100, 101\}$, the orthogonality violating strings, carries mass at most 0.01. Under the Haar null the 64 basis probabilities of the output state are flat $\text{Dirichlet}(1, \ldots, 1)$ (§C.5); each input register marginal q_y aggregates the 8 output coordinates of its input label, so the violating mass $Q = q_{010} + q_{011} + q_{100} + q_{101}$ aggregates 32 of the 64 coordinates and $Q \sim \text{Beta}(32, 32)$ by the Dirichlet aggregation property (remark of §C.9). The task fixes its qubit layout, so no relabeling union enters. The runner omits basis probabilities below 10^{-10}, at most 8 per string, so the true violating mass obeys $Q \le 4\,(0.01 + 8 \times 10^{-10}) = 0.04 + 32 \times 10^{-10}$. The bound drops the constraint $q_y \ge \theta_T$ on the valid strings $\{000, 001, 110, 111\}$, so it is not tight. Since acceptance places Q in $[0,\, 0.04 + 32 \times 10^{-10}]$ and the drop shift 32×10^{-10} does not change the printed value of the Beta tail, the published bound (63) follows a fortiori from

$$\begin{aligned} \varepsilon_{\text{T2.18}} \;&\le\; \Pr\big[\text{Beta}(32, 32) \le 0.04 + 32 \times 10^{-10}\big] \\ &=\; I_{0.04+32\times10^{-10}}(32, 32) \;<\; 4.97 \times 10^{-28} \;\le\; 2^{-3} = 0.125. \end{aligned} \tag{64}$$

The empirical Haar rates 1.34×10^{-3} for T2.17 and 1.30×10^{-3} for T2.18 in Tab. 7 price a wider event than these two proofs do. The bounds (62) and (64) price the runner's rule, with its tolerance 0.01 on

the violating strings. The calibration rates come from a sampler that applies the threshold $\theta_T = 0.1$ to both kinds of string, rejecting a run when any violating string reaches 0.1 or any valid string falls below 0.1; that acceptance event contains the runner's, so it is priced by the same argument with 0.1 in place of 0.01: for T2.17 the two violating strings give $Q \leq 0.2$ and $\varepsilon_{\text{T2.17}} \leq I_{0.2}(8, 8) \approx 4.24 \times 10^{-3}$. The rate 1.34×10^{-3} is consistent with this superset bound, and 1.30×10^{-3} with the corresponding bound for T2.18, where the four violating strings give $Q \leq 0.4$ and $I_{0.4}(32, 32) \approx 5.37 \times 10^{-2}$; neither speaks to the runner bounds 1.45×10^{-10} and 4.96×10^{-28}, which price the tolerance 0.01.

The two bounds (62) and (64) price the event that the shipped test decides on the violating strings, where it is stricter than the model; for these two rows the $\log_{10} \varepsilon_{\text{s}}$ column of Tab. 6 reports that event, as its caption states row by row for every instrument.

The bound applies to the custom verification type, T2.18 row in the mapping table, with $d = 64$ and $n_{\text{input}} = 3$. The tabulated $\varepsilon_{\text{theo}} = 2^{-3} = 0.125$, above the 0.01 escalation threshold, is the conservative value, so T2.18 remains an a priori escalation candidate; the value $I_{0.04+32\times10^{-10}}(32, 32) \approx 4.96 \times 10^{-28}$ of the proof is the entry of the $\log_{10} \varepsilon_{\text{s}}$ column of Tab. 6. The empirical rate of Tab. 7 is a pooled estimate over a six shard run and is discussed with that of T2.17 above.

D Empirical Haar-Sampling Calibration

The calibration reports, per task, how often a Haar random reference input satisfies a sampled predicate; it is a statement about a random baseline, not about the models evaluated in this paper.

For each of the 86 tasks a sampled predicate is evaluated on random reference draws, Haar random for every task but T6.6, whose sampler flips fair bits instead, and its empirical hit rate $\hat{\varepsilon}$ is reported with its Wilson 95% upper bound, or with a Clopper–Pearson bound for T2.12, T2.22 and T6.15; what the predicate is, and how far it is from the shipped test, is stated under *Scope of the Calibration* below. The sample size N follows the escalation split and, below it, a band rule on the closed form bound: the 34 tasks whose a priori bound exceeds 0.01 use $N \geq 10^9$, while the bound certified regular tasks are spot checked at $N = 10^9$ when $\varepsilon_{\text{theo}} < 10^{-5}$, at $N = 10^7$ when it lies between 10^{-5} and 10^{-3}, and at $N = 10^6$ above. The run scripts of 48 of these tasks set N by this rule; those of T2.12, T2.22 and T6.15 set $N = 10^9$ for their L7 Monte Carlo rows, and T6.19 keeps the $N = 10^9$ of the escalation pass in which it was sampled under its earlier $K{=}1$ bound. The per task N is listed in Tab. 7. The choice of N is not explained by the expected hit count $N\varepsilon_{\text{theo}}$: 67 of the 86 rows return zero hits, among them 14 rows whose expected count $N\varepsilon_{\text{theo}}$ lies between 248 and 7,900 (the largest is T5.7 at 7867.13) and a further 29 rows where it lies between 4.03×10^6 and 2.5×10^8. At an expected count of at least 248 the probability of observing no hit at all is at most e^{-248}, so these rows are not a sampling fluctuation. They show that $\varepsilon_{\text{theo}}$ is not the hit probability of the sampled predicate, and not for one reason. On the 21 functional rows among the 43 the reason is the event mismatch described under *Scope of the Calibration* below: the sampled event is not the event that $\varepsilon_{\text{theo}}$ prices, and $\varepsilon_{\text{theo}}$ is by a wide margin the larger of the two. The other 22 of the 43, the L5 rows, T6.8 and T2.10, are not functional; for them the zero counts show only that the bound is loose for the sampled predicate. The main sampler is a single threaded per sample Python loop (the re-runs of T2.22 and T6.15 draw in vectorized blocks, and T2.18 ran in six shards); the wall clock time of the original run is not reproducible from the released sampler and is not stated. The pass flag is set when the upper bound stays below the 0.01 QuARC threshold. The T6.15 row samples the test of the dataset version before v1.2.0 (see the discussion of T6.15 below). In the consistency column the label low rate marks the rows that the calibration artifact labels as low rate, and T2.10, whose plain ok label rests on zero hits; other rows with ≤ 5 hits keep the label the artifact gives them.

Sampling Methodology

For each task T with full-space dimension $d = 2^{n_{\text{qubits}}}$:

1. *Reference draw.* The main sampler (`haar_sampler.py`) uses two primitives. Wherever a Haar random pure state $|\psi_i\rangle \in \mathbb{C}^d$ is required, it is drawn directly as a standard complex Gaussian vector in $\mathbb{C}^d$ normalized to unit length. This is the uniform measure on the unit sphere S^{2d-1} and is equal in law to a column of a $d \times d$ Haar random unitary, but no unitary is formed. Wherever a full Haar random unitary is required, that is for the unitary fidelity, parametric unitary fidelity and code space unitary fidelity families, it is constructed

by QR decomposing a complex Ginibre matrix and phase aligning the diagonal of R to be positive Mezzadri (2006). For parametric tasks the sampler evaluates the predicate of the corresponding non parametric family; it draws no parameter vector. The task specific samplers of T2.17 and T2.18 instead take the first column of a Haar unitary constructed the same way, which has the same law, and the sampler of T2.10 uses a full Haar unitary.

2. *Acceptance evaluation.* The Haar draw is tested against an analytic predicate derived from the task specification, for example $|\langle 0|\psi\rangle|^2 \geq F^\star$ for state fidelity, $|\mathrm{tr}\, U|^2/d^2 \geq F^\star$ for unitary fidelity, or the mass of a uniformly drawn set of k_s basis indices for the functional family, with k_s the subset size the run script passes, which can differ from the K of Tab. 6 (see below). Neither the canonical solution nor the shipped runner is executed, so the predicate is in general not the runner's test (*Scope of the Calibration* below); a hit, $\mathbf{1}_{x\in S(T)}$ for the set $S(T)$ of draws that the predicate accepts, is recorded.

3. *Estimator.* The empirical hit rate of the sampled predicate is $\hat{\varepsilon} = (\sum_{i=1}^{N} \mathbf{1}_{x_i\in S(T)})/N$ for the per task sample size N (Tab. 7). The Wilson 95% upper bound is

$$\mathrm{CI}_{\mathrm{upper}} \;=\; \frac{\hat{\varepsilon} + z^2/(2N) + z\sqrt{\hat{\varepsilon}(1-\hat{\varepsilon})/N + z^2/(4N^2)}}{1 + z^2/N} \tag{65}$$

with z the exact standard normal 0.975 quantile 1.959963984540054, which is what every generator in the release computes with. The rounded 1.96 leaves every bound printed to three significant figures unchanged, but it disagrees in the last digit of four of the six six figure lower bounds in App. E, and at $8/9$ it returns 56.4994 against 56.5000, so it can move a value rounded to an integer.

4. *Pass criterion.* The task is flagged as claim met iff $\mathrm{CI}_{\mathrm{upper}} \leq 0.01$. For tasks with $\hat{\varepsilon} = 0$ the Wilson upper bound reduces to

$$\hat{\varepsilon} = 0 \implies \mathrm{CI}_{\mathrm{upper}} = \frac{z^2}{N + z^2} \approx 3.84\times 10^{-9}\ (N{=}10^9), \quad 3.84\times 10^{-6}\ (N{=}10^6), \tag{66}$$

with z as above: the low rate regime; the per task value follows the task's N in Tab. 7.

5. *Determinism.* Every row of `haar_calibration.csv` records a seed and has a sampler version field. The default seed is $s_T = (\mathrm{tier}\cdot 10^4 + \mathrm{sub})\cdot 10^3 + 42$; 8 rows record a different seed, and 26 rows leave the version field empty.

Scope of the Calibration

The calibration is a proxy measurement, and this subsection states precisely what it does and does not certify. It is added here because the properties recorded below are not visible from Tab. 7 alone.

The sampled predicate is not the shipped test function. For every one of the 86 rows the calibration sampler evaluates a predicate of its own, not the shipped runner: an analytic predicate derived from the task specification, and for T6.6 not even that, since its sampler draws no state at all (see below). The two acceptance regions are close only in special cases: the runner adds conditions the analytic predicates do not carry, so the two accepting sets are different events and their raw agreement is dominated by shared rejection. On the six functional rows for which the comparison has been carried out (T2.1, T2.2, T2.3, T2.11, T2.19 and T6.1, released as `runner_event_bounds.csv`), the published $\varepsilon_{\mathrm{theo}}$ sits between a factor 1.0×10^3 and a factor 2.1×10^{36} *above* an upper bound on the probability of the event the runner actually tests, computed from the Dirichlet law of the output distribution under the same Haar null; the assumption light fallback in the same file, which keeps only the $\mathrm{Beta}(1, d-1)$ marginal of one peak coordinate, without disjointness and without the argmax reduction, still leaves a factor of at least 40.8. The remaining functional rows have not been checked this way. The Lemma inequalities are therefore not violated anywhere; what is wrong is the reading of ε_T as "the false positive rate of the shipped runner". The correct reading is that ε_T is an analytically derived bound on a model acceptance event, and that the calibration samples a predicate of its own under the same Haar null. No result of this paper is withdrawn by that distinction. The sharpened entries of the $\log_{10}\varepsilon_\mathrm{s}$ column of Tab. 6 price, row by row, the event its caption names: for the L9 rows the conjunction over the test cases under the tolerances of the shipped runner, for T2.17 and T2.18 the rule the shipped test enforces, for T6.6 the stabilizer condition at the identity qubit

layout, and for the L5 rows the union over the admitted qubit relabelings. A direct empirical false positive rate of the shipped runner is not among the quantities this benchmark reports.

Two mechanisms account for most of the gap. First, for the functional Case 3 family the runner additionally requires the candidate output distribution to stay within total variation distance 0.05 of the canonical solution's distribution, a gate that $\varepsilon_{\text{theo}}$ does not price in and that shrinks the acceptance region by orders of magnitude; the chain bound of §C.11 prices it (rule 3 of L9c), and its values stand in the $\log_{10}\varepsilon_{\text{s}}$ column of Tab. 6. Second, five rows carry an L2, L3 argmax assignment in Tab. 6 that does not describe the implemented test at all: every case of T2.9, T2.21, T2.24 and T6.18 is a distribution band check, accepting when each prescribed probability lies within 0.05 of its target, and every case of T6.10 an ancilla marginal band check, none of them with an argmax. For T2.9, T2.21 and T2.24 the L9 value in the $\log_{10}\varepsilon_{\text{s}}$ column of Tab. 6 prices these band checks (rule 2 of L9c). All five cases of T6.10 are ancilla marginal band checks, accepting when the ancilla outcome probability lies within 0.1 of its expected value, with no argmax involved; under the Haar null that ancilla marginal is $\mathrm{Beta}(16,16)$ and the per case acceptance probability is 6.85×10^{-9}, far below the tabulated 2.48×10^{-5}. All four cases of T6.18 are single component distribution band checks, accepting when the probability of one prescribed basis index is within 0.05 of 1, again with no argmax; the marginal of Eq. (9) at $d=8$ is $\mathrm{Beta}(1,7)$, so the per case acceptance probability is $(1-0.95)^7\approx 7.81\times10^{-10}$ against the tabulated 5.71×10^{-2}. In the rows of T6.10 and T6.18 the tabulated $\varepsilon_{\text{theo}}$ is retained as a conservative placeholder. For T6.18 the L9 value in the $\log_{10}\varepsilon_{\text{s}}$ column of Tab. 6 prices the implemented band checks of all four cases (rule 2 of L9c); T6.10 does not escalate, and no sharper bound is tabulated for it.

22 **rows certify an event of probability exactly zero.** For the functional rows configured with $F^\star=1.0$ and a sampled subset size $k_{\text{s}}<d$ the sampled predicate asks that k_{s} basis amplitudes, a set drawn afresh for each sample, carry the entire probability mass. Here k_{s} is the subset size the run script passes to the sampler; among these rows it equals the K of Tab. 6 except on T2.1, T2.3, T2.9, T2.11, T2.19, T2.21, T2.24, T2.26 and T6.16, where it is one, or two for T2.21 and T2.24. That set has Haar measure exactly zero, so the zero hit result is a property of the predicate and not evidence about the task. The rows concerned are T2.1, T2.3, T2.9, T2.11, T2.19, T2.21, T2.24, T2.26, T5.2, T5.5, T5.6, T5.7, T5.10, T5.12, T5.13, T5.14, T6.1, T6.9, T6.10, T6.12, T6.13 and T6.16, 22 in total. They appear in Tab. 7 with claim met set, and they are counted in the headline calibration result, but they carry no empirical content: an empty measurement cannot fail. Readers should discount them accordingly. The remaining rows are unaffected.

What the empty rows cost depends on the bound. 12 of the 22, the eight T5 rows together with T6.9, T6.10, T6.12 and T6.13, carry a closed form bound below 0.01 and are certified by that bound alone; the void calibration withdraws a confirmation they did not need. The other 10 — T2.1, T2.3, T2.9, T2.11, T2.19, T2.21, T2.24, T2.26, T6.1 and T6.16 — are escalation rows. Their bounds run from 0.03125 to 0.25, every one above the threshold, which is why they were sent to calibration at all, and for them the empty measurement is the whole of the $\varepsilon\le0.01$ entry. 5 of those 10 — T2.9, T2.21, T2.24, T2.26 and T6.16 — entered this set with dataset v1.1.0, which turned them from single case state fidelity into functional tasks; their calibration rows were re-run with the functional sampler after that change. T2.2 stands with those 10 without being a null set. Its bound prices the argmax event at 0.0625, which at $N{=}10^9$ would have produced some 6×10^7 hits; the run recorded none, so the sampled predicate was not that event but the $F^\star=0.99$ mass condition, of Haar mass $(1-0.99)^{15}=10^{-30}$ and expectation 10^{-21} hits. Its zero was as fixed in advance as theirs. 11 rows of Tab. 7 therefore report a verdict that no draw could have overturned, on tasks whose bound does not reach the threshold on its own.

The direction recorded above concerns the deployed acceptance event, while the $\varepsilon\le0.01$ screen is stated on the model event, and a range of factors across rows decides nothing row by row. For these six a bound on the deployed event is available. Its definition was read off the shipped runner for one test case of each row (input 0; both cases for T6.1), the qubit relabeling retry included even for the rows whose layout is fixed, which can only enlarge the bound, and its Haar mass evaluated in closed form under the null model that the candidate is a Haar random unitary on exactly n qubits: 2.93×10^{-18} for T2.1, 9.77×10^{-19} for T2.2, 2.93×10^{-18} for T2.3, 2.50×10^{-4} for T2.11, 1.49×10^{-38} for T2.19 and 2.02×10^{-17} for T6.1. The case it models is still among the cases dataset v1.2.0 prescribes, with the same canonical output, and the runner requires every case to

pass, so each value bounds the deployed event from above; it is not its exact probability. Each was confirmed independently by Monte Carlo, each lies strictly below the $\varepsilon_{\text{theo}}$ published for its row, so no Lemma inequality is affected, and all six sit under the threshold, the largest of them, T2.11, by a factor 40. The chain bound of §C.11, which prices all cases of dataset v1.2.0, is smaller still on five of the six rows (Tab. 6). An assumption light fallback, a union over the relabelled events using one necessary condition per event, without disjointness and without the argmax reduction, keeps all six under 0.01 as well, the loosest at 1.53×10^{-3}. The derivation and its agreement checks against the runner were computed against an earlier dataset version; they are released, with their run log, under `artifacts/lemma_bounds/runner_event_bounds/`; a change to the runner's decision layer would require re-deriving them.

Three limits on that. It is analytic, not measured: the calibration draw stays empty, and nothing above was obtained from a model output. It does not restate $\varepsilon_{\text{theo}}$, which remains 0.0625, 0.0625, 0.0625, 0.25, 0.03125 and 0.07143 on these rows, and it is not a false positive rate of the shipped runner: it prices one acceptance event under one null model, and a failed circuit extraction, a candidate on a different number of qubits, a non unitary circuit and the hardware pre check all sit outside it. Nor does it explain the empty draw. A mass of 2.50×10^{-4} would have put some 2.5×10^5 hits into T2.11 at $N{=}10^9$; the zero came from the sampled predicate, not from the event the runner tests. What the six rows carry is a sub threshold bound on the deployed event, and still not a calibrated rate below 0.01. A sampler that draws the argmax event the published bound prices would settle that; that run is not part of this release.

Seven rows validate the bound formula against itself. The rows T3.5, T3.15, T5.3, T5.4, T5.8 and T5.11 were re-run with a dedicated permutation sampler after a defect in the first sampler version. That replacement draws a Haar random pure state, sorts its d computational basis probabilities in descending order, and records a hit when the resulting ranking is the identity permutation. The basis probabilities of a Haar state are exchangeable, so their ranking is uniform on S_d and the hit probability is exactly $1/d!$ by construction, which is also the closed form assigned to these rows in Tab. 6. T6.6 is the seventh row of this kind, and the plainest one: its sampler draws $n_{\text{stab}} = 4$ fair bits and records a hit on the all zero string, so it is Bernoulli(2^{-4}) by construction, which is the closed form of Eq. (29). No quantum state enters it, and neither does $d = 32$, the threshold $F^\star = 0.99$, or the $[\![5, 1, 3]\!]$ code itself; the per task run script passes d and $F^\star$ on the command line, and the sampling routine accepts neither argument. The agreement between $\hat{\varepsilon}$ and $\varepsilon_{\text{theo}}$ on these seven rows is therefore a tautology and carries no information about the tasks; it confirms only that the sampler implements the formula. The closed form bounds themselves are untouched by this, and they are derived independently in App. C §C.6 and §C.8.

The five sigma consistency rule is not in force for most of the table. The main sampler declares a consistency criterion $|\hat{\varepsilon} - \varepsilon_{\text{theo}}| \leq 5\hat{\sigma}$ for its Wilson rows and flags rows that violate it. Of the 86 rows, 72 fail that criterion, in most cases because $\hat{\varepsilon} = 0$ makes $\hat{\sigma} = 0$ and the comparison undefined; 31 of them carry a `flag`, and the remaining 41 carry an `ok` prefixed label instead. For 38 of these 41 the label (`ok_low_rate`, `ok_zero_conservative`, and for T6.15 an event identification label) is not produced by the rule of any released sampler, so it replaces the declared rule rather than applying it. The other three, T2.10, T2.17 and T2.18, come from custom samplers that label a row consistent when $\hat{\varepsilon}$ does not exceed $\varepsilon_{\text{theo}}$ (for T2.17 and T2.18 plus a confidence half width). T2.17 and T2.18 exceed the five sigma tolerance by more than four orders of magnitude and are carried as consistent under the one sided rule of their own samplers. Their deviations reflect that $\varepsilon_{\text{theo}}$ is not the hit probability of the sampled predicate, not a numerical error.

Notation. $\varepsilon_{\text{theo}}$ denotes the analytically derived bound of App. B throughout, and it heads that column in Tab. 6, in Tab. 7 and in Fig. 3 alike, with one exception: for T2.12 and T6.15, Tab. 7 carries the bound of the earlier event that their calibration sampled, before dataset v1.2.0 narrowed both tests. It is not an estimate and carries no hat. The hat is reserved throughout this paper for empirical quantities, as in $\hat{\varepsilon}$ and $\hat{\varepsilon}_{\text{MC}}$. Numbers printed in this appendix and in Tab. 7 are rounded to the nearest value at the digits shown; a bound stated with at most or at least is rounded in its safe direction instead. The exact values follow from the released artifacts, for T6.4 and T6.5 from the closed form of the lemma mapping.

DISCREPANCIES

T2.22 (Grover, $K{=}2$). The apparent discrepancy between the theoretical value and the empirical estimate for T2.22 is an event identification issue; no bound revision is required. The runner accepts T2.22 only if *each* of the two marked outcomes carries probability at least 0.45 *and* their combined mass reaches the total peak probability of 0.9 that the dataset declares. Under the flat Dirichlet null this conjunction has the exact probability $(1 - 0.45 - 0.45)^7 = 10^{-7}$ (remark in App. C §C.9), the value carried as $\varepsilon_{\text{theo}}$ in Tab. 7. The first functional sampler, however, measured a *proxy* event: it drew a uniformly random two element subset of the $d = 8$ outcomes and counted a hit when the aggregate mass of the subset reached $F^\star = 0.9$, with no per outcome thresholds. That aggregate event is a strict superset of the runner's conjunction event, and its probability is also exact under the null: by Dirichlet aggregation the sum of $K = 2$ flat Dirichlet coordinates on Δ^7 follows $\mathrm{Beta}(2,\,6)$ with density $42\,x(1-x)^5$, so

$$\Pr[\mathrm{Beta}(2,6) \geq 0.9] \;=\; 42\int_0^{0.1}(1-u)\,u^5\,du \;=\; 42\left(\frac{0.1^6}{6} - \frac{0.1^7}{7}\right) \;=\; 6.40\times 10^{-6} \qquad (67)$$

exactly. The original measurement, $\hat{\varepsilon} = 6.37{\times}10^{-6}$ ($N{=}10^9$, $\hat{\sigma} = 7.98{\times}10^{-8}$, $\mathrm{CI}_{\text{upper}} \approx 6.5{\times}10^{-6}$, interval containing $6.40{\times}10^{-6}$), is therefore consistent with the sampler's own null. That run predates the released artifact tree and its log is not archived; the repeat below, which is, measures a proxy event of the same law at 6.45×10^{-6} and carries the statement on its own. Because the sampled proxy event contains the runner's acceptance event under the same null, its measured rate is also an empirical upper estimate for the probability of that event under the Haar null; it is not a measurement of the shipped runner on candidate circuits. A direct re-calibration with the conjunction predicate itself (sampler `v2_conjunction`, the same seed and a different sampling path, $N{=}10^9$) measures the conjunction event at 108 hits, $\hat{\varepsilon} = 1.08\times10^{-7}$ with Clopper–Pearson 95% upper bound 1.30×10^{-7}, consistent with the exact value 10^{-7}; the same run re-measures the proxy event at 6.45×10^{-6}, consistent with 6.40×10^{-6} on that different sampling path. Tab. 7 reports the direct measurement.

T6.15 (Shor's algorithm, $N{=}15$). The calibration row of T6.15 was run before the test was narrowed in dataset v1.2.0. Its sampler (`v2_haar_marginal`, seed convention unchanged, $N{=}10^9$) draws genuine Haar states on $\mathbb{C}^{128}$ and applies the earlier test: four peaks of the marginal on the three counting qubits of least significance, at least 0.15 each, and total peak mass at least 0.9. It records $46{,}783$ hits on the four peak event without the total mass condition ($\hat{\varepsilon} = 4.68\times10^{-5}$) and zero hits on the full earlier event. The shipped test of dataset v1.2.0 asks for sixteen peaks of at least 0.0375 each on the marginal over all seven qubits, together with total peak mass at least 0.9. The sixteen indices fall four each onto the four earlier peaks, so every state the shipped test accepts also passes the earlier one: the sampled event contains the shipped event, and the zero hit result is consistent with both. In Tab. 6 the $\varepsilon_{\text{theo}}$ of T6.15 is the closed form of the shipped event, 1.75×10^{-101} (remark in App. C §C.9); its calibration row in Tab. 7 carries the bound 1.76×10^{-30} of the earlier event, and no Monte Carlo number enters either.

Table 7: Empirical Haar sampling calibration. Values are rounded to the nearest value at the digits shown, a bound stated with at most in its safe direction; the exact values follow from the released artifacts, for T6.4 and T6.5 from the closed form of the lemma mapping. Haar random reference inputs are drawn per task (T6.6 excepted, see below) at the sample size N shown—escalation tasks (a priori bound above 0.01) at $N{=}10^9$, bound-certified regular tasks spot-checked at $N{=}10^6$ to 10^9—and the empirical hit rate $\hat{\varepsilon}$ of the sampled predicate is reported with its Wilson 95% upper bound, except for T2.12, T2.22 and T6.15, which the calibration artifact records with a Clopper–Pearson upper bound; that is why T6.15 shows 3.69×10^{-9} at zero hits and $N{=}10^9$ where the Wilson rows show 3.84×10^{-9}. The claim met column is true when the upper bound stays below the 0.01 QuARC threshold. In the consistency column the label low rate marks the rows that the calibration artifact labels as low rate, and T2.10, whose plain ok label rests on zero hits; other rows with ≤ 5 hits keep the label the artifact gives them. Seven rows came out with $\hat{\varepsilon}$ above $\varepsilon_{\text{theo}}$; a dagger on that label marks the six of them whose label is a bare ok, T2.22 carrying its own label instead and treated separately below. There $\varepsilon_{\text{theo}}$ is the exact hit probability of the sampled predicate and not a ceiling on it—the Fubini–Study volume (4) for T1.2, T1.10 and T4.8, a sampler that implements the closed form by construction for T5.3, T5.4 and T6.6 (App. D)—so a rate above it is scatter, not a violated bound; the size of the T4.8 excess is a separate question, treated below. On T6.6 the excess does not show in the printed cells, which both read 6.25×10^{-2}: it is 62,512,070 hits at $N{=}10^9$ against the 62,500,000 that the exact $\varepsilon_{\text{theo}}$ predicts. That excess is 1.58 standard deviations in the artifact's own $\hat{\sigma}$ convention, one sided $p = 0.06$ under the normal approximation, so it is scatter of the size an equality predicts and not a signal; the No in the claim met column records that the Wilson upper bound lies above the 0.01 threshold—which the a priori bound $2^{-n_{\text{stab}}} = 0.0625$ announces before any sample is drawn—and not a bound that the sampling exceeded. That a priori bound belongs to the sampled surrogate: for the acceptance event the shipped test decides, the corresponding bound is 4.27×10^{-29} at the identity qubit layout, and its union over the 120 qubit relabelings is at most 5.2×10^{-27}, against the tabulated 0.0625, so the one row that misses the threshold misses it under the sampled model and not under the released test. What the table certifies needs stating precisely. The sampler applies analytic predicates to Haar draws and never calls the shipped test function, so the set it accepts and the set the runner accepts are different events. How far the two diverge varies by row: of the 34 escalation rows the sampler draws the quantity the shipped test decides on for 19 (for T2.17, T2.18, T4.4, T4.6 and the L5 rows without a fixed layout under a rule other than the test's, App. C), draws a Haar state but evaluates a different event for 14, all of them functional, and for one, T6.6, draws no random state at all—it flips n_{stab} fair bits whose expectation is the very formula $2^{-n_{\text{stab}}}$ that the row is checking. Five of the fourteen (T2.9, T2.21, T2.24, T2.26, T6.16) moved from state fidelity to functional with dataset v1.1.0; their rows were re-run with the functional sampler after that change. On the six functional rows released as `runner_event_bounds.csv` (among them T2.2, 9.77×10^{-19} against the published 6.25×10^{-2}) the upper bound on the deployed acceptance event lies orders of magnitude below the published bound, so the lemma inequalities are not violated on these rows; for the one parametric fidelity row checked, T4.6, an empirical confidence bound of 7.68×10^{-9} against 6.31×10^{-2} points the same way, but it is not a proof (App. C). The direction is not established on every row: for T1.17, for example, the deployed event is the union of the relabeled copies of the published event (App. C), so its probability is not smaller than the published value. This table reports, under a Haar null model, how often the sampled predicate of each row is met; it does not confirm the bounds row by row and is not a measurement of the shipped test function's false positive rate. T2.10, T2.17, T2.18 use Custom Bound samplers (App. C), and T2.18 is the pooled estimator over a six shard run. The T2.22 row reports the direct re-calibration with the harness conjunction predicate (108 hits at $N{=}10^9$, consistent with the exact value 10^{-7}); the original run had measured an aggregate proxy event (exact value 6.40×10^{-6}, measured 6.37×10^{-6}, re-confirmed at 6.45×10^{-6} with the same seed on a different sampling path). The T2.12 and T6.15 rows were sampled before dataset v1.2.0 narrowed both tests, and their $\varepsilon_{\text{theo}}$ cell is the bound of the earlier event the row samples (10^{-7} and 1.76×10^{-30}); the shipped tests accept a subset of those events, and their closed forms, 10^{-14} and 1.75×10^{-101}, stand in Tab. 6. For T6.15 genuine Haar sampling at $N{=}10^9$ measures the four peak event without the total mass condition at 4.68×10^{-5} and records zero hits on the full earlier event. Both rows appear in the calibration discrepancy record released with the artifacts (`calibration_discrepancies.md`) with their sample size, hit count and deviation; the reasoning behind the two resolutions is in App. D rather than in that file. T4.8 is the only row that deviates from an exact $\varepsilon_{\text{theo}}$ by more than three standard deviations (1106 hits at $N{=}10^9$ against the exact 10^{-6}: $z = 3.19$ in the artifact's own $\hat{\sigma}$ convention, exact binomial $p = 5.1 \times 10^{-4}$ one sided and $p = 9.5 \times 10^{-4}$ two sided; one sided is the question asked here, since only an excess over $\varepsilon_{\text{theo}}$ would matter). Nineteen rows record a non-zero hit count and therefore admit such a test at all; across that family the Bonferroni corrected value is $p = 0.0097$ one sided and $p = 0.018$ two sided, so the excess stays significant at the 1% level under the one sided reading and at the 5% level under the two sided one, and is not explained away as sampling noise. The row is retained because what this table certifies does not depend on it: at 1106 hits the Wilson upper bound is 1.17×10^{-6}, a factor 8524 below the 0.01 threshold, so the claim holds whatever the exact value of $\varepsilon_{\text{theo}}$ is. The excess itself remains unexplained. The T6.19 row was calibrated a priori under $K{=}1$; with the corrected $K{=}2$ its bound is 4.03×10^{-3} and it leaves the escalation set. The $K{=}2$ event has since been re-measured under the same protocol ($N{=}10^9$, seed 60019042, a re-run sampler version recorded in the artifact, log released under `artifacts/haar_calibration/logs/`); it again yields zero hits, so $\hat{\varepsilon}$, the Wilson upper bound and the consistency label are unchanged.

Task	N	$\hat{\varepsilon}$	CI_{upper}	$\varepsilon_{\text{theo}}$	Claim met	Consistency
T1.2	10^9	$1.03e{-6}$	$1.09e{-6}$	$1.00e{-6}$	Yes	ok†
T1.5	10^9	0	$3.84e{-9}$	$1.00e-14$	Yes	low rate
T1.8	10^9	0	$3.84e{-9}$	$1.00e-14$	Yes	low rate
T1.10	10^9	$1.03e{-6}$	$1.10e{-6}$	$1.00e{-6}$	Yes	ok†
T1.11	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T1.12	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T1.14	10^9	0	$3.84e{-9}$	$1.00e-62$	Yes	low rate
T1.15	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T1.17	10^9	0	$3.84e{-9}$	$1.00e-62$	Yes	low rate
T2.1	10^9	0	$3.84e{-9}$	$6.25e{-2}$	Yes	flag
T2.2	10^9	0	$3.84e{-9}$	$6.25e{-2}$	Yes	flag
T2.3	10^9	0	$3.84e{-9}$	$6.25e{-2}$	Yes	flag
T2.6	10^9	$9.76e{-7}$	$1.04e{-6}$	$1.00e{-6}$	Yes	ok
T2.7	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T2.9	10^9	0	$3.84e{-9}$	$1.25e{-1}$	Yes	flag
T2.10	10^9	0	$3.84e{-9}$	$5.00e{-2}$	Yes	low rate
T2.11	10^9	0	$3.84e{-9}$	$2.50e{-1}$	Yes	flag
T2.12	10^9	$9.50e{-8}$	$1.16e{-7}$	$1.00e{-7}$	Yes	ok, L7 within CI
T2.16	10^6	0	$3.84e{-6}$	$3.95e{-3}$	Yes	ok, zero

(continued on next page)

(continued)

Task	N	$\hat{\varepsilon}$	$\mathrm{CI}_{\mathrm{upper}}$	$\varepsilon_{\mathrm{theo}}$	Claim met	Consistency
T2.17	10^9	$1.34e{-3}$	$1.34e{-3}$	$2.50e{-1}$	Yes	ok
T2.18	10^9	$1.30e{-3}$	$1.30e{-3}$	$1.25e{-1}$	Yes	ok
T2.19	10^9	0	$3.84e{-9}$	$3.13e{-2}$	Yes	flag
T2.21	10^9	0	$3.84e{-9}$	$7.14e{-2}$	Yes	flag
T2.22	10^9	$1.08e{-7}$	$1.30e{-7}$	$1.00e{-7}$	Yes	direct re-run; v1 proxy $6.37e{-6}$
T2.23	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T2.24	10^9	0	$3.84e{-9}$	$7.14e{-2}$	Yes	flag
T2.25	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T2.26	10^9	0	$3.84e{-9}$	$6.25e{-2}$	Yes	flag
T3.2	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T3.3	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T3.5	10^7	$2.21e{-5}$	$2.52e{-5}$	$2.48e{-5}$	Yes	ok
T3.6	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T3.8	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T3.11	10^9	0	$3.84e{-9}$	$6.26e{-2}$	Yes	flag
T3.12	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T3.13	10^6	0	$3.84e{-6}$	$3.95e{-3}$	Yes	ok, zero
T3.14	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T3.15	10^7	$2.36e{-5}$	$2.68e{-5}$	$2.48e{-5}$	Yes	ok
T4.2	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T4.3	10^9	0	$3.84e{-9}$	$1.00e-14$	Yes	low rate
T4.4	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T4.5	10^9	0	$3.84e{-9}$	$1.00e-14$	Yes	low rate
T4.6	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T4.7	10^9	0	$3.84e{-9}$	$1.00e-14$	Yes	low rate
T4.8	10^9	$1.11e{-6}$	$1.17e{-6}$	$1.00e{-6}$	Yes	ok†
T4.9	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T4.10	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T4.11	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T4.12	10^6	0	$3.84e{-6}$	$3.95e{-3}$	Yes	ok, zero
T4.13	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T4.14	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T4.15	10^9	$9.60e{-7}$	$1.02e{-6}$	$1.00e{-6}$	Yes	ok
T4.16	10^6	0	$3.84e{-6}$	$3.95e{-3}$	Yes	ok, zero
T4.17	10^9	0	$3.84e{-9}$	$1.00e-30$	Yes	low rate
T5.1	10^9	$3.45e{-7}$	$3.83e{-7}$	$5.71e{-2}$	Yes	flag
T5.2	10^7	0	$3.84e{-7}$	$6.22e{-4}$	Yes	ok, zero
T5.3	10^7	$2.62e{-5}$	$2.96e{-5}$	$2.48e{-5}$	Yes	ok†
T5.4	10^7	$2.66e{-5}$	$3.00e{-5}$	$2.48e{-5}$	Yes	ok†
T5.5	10^9	0	$3.84e{-9}$	$3.28e-14$	Yes	low rate
T5.6	10^9	0	$3.84e{-9}$	$2.66e{-8}$	Yes	ok, zero
T5.7	10^7	0	$3.84e{-7}$	$7.87e{-4}$	Yes	ok, zero
T5.8	10^9	0	$3.84e{-9}$	$3.80e-36$	Yes	low rate
T5.9	10^9	$3.54e{-7}$	$3.93e{-7}$	$5.71e{-2}$	Yes	flag
T5.10	10^9	0	$3.84e{-9}$	$7.61e{-7}$	Yes	ok, zero
T5.11	10^7	$2.24e{-5}$	$2.55e{-5}$	$2.48e{-5}$	Yes	ok
T5.12	10^9	0	$3.84e{-9}$	$2.66e{-8}$	Yes	ok, zero
T5.13	10^9	0	$3.84e{-9}$	$1.59e-24$	Yes	low rate
T5.14	10^9	0	$3.84e{-9}$	$3.28e-14$	Yes	low rate
T6.1	10^9	0	$3.84e{-9}$	$7.14e{-2}$	Yes	flag
T6.2	10^7	0	$3.84e{-7}$	$1.11e{-4}$	Yes	ok, zero
T6.3	10^9	0	$3.84e{-9}$	$1.58e{-2}$	Yes	flag
T6.4	10^9	0	$3.84e{-9}$	$1.00e-1533$	Yes	low rate
T6.5	10^9	0	$3.84e{-9}$	$1.00e-381$	Yes	low rate
T6.6	10^9	$6.25e{-2}$	$6.25e{-2}$	$6.25e{-2}$	No	ok†
T6.8	10^9	0	$3.84e{-9}$	$6.31e{-2}$	Yes	flag
T6.9	10^7	0	$3.84e{-7}$	$2.48e{-5}$	Yes	ok, zero

(continued on next page)

(continued)

Task	N	$\hat{\varepsilon}$	$\mathrm{CI}_{\mathrm{upper}}$	$\varepsilon_{\mathrm{theo}}$	Claim met	Consistency
T6.10	10^7	0	$3.84e{-7}$	$2.48e{-5}$	Yes	ok, zero
T6.11	10^6	0	$3.84e{-6}$	$4.03e{-3}$	Yes	ok, zero
T6.12	10^9	0	$3.84e{-9}$	$6.30e{-6}$	Yes	ok, zero
T6.13	10^7	0	$3.84e{-7}$	$6.22e{-4}$	Yes	ok, zero
T6.14	10^7	$9.92e{-4}$	$1.01e{-3}$	$1.00e{-3}$	Yes	ok
T6.15	10^9	0	$3.69e{-9}$	$1.76e-30$	Yes	direct re-run; exact bound
T6.16	10^9	0	$3.84e{-9}$	$6.25e{-2}$	Yes	flag
T6.17	10^7	0	$3.84e{-7}$	$6.11e{-5}$	Yes	ok, zero
T6.18	10^9	$1.94e{-4}$	$1.95e{-4}$	$5.71e{-2}$	Yes	flag
T6.19	10^9	0	$3.84e{-9}$	$4.03e{-3}$	Yes	flag

E Mutation Calibration of the Acceptance Region

Corpus and operators. Every first order mutant of the 86 canonical solutions is generated under three operators; nothing is sampled. M1 deletes one gate, 778 mutants. M2 swaps one qubit of a multi-qubit gate with its last qubit, 389 mutants; the tables call it the control target swap after its effect on a controlled gate. M3 rescales one angle by the factor 0.95 or 1.05, 414 mutants. The 414 are 207 angle slots at both factors; the corpus's one angle slot of value zero is skipped, since rescaling it changes nothing. Where the rescaled angle is a number, the shift is 0.0169 to 0.1571 rad, median 0.0785. This gives 1,581 mutants, 0 rejected in pre-flight. A ground truth computed outside the runner assigns each mutant its fidelity F to the canonical solution under the best permitted relabeling. It calls a mutant equivalent (133) if its fidelity under the identity layout already reaches $1 - 10^{-9}$ and relabeling (14) if only a relabeling does. Every other mutant is within tolerance (342) if F reaches the fidelity threshold of its task and genuinely wrong (1,092) otherwise. On fidelity graded tasks the ground truth measures the same fidelity the runner computes, and on the 886 fidelity graded mutants that reach $A(T)$ the two agree to within 3.331×10^{-15}. On predicate graded tasks that threshold is $1 - 10^{-9}$ itself, so no predicate graded mutant is within tolerance. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below, to within, exceeds or an inequality sign, and each end of a stated span, is rounded in its safe direction instead, and an equality sign joins only exact values. Wilson lower bounds are therefore cut off at the digits shown.

The denominator. Of the 1,092 genuinely wrong mutants, 1,073 are kept: all 540 fidelity graded ones, whatever their output distribution, and the 533 predicate graded ones whose output distribution, under every permitted relabeling, differs from the canonical one by more than 10^{-12} on at least one of the inputs the ground truth evaluates. The other 19 are predicate graded and, on the inputs the ground truth evaluates and under a permitted relabeling, give the canonical output distribution to within 10^{-12}; they are not counted in the rates below. The runner accepts 11 of them, rejects 6 and stops 2 before $A(T)$; counting the 17 that reach $A(T)$ would give 959 of 1,006, a rate of 0.953280. Of the 1,073, 989 reach $A(T)$; a structural pre-check stops the other 84, all gate deletions, across 16 tasks. The region rejects 953 of the 989, a rate of 0.963600. The Wilson 95% lower bound is 0.950019. Apart from that comparison, every rejection rate in this appendix is over the 989 or over a stratum of them. Strata in Tab. 8.

What is still accepted. Within the denominator the region admits 36 mutants. All are predicate graded, all sit on functional tasks, and all are M3 rescalings of a numeric angle, 18 at each factor. They fall on three tasks, T6.16 (24), T2.9 (6) and T2.26 (6). Every fidelity graded mutant in the denominator is rejected, 471 of 471. The 518 predicate graded ones split into 482 rejected and the 36 accepted. By operator, gate deletion is rejected 658 of 658, the control target swap 275 of 275 and the angle rescaling 20 of 56. The released test for functional tasks has three kinds of test case, and none of them uses a fidelity threshold. Of the 329 test cases of the functional tasks, 300 require argmax agreement and a total variation distance of at most 0.05 to the canonical output. Another 24 require each listed output probability to match the dataset's distribution to within 0.05, and 5 require an ancilla probability to differ from the expected one by less than 0.1. The 24 accepted mutants on T6.16 pass cases of the first kind, the 6 on T2.9 cases of the second, and the 6 on T2.26 cases of both.

The genuinely wrong label instead asks for fidelity below $1 - 10^{-9}$, so the two criteria can disagree. Measured by the ground truth distance under the best permitted relabeling, the largest distance among the 36 accepted mutants is 0.0392. The smallest among the rejected functional mutants of the denominator is 0.146. Each of the 36 thus deviates from the canonical solution and still passes the tolerances its test declares; against the genuinely wrong label they are false acceptances.

What was not measured. Every mutant changes one gate, so this is no false positive rate against model produced circuits, whose errors change the approach. Most of the denominator lies far from the boundary of $A(T)$: the median F over the 989 is 0.2500. The 36 accepted mutants are the only ones in $[0.99, 1.00)$, and no other has F above 0.9809. The tolerance 0.05 of both case kinds lies in the gap between the largest accepted and the smallest rejected distance above, and no functional mutant of the denominator falls into it. The run therefore shows that the accepted mutants stay within the tolerance, but it does not probe where the tolerance takes effect. That it holds says nothing about whether 0.05 is the right tolerance. The deviation at which $A(T)$ begins to accept is not measured, since no amplitude sweep was run. The empty circuit is accepted on 0 of 86 tasks (control N1), and on 23 of them the structural pre-check already stops it, so there the control says nothing about the acceptance threshold. One run, one machine, one library version; repeatability is not claimed. Every interval treats mutants as independent, while they cluster by task: the 989 sit on 85 tasks, T1.15 alone carrying 62, and T4.7 none.

F Composite-Task Validation

The registry marks 7 of the 86 tasks as composite, each splitting into $k \geq 2$ sub components given as gate lists that match the canonical solution gate by gate. Each test compares the output states of the canonical solution's gate lists and of a variant on a single random input state; the comparison succeeds when they reach a fidelity of at least 0.99. The pipeline therefore tests the decomposition of the canonical solutions, not the runner and not model output. A test passes when its comparison ends as the test expects. The positive tests are the unchanged canonical solution (gold), k variants that add an X gate pair before each controlled gate of a single sub component, and a variant applying this across all components; each expects success. All of these leave the unitary unchanged. The candidate negative tests, called E tests, edit the sub component structure: k omissions of one sub component (E1), k flips of every CNOT direction within one sub component (E2), and $k - 1$ reorderings of adjacent sub components (E3); each expects failure. This gives $3k - 1$ candidates per task. An E2 candidate has no mutation when its sub component holds no CNOT, and an E3 candidate leaves the unitary unchanged when the reordered sub components commute. Such candidates are recorded as skipped, their pass flag set without a test and left out of every executed count.

Over the 7 tasks, 15 of the 59 candidates are skipped, 7 of them E2 and 8 E3, leaving 44 executed E tests and 94 executed tests in total. Every executed E test changes the unitary. The highest fidelity among them stays below 0.9256. A task passes when all its positive and executed E tests pass, and all 7 tasks pass. The commutativity flag in Tab. 9 marks a task commutative when every adjacent pair of its sub components commutes, that is, when all its E3 candidates are skipped. T6.19 is therefore marked non commutative, although 1 of its adjacent pairs commutes. The P tests are error injections: they add an X (P1) or a Z (P2) on qubit $42 \bmod n$, where n is the task's qubit count, after the second sub component. Every P test except P1 on T6.19 expects the comparison to fail, and all 14 comparisons fail; the P tests are not part of the task criterion. For T6.19, P1 expects the correction to restore the output, but the test adds a bit flip on q_2 to the flip on q_1 that the task itself injects. The comparison fails with the task's correction and with a full majority vote decoder alike. Two flips exceed what the three qubit bit flip code corrects, but the comparison fails even for a single flip that the full decoder corrects. The test starts from a random state on all 5 qubits of T6.19, so the decoder reads a random syndrome, and the comparison also covers the syndrome qubits. The failure therefore says nothing about the correction circuit of the task. The construction script and the sub component definitions of all 7 tasks are released with the benchmark, and the script simulates with numpy alone and uses fixed seeds. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below, to within, exceeds or an inequality sign, and each end of a stated span, is rounded in its safe direction instead, and an equality sign joins only exact values.

Table 8: Mutation calibration of the acceptance region. Cells of the form k/n are counts over n; the rate blocks count rejections, the controls acceptances. The 1,073 (D3) are the genuinely wrong mutants that are fidelity graded or whose output distribution differs from the canonical one by more than 10^{-12}; the other 19 of the 1,092 are predicate graded and, on the inputs the ground truth evaluates and under a permitted relabeling, give the canonical output distribution to within 10^{-12}. The 84 mutants between the 1,073 and the 989 never reach $A(T)$: a structural pre-check over 16 tasks stops them first. Wilson is the score interval, two sided 95%, lower bound, cut off at the digits shown; rates are rounded to the nearest value; "—" stands where no binomial proportion is involved, at the macro mean and at the four controls, which check the setup on all 86 tasks rather than estimate a rate, and throughout the per task block. Five qualifications do not show in the cells. †The fidelity row is close to tautological: the ground truth and the test statistic are the same quantity computed twice, agreeing to 3.331×10^{-15} over the 886 mutants of the 43 fidelity graded tasks that reach $A(T)$, 471 of them genuinely wrong, so such a mutant must fail, and only the 43 predicate graded tasks carry information. The corpus is skewed, T1.15 alone contributing 151 of the 1,581 mutants, which is why the per task block precedes the pooled rate. T4.7 is the one task without a denominator. The operators differ in reach: M1 deletes an entire gate where M3 rescales one angle, shifting it by 0.0169 to 0.1571 rad where that angle is a number, a span inside the acceptance radius at $\theta_T = 0.99$ (§3.3); on a fidelity graded task at that threshold all 142 such mutants pass, the lowest at $F \approx 0.9953$. 342 of the 414 are within tolerance and 56 reach the denominator: 20 free parameter rescalings on fidelity graded tasks, all rejected, the highest at $F \approx 0.9662$, and 36 rescalings of a numeric angle on predicate graded tasks, all accepted, the lowest at $F \approx 0.9938$ and the highest at $F \approx 0.9996$; so no fidelity graded mutant of that row comes near the boundary at 0.99. One run, one machine, one library version, recorded in `artifacts/mutation_calibration/` with a record of the dataset version it ran against, v1.2.0. The median F over the run's denominator is 0.2500; the 36 accepted mutants are the only ones in $[0.99, 1.00)$, and no other has F above 0.9809, so near $F = 1$ these rates rest on those 36 alone.

Stratum	Count / n	Rate	Wilson 95% lower
Corpus and denominator (mutants)			
First order mutants generated	1,581	—	—
Genuinely wrong (ground truth)	1,092	—	—
Fidelity graded or output differs (D3)	1,073	—	—
Reaching $A(T)$	989	—	—
Per task (85 tasks with a denominator)			
Tasks rejecting every mutant	82/85	—	—
Macro mean of task rates	—	0.984923	—
Rejection rate			
All	953/989	0.963600	0.950019
Fidelity graded†	471/471	1.000000	0.991910
Predicate graded	482/518	0.930502	0.905285
M1 gate deletion	658/658	1.000000	0.994195
M2 control target swap	275/275	1.000000	0.986223
M3 angle rescaling	20/56	0.357143	0.244565
Mandatory controls (all 86 tasks)			
P1 canonical solution accepted	86/86	—	—
P2 repeat, fidelity bit ident.	86/86	—	—
N1 empty circuit accepted	0/86	—	—
N2 identity gates only accepted	0/86	—	—

Table 9: Composite task validation summary. Each composite task is tested with $k+2$ positive tests (gold canonical, k single component variants, all component variant), $3k-1$ candidate negative E tests (E1: k sub component omissions, E2: k CNOT direction flips, E3: $k-1$ reorderings of adjacent sub components), plus two P tests (P1: bit flip error injection, P2: phase flip). The E-Neg. column gives executed over candidate tests; a candidate whose mutation is absent (E2: no CNOT in the sub component) or leaves the unitary unchanged (E3: the two sub components commute) is recorded as skipped and never executed, as described in the text of this appendix. Exec. and Passed count executed tests only. Pass criterion: every positive and executed E test passes, that is, its comparison succeeds for positive tests and fails for E tests; the P tests are not part of the criterion. The one failed test is P1 on T6.19, which adds a bit flip on q_2 to the flip on q_1 that the task itself injects; the text of this appendix explains why this failure says nothing about the correction circuit of the task.

Task	k	Pos.	E-Neg. (exec./cand.)	P-tests	Exec.	Passed	Comm.
T5.1	2	4	4/5	2	10	10/10	*non-commutative*
T5.2	3	5	7/8	2	14	14/14	*non-commutative*
T5.10	2	4	2/5	2	8	8/8	*commutative*
T5.13	3	5	7/8	2	14	14/14	*non-commutative*
T6.9	4	6	8/11	2	16	16/16	*commutative*
T6.11	4	6	8/11	2	16	16/16	*commutative*
T6.19	4	6	8/11	2	16	15/16	*non-commutative*
Total	—	36	44/59	14	94	93/94	—

G Multi-Source LLM-Output Validation

The runner's accept and reject decisions on the baseline's own circuits are compared with a separate check by three SDKs, the three SDK check. Each circuit is loaded from the generated Qiskit code and simulated three times: by Qiskit's statevector class, by Cirq after an OpenQASM 2 export, and by TKET after the `pytket-qiskit` converter. One shared evaluator turns each result into a verdict; for unitary tasks the Cirq and TKET verdicts both come from one numpy construction of the unitary from the Qiskit code. The three SDK check also contains a PennyLane path, which its consensus does not use because of an `autoray` import conflict in the test environment. Of the 3,870 baseline generations, 3,803 contain extracted code and 67 contain none. The 189 extraction failures the runner records are these 67 and 122 responses whose code yields no usable module level circuit; none of the 122 reaches a unanimous verdict. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below, to within, exceeds or an inequality sign, and each end of a stated span, is rounded in its safe direction instead, and an equality sign joins only exact values.

On 1,100 generations all three SDKs reach a pass or fail verdict, and on 1,073 of them (97.55%) the three agree. On the 15 unitary tasks among them, which carry 610 of the 1,073 unanimous verdicts, the Cirq and TKET verdicts are one computation, so there a unanimous verdict is the agreement of two computations, not three. The runner agrees with that unanimous verdict on 1,071 of the 1,073 (99.81%). Both disagreements are runner acceptances that all three SDKs reject, both on T6.6, the only stabilizer eigenvalue task. The runner rejects none of the 991 unanimous acceptances.

The unanimous verdicts divide into 991 acceptances and 82 rejections. A runner more permissive than the three SDK check can show only among the rejections; permissiveness that the two share stays invisible here. Of these 82 the runner accepts 2 (2.44%), with a 95% Wilson interval from 0.67% to 8.47%. Both of the 2 prepare a state that the four stated stabilizers fix only after a relabeling of the qubits; the runner applies such a relabeling, while the three SDK check relabels only on state and unitary tasks of at most 5 qubits, not on its stabilizer path. Whether a relabeled code state solves T6.6 is a question of the task definition, which this appendix does not settle. Over all 3,870 runner verdicts, the three seeds of one model on one task share an intraclass correlation of 0.75. The 82 fall into 48 clusters of one model and one task, so the interval, which treats them as independent, can be too narrow; this appendix does not model by how much.

The unanimous verdicts lie on 29 of the 86 tasks. For 20 of the other 57 the three SDK check has no comparison path by design. These 20 are the 12 parametric state fidelity tasks, the 4 parametric

unitary fidelity tasks, the 3 custom tasks and the 1 code space unitary fidelity task. The other 37 drop out by implementation. On 30 of the 35 functional tasks only the Qiskit path can start from a basis input, so the Cirq and TKET paths return no verdict. On the other 5 the check has no evaluation for the form of their test cases, and all three paths return an error. Two further tasks drop out because two of the three paths return an error on every generation. On T2.22 every generation measures its qubits, which the Qiskit and TKET paths reject; T6.17 acts on 7 qubits, beyond the size up to which the Cirq and TKET paths build a unitary. The agreement figures therefore cover none of the functional tasks. The per generation table ships in the release as `baseline/consensus_matrix.csv` (3,803 rows, one per generation with extracted code), with the consolidated write up beside it in `consensus_results.md`.

H STANDARD PROMPT TEMPLATE

A single minimal prompt template applies to all 86 tasks and all 15 models of the baseline. The prompt is a system instruction asking for Qiskit code that builds a module level variable named `qc`, followed by the task description and a one line output reminder as the user message. By design the prompt is spartan. It includes no chain of thought scaffolding, few shot examples, JSON schema, or tool descriptions. The template adds no hardware context and no task modes; a task description may itself state a hardware constraint, such as nearest neighbor connectivity.

Only the task description varies; the system instruction is fixed across tasks. That variation is not only one of wording: 10 of the 86 task descriptions carry a line forbidding a prebuilt state preparation. T1.2 is one of them, and the other 76 carry no such line. The ten tasks are T1.2, T1.5, T1.8, T1.10, T1.14, T1.15, T6.4, T6.5, T6.6 and T6.14. Every claim in this paper about a ban on prebuilt state preparation is a claim about those ten and about no other task; other descriptions forbid specific gates instead. The prompt below is therefore typical in form and not in content.

Both parts are reproduced below without abridgement or edit, with the user message shown for T1.2 as an example; the user message is the task description followed by one appended output reminder.

System message:

```
You are a Qiskit quantum-circuit generator.  For the given
task, write Python code using Qiskit that builds the circuit
as a module-level variable named `qc` — define it at the
top level, not inside a function or class — and include the
necessary imports.  Return ONLY the code, no explanation.
```

User message, task T1.2:

```
Create a quantum circuit that prepares the Bell state |Phi->
= (|00> - |11>) / sqrt(2).

Requirements:
- Use 2 qubits
- Start from the |00> state
- Note the minus sign between the basis states
- Build the circuit from gates:  do not use `initialize`,
`StatePreparation`/`prepare_state`, or a hand-written
`unitary` matrix.

Return a Qiskit QuantumCircuit object.

Return only the Qiskit code (a single code block).
```

The system message is fixed in `baseline/baseline.py`. The task description is the `prompt` field of T1.2 in the benchmark's task file, and the closing reminder is appended by the same script for every task. These parts agree with the `system_prompt` and `user_prompt` fields recorded for the 3,870 baseline generations, so what is printed here is what the models of the baseline received.

I TIER BREAKDOWN

This appendix holds the per model record behind §6: the pass rate of every model in every tier, all 105 pairwise contrasts of the panel with their bootstrap intervals, the split by task provenance, and the served model identifiers. App. K reads these numbers. A tier here is one of the six task categories of §4.1. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below, to within, exceeds or an inequality sign, and each end of a stated span, is rounded in its safe direction instead, and an equality sign joins only exact values. The interval ends of Tab. 12 are the one exception, stated in its caption.

Tab. 10 reports the tier pass rate of every model. Pooled over the 15 models the profile does not fall with the tier number. It is highest at Tier 5 (93.2%) and lowest at Tier 4 (56.7%, the variational tasks), with Tier 6 at 78.1% and Tiers 1 to 3 between 89.0% and 90.7%. Tier 4 is the lowest tier in all 15 rows; Tier 5 is the highest, ties included, in 10 of them.

Part of the Tier 4 result depends on how the runner binds parameters. It binds the test values to circuit parameters by name and binds every parameter whose name does not match to 0.0; only if no name matches and the counts agree does it bind in sorted name order instead. Both rules can decide a verdict without the circuit deciding it. Over the 720 Tier 4 generations, 408 of which pass, the sorted order binding carries 55 accepted generations on 9 tasks, and renaming the parameters turns 9 failures on T4.15 into passes. Whether the 55 would fail without the fallback is not measured. Three Tier 4 tasks, T4.11, T4.12 and T4.15, are solved by no model on a majority of its samples, and App. K.5 separates the causes: T4.11 and T4.12 are under specified, and on T4.15 the naming decides 9 of the 44 failures.

Table 10: Tier stratified pass rate of every model, as a percentage of the generations of that model in that tier (three samples per task); the pooled row is the rate over all 15 models. The profile does not fall with the tier number: Tier 4 is the lowest tier in every row, and pooled, Tier 5 is the highest, above Tiers 1 to 4. Tier 4 carries a caveat that is not difficulty alone: over the 720 Tier 4 generations the runner's sorted order binding of parameter values carries 55 accepted generations on 9 tasks, and 9 further generations on one task pass after a pure rename (§7). Every Tier 4 task carries a parametric test and every parametric task is in Tier 4, so the tier reading and the verification type reading cannot be separated here (Fig. 7).

Model	T1	T2	T3	T4	T5	T6
Gemini 3.8 Flash	100.0	98.2	96.7	68.8	100.0	90.7
Grok 4.6	100.0	93.0	100.0	64.6	100.0	92.6
Qwen3.8 Max	100.0	94.7	100.0	64.6	100.0	87.0
GLM 5.3	100.0	98.2	96.7	64.6	100.0	83.3
GPT 6 Astra	100.0	100.0	90.0	62.5	100.0	83.3
Fable 5.1	85.2	96.5	100.0	60.4	100.0	87.0
DeepSeek V4 Pro	100.0	87.7	96.7	62.5	97.6	81.5
Opus 5	88.9	91.2	96.7	68.8	92.9	79.6
Kimi K3	96.3	94.7	96.7	56.2	97.6	79.6
DeepSeek V4.1 Flash	96.3	89.5	100.0	45.8	100.0	77.8
Sonnet 4.6	88.9	91.2	83.3	45.8	100.0	68.5
Gemini 3 Flash	85.2	91.2	80.0	54.2	85.7	75.9
GPT OSS 120B	81.5	75.4	80.0	56.2	90.5	68.5
R1 Distill 70B	74.1	70.2	56.7	35.4	71.4	68.5
Qwen3 Coder	63.0	63.2	66.7	39.6	61.9	48.1
Pooled	90.6	89.0	89.3	56.7	93.2	78.1

Table 11: Every pairwise pass rate contrast of the 15 model panel, $\binom{15}{2} = 105$ pairs. The primary test is the task paired exact McNemar test over all 86 tasks: per task each model is dichotomised by majority vote over its three samples, the discordant column $d_{a>b}+d_{b>a}$ counts the tasks won by exactly one model, and the p value is the two sided exact binomial test on the discordant pairs, Benjamini–Hochberg adjusted at $\alpha = 0.05$ over $m = 105$ contrasts. The pooled Fisher exact p values on the 2×2 pass and fail table over 258 generations per model are a descriptive secondary; they treat the three samples of a task as independent. On this corpus they select the same *number* of contrasts as the paired test, 49 of 105, but not the same *set*. The Δpp column is the difference of pooled generation level pass rates, first model minus second. Cliff's δ is computed unpaired over the 86 per task pass@1 rates, a different unit of analysis from Δpp, and carries its magnitude bin. The evidence column follows the paired test: a contrast it separates is Clear Evidence if $|\delta| \geq 0.33$ and Evidence Present otherwise; a contrast it does not separate is a Trend if $|\Delta\text{pp}| > 2$ and Below Resolution otherwise. A per sample sensitivity analysis, three McNemar tests per pair, one per sample index, disagrees with the majority vote test on 24 contrasts when both are read at the same unadjusted $\alpha = 0.05$. On all 12 of them that the paired test separates, it is the majority vote that is significant where a single sample is not; the reverse direction occurs only among contrasts the paired test does not separate. Fig. 4 shows which pairs are separated. The smallest separating advantage on this corpus is 7 purely one sided tasks, 8.14 percentage points of the 86. That is not a threshold: a contrast of 10 tasks won against 3 lost does not separate, because the exact test scores the one sidedness and not the margin. The 56 of 105 contrasts that do not separate therefore do not show two models to be equally good; they show their difference to lie below the resolution of this design.

Pair (a vs b)	Δpp	disc.	p_{McN} (BH)	rej.	p_{Fisher} (BH)	Cliff δ (mag)	Evidence
Gemini 3.8 Flash vs Grok 4.6	+1.2	1+1	1.000 (1.000)	no	0.760 (0.840)	+0.022 (negl)	Below Resolution
Gemini 3.8 Flash vs Qwen3.8 Max	+1.9	3+2	1.000 (1.000)	no	0.548 (0.640)	+0.033 (negl)	Below Resolution
Gemini 3.8 Flash vs GLM 5.3	+2.3	2+3	1.000 (1.000)	no	0.457 (0.565)	+0.094 (negl)	Trend
Gemini 3.8 Flash vs GPT 6 Astra	+3.1	4+1	0.375 (0.518)	no	0.306 (0.407)	+0.016 (negl)	Trend
Gemini 3.8 Flash vs Fable 5.1	+3.9	3+1	0.625 (0.746)	no	0.195 (0.293)	+0.086 (negl)	Trend
Gemini 3.8 Flash vs DeepSeek V4 Pro	+5.8	6+1	0.125 (0.202)	no	0.052 (0.090)	+0.089 (negl)	Trend
Gemini 3.8 Flash vs Opus 5	+6.2	7+2	0.180 (0.277)	no	0.039 (0.074)	+0.081 (negl)	Trend
Gemini 3.8 Flash vs Kimi K3	+6.2	3+1	0.625 (0.746)	no	0.039 (0.074)	+0.121 (negl)	Trend
Gemini 3.8 Flash vs DeepSeek V4.1 Flash	+8.9	6+0	0.031 (0.062)	no	0.004 (0.009)	+0.145 (negl)	Trend
Gemini 3.8 Flash vs Sonnet 4.6	+13.2	12+0	0.0005 (0.002)	yes	< 0.0001 (0.0001)	+0.162 (small)	Evidence Present
Gemini 3.8 Flash vs Gemini 3 Flash	+13.2	11+0	0.0010 (0.003)	yes	< 0.0001 (0.0001)	+0.163 (small)	Evidence Present
Gemini 3.8 Flash vs GPT OSS 120B	+17.4	12+0	0.0005 (0.002)	yes	< 0.0001 (< 0.0001)	+0.270 (small)	Evidence Present
Gemini 3.8 Flash vs R1 Distill 70B	+29.1	27+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.358 (medium)	Clear Evidence
Gemini 3.8 Flash vs Qwen3 Coder	+35.7	30+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.441 (medium)	Clear Evidence
Grok 4.6 vs Qwen3.8 Max	+0.8	3+2	1.000 (1.000)	no	0.884 (0.937)	+0.011 (negl)	Below Resolution
Grok 4.6 vs GLM 5.3	+1.2	2+3	1.000 (1.000)	no	0.772 (0.843)	+0.071 (negl)	Below Resolution
Grok 4.6 vs GPT 6 Astra	+1.9	4+1	0.375 (0.518)	no	0.569 (0.656)	−0.006 (negl)	Below Resolution
Grok 4.6 vs Fable 5.1	+2.7	3+1	0.625 (0.746)	no	0.400 (0.506)	+0.064 (negl)	Trend
Grok 4.6 vs DeepSeek V4 Pro	+4.7	5+0	0.062 (0.117)	no	0.136 (0.219)	+0.068 (negl)	Trend
Grok 4.6 vs Opus 5	+5.0	6+1	0.125 (0.202)	no	0.106 (0.174)	+0.059 (negl)	Trend
Grok 4.6 vs Kimi K3	+5.0	3+1	0.625 (0.746)	no	0.106 (0.174)	+0.099 (negl)	Trend
Grok 4.6 vs DeepSeek V4.1 Flash	+7.8	6+0	0.031 (0.062)	no	0.014 (0.030)	+0.124 (negl)	Trend
Grok 4.6 vs Sonnet 4.6	+12.0	12+0	0.0005 (0.002)	yes	0.0002 (0.0007)	+0.141 (negl)	Evidence Present
Grok 4.6 vs Gemini 3 Flash	+12.0	11+0	0.0010 (0.003)	yes	0.0002 (0.0007)	+0.142 (negl)	Evidence Present
Grok 4.6 vs GPT OSS 120B	+16.3	12+0	0.0005 (0.002)	yes	< 0.0001 (< 0.0001)	+0.249 (small)	Evidence Present
Grok 4.6 vs R1 Distill 70B	+27.9	27+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.340 (medium)	Clear Evidence
Grok 4.6 vs Qwen3 Coder	+34.5	30+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.423 (medium)	Clear Evidence
Qwen3.8 Max vs GLM 5.3	+0.4	0+2	0.500 (0.656)	no	1.000 (1.000)	+0.059 (negl)	Below Resolution
Qwen3.8 Max vs GPT 6 Astra	+1.2	3+1	0.625 (0.746)	no	0.779 (0.843)	−0.016 (negl)	Below Resolution
Qwen3.8 Max vs Fable 5.1	+1.9	3+2	1.000 (1.000)	no	0.580 (0.662)	+0.052 (negl)	Below Resolution

(continued)

Pair (a vs b)	Δpp	disc.	p_{McN} (BH)	rej.	p_{Fisher} (BH)	Cliff δ (mag)	Evidence
Qwen3.8 Max vs DeepSeek V4 Pro	+3.9	4+0	0.125 (0.202)	no	0.229 (0.338)	+0.057 (negl)	Trend
Qwen3.8 Max vs Opus 5	+4.3	6+2	0.289 (0.422)	no	0.184 (0.280)	+0.048 (negl)	Trend
Qwen3.8 Max vs Kimi K3	+4.3	3+2	1.000 (1.000)	no	0.184 (0.280)	+0.087 (negl)	Trend
Qwen3.8 Max vs DeepSeek V4.1 Flash	+7.0	5+0	0.062 (0.117)	no	0.030 (0.062)	+0.112 (negl)	Trend
Qwen3.8 Max vs Sonnet 4.6	+11.2	12+1	0.003 (0.009)	yes	0.0007 (0.002)	+0.131 (negl)	Evidence Present
Qwen3.8 Max vs Gemini 3 Flash	+11.2	11+1	0.006 (0.016)	yes	0.0007 (0.002)	+0.131 (negl)	Evidence Present
Qwen3.8 Max vs GPT OSS 120B	+15.5	11+0	0.0010 (0.003)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.238 (small)	Evidence Present
Qwen3.8 Max vs R1 Distill 70B	+27.1	26+0	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.330 (medium)	Clear Evidence
Qwen3.8 Max vs Qwen3 Coder	+33.7	29+0	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.414 (medium)	Clear Evidence
GLM 5.3 vs GPT 6 Astra	+0.8	4+0	0.125 (0.202)	no	0.889 (0.937)	−0.074 (negl)	Below Resolution
GLM 5.3 vs Fable 5.1	+1.6	4+1	0.375 (0.518)	no	0.681 (0.760)	−0.004 (negl)	Below Resolution
GLM 5.3 vs DeepSeek V4 Pro	+3.5	6+0	0.031 (0.062)	no	0.288 (0.399)	+0.004 (negl)	Trend
GLM 5.3 vs Opus 5	+3.9	7+1	0.070 (0.123)	no	0.235 (0.338)	−0.005 (negl)	Trend
GLM 5.3 vs Kimi K3	+3.9	4+1	0.375 (0.518)	no	0.235 (0.338)	+0.031 (negl)	Trend
GLM 5.3 vs DeepSeek V4.1 Flash	+6.6	7+0	0.016 (0.035)	yes	0.043 (0.080)	+0.061 (negl)	Evidence Present
GLM 5.3 vs Sonnet 4.6	+10.9	13+0	0.0002 (0.0010)	yes	0.001 (0.003)	+0.087 (negl)	Evidence Present
GLM 5.3 vs Gemini 3 Flash	+10.9	12+0	0.0005 (0.002)	yes	0.001 (0.003)	+0.086 (negl)	Evidence Present
GLM 5.3 vs GPT OSS 120B	+15.1	13+0	0.0002 (0.0010)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.196 (small)	Evidence Present
GLM 5.3 vs R1 Distill 70B	+26.7	28+0	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.306 (small)	Evidence Present
GLM 5.3 vs Qwen3 Coder	+33.3	31+0	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.393 (medium)	Clear Evidence
GPT 6 Astra vs Fable 5.1	+0.8	2+3	1.000 (1.000)	no	0.892 (0.937)	+0.067 (negl)	Below Resolution
GPT 6 Astra vs DeepSeek V4 Pro	+2.7	3+1	0.625 (0.746)	no	0.432 (0.540)	+0.069 (negl)	Trend
GPT 6 Astra vs Opus 5	+3.1	4+2	0.688 (0.811)	no	0.362 (0.470)	+0.062 (negl)	Trend
GPT 6 Astra vs Kimi K3	+3.1	1+2	1.000 (1.000)	no	0.362 (0.470)	+0.101 (negl)	Trend
GPT 6 Astra vs DeepSeek V4.1 Flash	+5.8	5+2	0.453 (0.602)	no	0.080 (0.135)	+0.123 (negl)	Trend
GPT 6 Astra vs Sonnet 4.6	+10.1	10+1	0.012 (0.029)	yes	0.003 (0.007)	+0.139 (negl)	Evidence Present
GPT 6 Astra vs Gemini 3 Flash	+10.1	9+1	0.021 (0.047)	yes	0.003 (0.007)	+0.140 (negl)	Evidence Present
GPT 6 Astra vs GPT OSS 120B	+14.3	10+1	0.012 (0.029)	yes	$<$ 0.0001 (0.0001)	+0.243 (small)	Evidence Present
GPT 6 Astra vs R1 Distill 70B	+26.0	25+1	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.326 (small)	Evidence Present
GPT 6 Astra vs Qwen3 Coder	+32.6	28+1	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.407 (medium)	Clear Evidence
Fable 5.1 vs DeepSeek V4 Pro	+1.9	5+2	0.453 (0.602)	no	0.605 (0.683)	+0.007 (negl)	Below Resolution
Fable 5.1 vs Opus 5	+2.3	6+3	0.508 (0.658)	no	0.521 (0.614)	−0.001 (negl)	Trend
Fable 5.1 vs Kimi K3	+2.3	2+2	1.000 (1.000)	no	0.521 (0.614)	+0.035 (negl)	Trend
Fable 5.1 vs DeepSeek V4.1 Flash	+5.0	6+2	0.289 (0.422)	no	0.138 (0.219)	+0.063 (negl)	Trend
Fable 5.1 vs Sonnet 4.6	+9.3	12+2	0.013 (0.030)	yes	0.007 (0.015)	+0.087 (negl)	Evidence Present
Fable 5.1 vs Gemini 3 Flash	+9.3	11+2	0.022 (0.048)	yes	0.007 (0.015)	+0.087 (negl)	Evidence Present
Fable 5.1 vs GPT OSS 120B	+13.6	12+2	0.013 (0.030)	yes	0.0001 (0.0004)	+0.192 (small)	Evidence Present
Fable 5.1 vs R1 Distill 70B	+25.2	26+1	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.297 (small)	Evidence Present
Fable 5.1 vs Qwen3 Coder	+31.8	29+1	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.383 (medium)	Clear Evidence
DeepSeek V4 Pro vs Opus 5	+0.4	3+3	1.000 (1.000)	no	1.000 (1.000)	−0.007 (negl)	Below Resolution
DeepSeek V4 Pro vs Kimi K3	+0.4	0+3	0.250 (0.380)	no	1.000 (1.000)	+0.028 (negl)	Below Resolution
DeepSeek V4 Pro vs DeepSeek V4.1 Flash	+3.1	3+2	1.000 (1.000)	no	0.399 (0.506)	+0.054 (negl)	Trend
DeepSeek V4 Pro vs Sonnet 4.6	+7.4	8+1	0.039 (0.076)	no	0.039 (0.074)	+0.078 (negl)	Trend
DeepSeek V4 Pro vs Gemini 3 Flash	+7.4	7+1	0.070 (0.123)	no	0.039 (0.074)	+0.079 (negl)	Trend
DeepSeek V4 Pro vs GPT OSS 120B	+11.6	7+0	0.016 (0.035)	yes	0.001 (0.004)	+0.180 (small)	Evidence Present
DeepSeek V4 Pro vs R1 Distill 70B	+23.3	22+0	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.281 (small)	Evidence Present
DeepSeek V4 Pro vs Qwen3 Coder	+29.8	25+0	$<$ 0.0001 ($<$ 0.0001)	yes	$<$ 0.0001 ($<$ 0.0001)	+0.366 (medium)	Clear Evidence

(continued)

Pair (a vs b)	Δpp	disc.	p_{McN} (BH)	rej.	p_{Fisher} (BH)	Cliff δ (mag)	Evidence
Opus 5 vs Kimi K3	+0.0	2+5	0.453 (0.602)	no	1.000 (1.000)	+0.036 (negl)	Below Resolution
Opus 5 vs DeepSeek V4.1 Flash	+2.7	5+4	1.000 (1.000)	no	0.472 (0.570)	+0.061 (negl)	Trend
Opus 5 vs Sonnet 4.6	+7.0	9+2	0.065 (0.121)	no	0.052 (0.090)	+0.082 (negl)	Trend
Opus 5 vs Gemini 3 Flash	+7.0	9+3	0.146 (0.232)	no	0.052 (0.090)	+0.083 (negl)	Trend
Opus 5 vs GPT OSS 120B	+11.2	10+3	0.092 (0.159)	no	0.002 (0.005)	+0.184 (small)	Trend
Opus 5 vs R1 Distill 70B	+22.9	25+3	< 0.0001 (0.0001)	yes	< 0.0001 (< 0.0001)	+0.279 (small)	Evidence Present
Opus 5 vs Qwen3 Coder	+29.5	28+3	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.363 (medium)	Clear Evidence
Kimi K3 vs DeepSeek V4.1 Flash	+2.7	6+2	0.289 (0.422)	no	0.472 (0.570)	+0.028 (negl)	Trend
Kimi K3 vs Sonnet 4.6	+7.0	10+0	0.002 (0.005)	yes	0.052 (0.090)	+0.052 (negl)	Evidence Present
Kimi K3 vs Gemini 3 Flash	+7.0	9+0	0.004 (0.010)	yes	0.052 (0.090)	+0.052 (negl)	Evidence Present
Kimi K3 vs GPT OSS 120B	+11.2	10+0	0.002 (0.005)	yes	0.002 (0.005)	+0.158 (small)	Evidence Present
Kimi K3 vs R1 Distill 70B	+22.9	25+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.262 (small)	Evidence Present
Kimi K3 vs Qwen3 Coder	+29.5	28+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.348 (medium)	Clear Evidence
DeepSeek V4.1 Flash vs Sonnet 4.6	+4.3	7+1	0.070 (0.123)	no	0.267 (0.374)	+0.027 (negl)	Trend
DeepSeek V4.1 Flash vs Gemini 3 Flash	+4.3	7+2	0.180 (0.277)	no	0.267 (0.374)	+0.027 (negl)	Trend
DeepSeek V4.1 Flash vs GPT OSS 120B	+8.5	6+0	0.031 (0.062)	no	0.025 (0.052)	+0.127 (negl)	Trend
DeepSeek V4.1 Flash vs R1 Distill 70B	+20.2	21+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.235 (small)	Evidence Present
DeepSeek V4.1 Flash vs Qwen3 Coder	+26.7	24+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.321 (small)	Evidence Present
Sonnet 4.6 vs Gemini 3 Flash	+0.0	2+3	1.000 (1.000)	no	1.000 (1.000)	+0.001 (negl)	Below Resolution
Sonnet 4.6 vs GPT OSS 120B	+4.3	4+4	1.000 (1.000)	no	0.302 (0.406)	+0.093 (negl)	Trend
Sonnet 4.6 vs R1 Distill 70B	+15.9	16+1	0.0003 (0.001)	yes	0.0001 (0.0004)	+0.195 (small)	Evidence Present
Sonnet 4.6 vs Qwen3 Coder	+22.5	19+1	< 0.0001 (0.0002)	yes	< 0.0001 (< 0.0001)	+0.280 (small)	Evidence Present
Gemini 3 Flash vs GPT OSS 120B	+4.3	4+3	1.000 (1.000)	no	0.302 (0.406)	+0.093 (negl)	Trend
Gemini 3 Flash vs R1 Distill 70B	+15.9	16+0	< 0.0001 (0.0001)	yes	0.0001 (0.0004)	+0.193 (small)	Evidence Present
Gemini 3 Flash vs Qwen3 Coder	+22.5	19+0	< 0.0001 (< 0.0001)	yes	< 0.0001 (< 0.0001)	+0.278 (small)	Evidence Present
GPT OSS 120B vs R1 Distill 70B	+11.6	16+1	0.0003 (0.001)	yes	0.006 (0.014)	+0.120 (negl)	Evidence Present
GPT OSS 120B vs Qwen3 Coder	+18.2	19+1	< 0.0001 (0.0002)	yes	< 0.0001 (< 0.0001)	+0.209 (small)	Evidence Present
R1 Distill 70B vs Qwen3 Coder	+6.6	8+5	0.581 (0.744)	no	0.152 (0.238)	+0.088 (negl)	Trend

Table 12: Unpaired percentile bootstrap intervals on the pass rate difference of every contrast of Tab. 11, reported for description only. Each interval resamples the 86 per task pass@1 rates of the two models independently, 10,000 resamples at seed 42 under the bootstrap implementation shipped with the release, and takes the 2.5 and 97.5 percentiles of the resampled difference. The interval ends are rounded to the nearest tenth of a percentage point, so an end can lie up to 0.05 points inside the computed interval. These intervals do not enter the evidence column of Tab. 11, which follows the paired McNemar test. Whether the interval excludes zero disagrees with that test on 9 of the 105 contrasts, marked †: on eight of the nine the paired test separates a pair whose interval covers zero, and on one it is the other way round. The resampling is unpaired although both models are evaluated on the same 86 tasks; this is the form the released implementation computes, so the count of nine describes that estimator and not paired resampling in general.

Pair (a vs b)	Δpp	Bootstrap 95% CI (pp)	excl. 0	McNemar BH rej.
Gemini 3.8 Flash vs Grok 4.6	+1.2	[−7.0, +9.3]	no	no
Gemini 3.8 Flash vs Qwen3.8 Max	+1.9	[−6.2, +10.1]	no	no
Gemini 3.8 Flash vs GLM 5.3	+2.3	[−5.4, +9.7]	no	no
Gemini 3.8 Flash vs GPT 6 Astra	+3.1	[−5.4, +12.0]	no	no
Gemini 3.8 Flash vs Fable 5.1	+3.9	[−4.3, +12.0]	no	no
Gemini 3.8 Flash vs DeepSeek V4 Pro	+5.8	[−2.7, +14.3]	no	no

(continued)

Pair (a vs b)	Δpp	Bootstrap 95% CI (pp)	excl. 0	McNemar BH rej.
Gemini 3.8 Flash vs Opus 5	+6.2	[−2.7, +15.1]	no	no
Gemini 3.8 Flash vs Kimi K3	+6.2	[−2.3, +14.7]	no	no
Gemini 3.8 Flash vs DeepSeek V4.1 Flash	+8.9	[+0.0, +17.8]	no	no
Gemini 3.8 Flash vs Sonnet 4.6	+13.2	[+3.5, +22.9]	yes	yes
Gemini 3.8 Flash vs Gemini 3 Flash	+13.2	[+3.5, +22.9]	yes	yes
Gemini 3.8 Flash vs GPT OSS 120B	+17.4	[+7.8, +27.1]	yes	yes
Gemini 3.8 Flash vs R1 Distill 70B	+29.1	[+18.2, +39.5]	yes	yes
Gemini 3.8 Flash vs Qwen3 Coder	+35.7	[+24.8, +46.5]	yes	yes
Grok 4.6 vs Qwen3.8 Max	+0.8	[−7.8, +8.9]	no	no
Grok 4.6 vs GLM 5.3	+1.2	[−7.0, +8.9]	no	no
Grok 4.6 vs GPT 6 Astra	+1.9	[−7.0, +10.9]	no	no
Grok 4.6 vs Fable 5.1	+2.7	[−5.8, +11.2]	no	no
Grok 4.6 vs DeepSeek V4 Pro	+4.7	[−4.3, +13.2]	no	no
Grok 4.6 vs Opus 5	+5.0	[−4.3, +14.3]	no	no
Grok 4.6 vs Kimi K3	+5.0	[−3.9, +13.6]	no	no
Grok 4.6 vs DeepSeek V4.1 Flash	+7.8	[−1.6, +16.7]	no	no
Grok 4.6 vs Sonnet 4.6	+12.0	[+2.3, +22.1]	yes	yes
Grok 4.6 vs Gemini 3 Flash	+12.0	[+1.9, +21.7]	yes	yes
Grok 4.6 vs GPT OSS 120B	+16.3	[+6.2, +26.4]	yes	yes
Grok 4.6 vs R1 Distill 70B	+27.9	[+16.7, +38.8]	yes	yes
Grok 4.6 vs Qwen3 Coder	+34.5	[+23.3, +45.3]	yes	yes
Qwen3.8 Max vs GLM 5.3	+0.4	[−7.8, +8.1]	no	no
Qwen3.8 Max vs GPT 6 Astra	+1.2	[−7.8, +10.5]	no	no
Qwen3.8 Max vs Fable 5.1	+1.9	[−6.6, +10.5]	no	no
Qwen3.8 Max vs DeepSeek V4 Pro	+3.9	[−5.0, +12.8]	no	no
Qwen3.8 Max vs Opus 5	+4.3	[−5.0, +13.6]	no	no
Qwen3.8 Max vs Kimi K3	+4.3	[−4.7, +13.2]	no	no
Qwen3.8 Max vs DeepSeek V4.1 Flash	+7.0	[−2.3, +16.3]	no	no
Qwen3.8 Max vs Sonnet 4.6	+11.2	[+1.2, +21.3]	yes	yes
Qwen3.8 Max vs Gemini 3 Flash	+11.2	[+1.2, +21.3]	yes	yes
Qwen3.8 Max vs GPT OSS 120B	+15.5	[+5.4, +25.6]	yes	yes
Qwen3.8 Max vs R1 Distill 70B	+27.1	[+16.3, +38.0]	yes	yes
Qwen3.8 Max vs Qwen3 Coder	+33.7	[+22.1, +45.0]	yes	yes
GLM 5.3 vs GPT 6 Astra	+0.8	[−7.8, +9.3]	no	no
GLM 5.3 vs Fable 5.1	+1.6	[−6.2, +9.7]	no	no
GLM 5.3 vs DeepSeek V4 Pro	+3.5	[−4.7, +12.0]	no	no
GLM 5.3 vs Opus 5	+3.9	[−4.7, +12.8]	no	no
GLM 5.3 vs Kimi K3	+3.9	[−4.3, +12.4]	no	no
GLM 5.3 vs DeepSeek V4.1 Flash †	+6.6	[−1.9, +15.5]	no	yes
GLM 5.3 vs Sonnet 4.6	+10.9	[+1.6, +20.6]	yes	yes
GLM 5.3 vs Gemini 3 Flash	+10.9	[+1.2, +20.9]	yes	yes
GLM 5.3 vs GPT OSS 120B	+15.1	[+5.4, +24.8]	yes	yes
GLM 5.3 vs R1 Distill 70B	+26.7	[+16.3, +37.2]	yes	yes
GLM 5.3 vs Qwen3 Coder	+33.3	[+22.1, +44.2]	yes	yes
GPT 6 Astra vs Fable 5.1	+0.8	[−8.5, +9.7]	no	no
GPT 6 Astra vs DeepSeek V4 Pro	+2.7	[−7.0, +12.0]	no	no
GPT 6 Astra vs Opus 5	+3.1	[−6.6, +12.8]	no	no
GPT 6 Astra vs Kimi K3	+3.1	[−6.2, +12.4]	no	no
GPT 6 Astra vs DeepSeek V4.1 Flash	+5.8	[−3.9, +15.5]	no	no
GPT 6 Astra vs Sonnet 4.6 †	+10.1	[−0.4, +20.5]	no	yes
GPT 6 Astra vs Gemini 3 Flash †	+10.1	[−0.8, +20.5]	no	yes
GPT 6 Astra vs GPT OSS 120B	+14.3	[+3.5, +24.8]	yes	yes
GPT 6 Astra vs R1 Distill 70B	+26.0	[+14.3, +37.2]	yes	yes
GPT 6 Astra vs Qwen3 Coder	+32.6	[+20.5, +43.8]	yes	yes
Fable 5.1 vs DeepSeek V4 Pro	+1.9	[−7.0, +10.9]	no	no
Fable 5.1 vs Opus 5	+2.3	[−7.0, +11.6]	no	no
Fable 5.1 vs Kimi K3	+2.3	[−6.6, +10.9]	no	no
Fable 5.1 vs DeepSeek V4.1 Flash	+5.0	[−4.3, +14.3]	no	no
Fable 5.1 vs Sonnet 4.6 †	+9.3	[−0.4, +19.4]	no	yes
Fable 5.1 vs Gemini 3 Flash †	+9.3	[−0.8, +19.4]	no	yes
Fable 5.1 vs GPT OSS 120B	+13.6	[+3.5, +23.6]	yes	yes
Fable 5.1 vs R1 Distill 70B	+25.2	[+14.3, +36.0]	yes	yes

(continued)

Pair (a vs b)	Δpp	Bootstrap 95% CI (pp)	excl. 0	McNemar BH rej.
Fable 5.1 vs Qwen3 Coder	$+31.8$	$[+20.5, +43.0]$	yes	yes
DeepSeek V4 Pro vs Opus 5	$+0.4$	$[-9.3, +10.1]$	no	no
DeepSeek V4 Pro vs Kimi K3	$+0.4$	$[-8.5, +9.7]$	no	no
DeepSeek V4 Pro vs DeepSeek V4.1 Flash	$+3.1$	$[-6.6, +12.8]$	no	no
DeepSeek V4 Pro vs Sonnet 4.6	$+7.4$	$[-3.1, +17.8]$	no	no
DeepSeek V4 Pro vs Gemini 3 Flash	$+7.4$	$[-3.1, +17.8]$	no	no
DeepSeek V4 Pro vs GPT OSS 120B	$+11.6$	$[+1.2, +22.1]$	yes	yes
DeepSeek V4 Pro vs R1 Distill 70B	$+23.3$	$[+12.0, +34.5]$	yes	yes
DeepSeek V4 Pro vs Qwen3 Coder	$+29.8$	$[+18.2, +41.5]$	yes	yes
Opus 5 vs Kimi K3	$+0.0$	$[-9.7, +9.3]$	no	no
Opus 5 vs DeepSeek V4.1 Flash	$+2.7$	$[-7.4, +12.4]$	no	no
Opus 5 vs Sonnet 4.6	$+7.0$	$[-3.9, +17.8]$	no	no
Opus 5 vs Gemini 3 Flash	$+7.0$	$[-4.3, +17.8]$	no	no
Opus 5 vs GPT OSS 120B †	$+11.2$	$[+0.4, +21.7]$	yes	no
Opus 5 vs R1 Distill 70B	$+22.9$	$[+11.2, +34.1]$	yes	yes
Opus 5 vs Qwen3 Coder	$+29.5$	$[+17.4, +41.1]$	yes	yes
Kimi K3 vs DeepSeek V4.1 Flash	$+2.7$	$[-7.0, +12.4]$	no	no
Kimi K3 vs Sonnet 4.6 †	$+7.0$	$[-3.1, +17.4]$	no	yes
Kimi K3 vs Gemini 3 Flash †	$+7.0$	$[-3.5, +17.4]$	no	yes
Kimi K3 vs GPT OSS 120B	$+11.2$	$[+0.8, +21.7]$	yes	yes
Kimi K3 vs R1 Distill 70B	$+22.9$	$[+11.6, +34.1]$	yes	yes
Kimi K3 vs Qwen3 Coder	$+29.5$	$[+17.8, +40.7]$	yes	yes
DeepSeek V4.1 Flash vs Sonnet 4.6	$+4.3$	$[-6.6, +15.1]$	no	no
DeepSeek V4.1 Flash vs Gemini 3 Flash	$+4.3$	$[-6.6, +15.1]$	no	no
DeepSeek V4.1 Flash vs GPT OSS 120B	$+8.5$	$[-2.3, +19.0]$	no	no
DeepSeek V4.1 Flash vs R1 Distill 70B	$+20.2$	$[+8.5, +31.8]$	yes	yes
DeepSeek V4.1 Flash vs Qwen3 Coder	$+26.7$	$[+14.7, +38.4]$	yes	yes
Sonnet 4.6 vs Gemini 3 Flash	$+0.0$	$[-11.6, +11.2]$	no	no
Sonnet 4.6 vs GPT OSS 120B	$+4.3$	$[-7.4, +15.9]$	no	no
Sonnet 4.6 vs R1 Distill 70B	$+15.9$	$[+3.5, +27.9]$	yes	yes
Sonnet 4.6 vs Qwen3 Coder	$+22.5$	$[+9.7, +34.9]$	yes	yes
Gemini 3 Flash vs GPT OSS 120B	$+4.3$	$[-7.4, +15.9]$	no	no
Gemini 3 Flash vs R1 Distill 70B	$+15.9$	$[+3.5, +28.3]$	yes	yes
Gemini 3 Flash vs Qwen3 Coder	$+22.5$	$[+10.1, +34.9]$	yes	yes
GPT OSS 120B vs R1 Distill 70B †	$+11.6$	$[-0.8, +23.7]$	no	yes
GPT OSS 120B vs Qwen3 Coder	$+18.2$	$[+5.8, +30.6]$	yes	yes
R1 Distill 70B vs Qwen3 Coder	$+6.6$	$[-6.6, +19.8]$	no	no

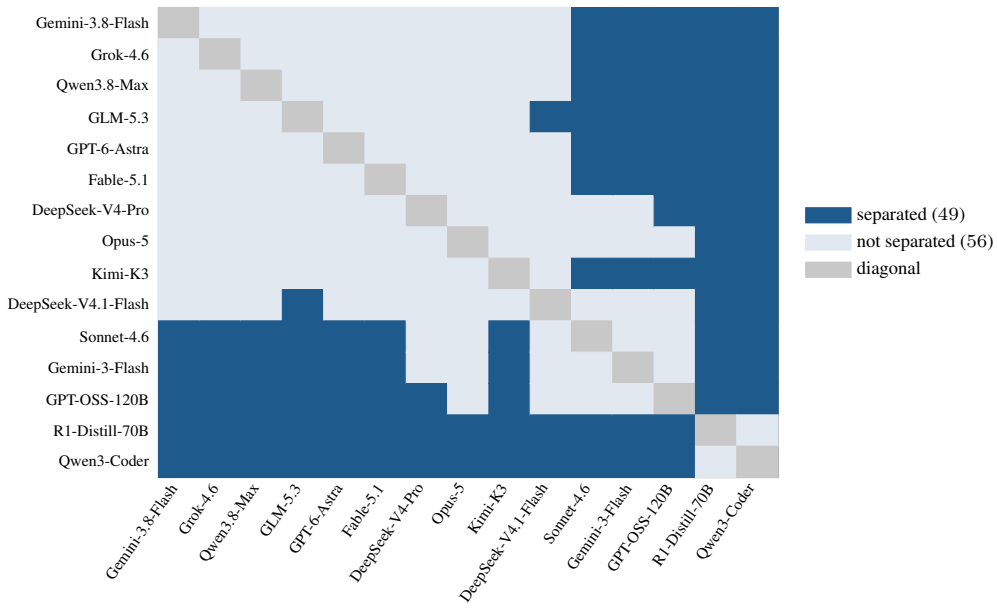


Figure 4: Which pairs of models the benchmark actually separates. Dark cells are the 49 of 105 contrasts that survive the task-paired exact McNemar test over all 86 tasks after Benjamini–Hochberg correction at $\alpha = 0.05$; pale cells are the 56 that do not, and the diagonal is not a contrast. Models are ordered as in Fig. 5, which is what makes the structure readable: the dark blocks sit away from the diagonal. Only *one* of the 14 adjacent pairs in that ordering separates, while the three weakest models are separated from most of the other models (10, 13 and 13 of their 14 contrasts, in ladder order). The benchmark resolves capability bands, not neighbouring models — a resolution claim that holds at 86 tasks and three samples and would need a larger corpus to sharpen. Every contrast with all its columns is in Tab. 11.

Table 13: Pass rates stratified by task provenance, from the Cite column of the task registry (Tab. 5): literature derived tasks follow a cited published construction, own constructions are specific to QuARC (T2.10 six state teleportation conjunction, T2.17 and T2.18 Simon variants with fixed hidden strings). Cells are the mean of per task pass@1 rates over three samples, over the full corpus (83 literature derived versus 3 own construction tasks) for every model. With three tasks the own construction stratum is descriptive only, and the two strata are not matched in difficulty, so the gap column must not be read causally. Six of the 15 models carry a negative gap, that is they do better on the own constructions than on the literature derived tasks.

Model	Scope	Literature derived	Own constr.	Gap (pp)
Gemini 3.8 Flash	T1–T6	91.2% (83)	100.0% (3)	−8.8
Grok 4.6	T1–T6	91.2% (83)	66.7% (3)	+24.5
Qwen3.8 Max	T1–T6	89.2% (83)	100.0% (3)	−10.8
GLM 5.3	T1–T6	88.8% (83)	100.0% (3)	−11.2
GPT 6 Astra	T1–T6	88.0% (83)	100.0% (3)	−12.0
Fable 5.1	T1–T6	87.1% (83)	100.0% (3)	−12.9
DeepSeek V4 Pro	T1–T6	85.9% (83)	77.8% (3)	+8.2
Opus 5	T1–T6	85.9% (83)	66.7% (3)	+19.3
Kimi K3	T1–T6	85.1% (83)	88.9% (3)	−3.7
DeepSeek V4.1 Flash	T1–T6	83.1% (83)	66.7% (3)	+16.5
Sonnet 4.6	T1–T6	78.7% (83)	66.7% (3)	+12.0
Gemini 3 Flash	T1–T6	79.5% (83)	44.4% (3)	+35.1
GPT OSS 120B	T1–T6	74.3% (83)	66.7% (3)	+7.6
R1 Distill 70B	T1–T6	62.7% (83)	55.6% (3)	+7.1
Qwen3 Coder	T1–T6	55.8% (83)	55.6% (3)	+0.3

Table 14: Served model identifiers and evaluation dates of the 15 model panel. All models are accessed via provider APIs with the serving endpoint pinned per request; provider side snapshot versions beyond these identifiers are not exposed, so the panel is identified by the named served models at the dates given. The identifier column is what the provider reported back per request, not what was asked for, and the dates are read from the records rather than from a run book.

Model	Served identifier	Evaluated
Gemini 3.8 Flash	`google/gemini-3.8-flash`	2026-09-11
Grok 4.6	`x-ai/grok-4.6`	2026-09-11
Qwen3.8 Max	`qwen/qwen3.8-max-0902`	2026-09-11–2026-09-12
GLM 5.3	`z-ai/glm-5.3`	2026-09-11–2026-09-12
GPT 6 Astra	`openai/gpt-6-astra`	2026-09-11
Fable 5.1	`anthropic/claude-fable-5.1`	2026-09-12
DeepSeek V4 Pro	`deepseek/deepseek-v4-pro-0813`	2026-09-11–2026-09-12
Opus 5	`anthropic/claude-opus-5`	2026-09-11
Kimi K3	`moonshotai/kimi-k3`	2026-09-11
DeepSeek V4.1 Flash	`deepseek/deepseek-v4.1-flash`	2026-09-11
Sonnet 4.6	`anthropic/claude-sonnet-4.6`	2026-09-10
Gemini 3 Flash	`google/gemini-3-flash-preview`	2026-09-10
GPT OSS 120B	`openai/gpt-oss-120b`	2026-09-10
R1 Distill 70B	`deepseek/deepseek-r1-distill-llama-70b`	2026-09-10–2026-09-11
Qwen3 Coder	`qwen/qwen3-coder-next`	2026-09-10

J Per Model Error Breakdown

Failing runs are of two kinds. In an extraction failure the reply yields no circuit object; in a semantic failure a circuit reaches the runner and fails verification. Every failing generation of the cohort is of one of the two kinds (App. L). Tab. 15 gives the per model split over the pooled generations.

The split differs strongly between models. Grok 4.6 and GPT 6 Astra have no extraction failure, while 66 of the 97 failures of R1 Distill 70B are extraction failures; 54 of those replies end with the finish reason `length` (App. L). Every other model fails more often on semantics than on extraction. Gemini 3.8 Flash has the fewest failures, 22 of its 258 generations, 2 of them extraction.

Table 15: Failure kinds per model, as generation counts over the 258 generations of each model. An extraction failure is a generation whose reply yields no circuit object; a semantic failure is a circuit rejected by the verification specification. Each row sums to the 258 generations of that model. Extraction accounts for none of the failures of Grok 4.6 and GPT 6 Astra and for 68% of those of R1 Distill 70B (App. K).

Model	n	Pass	Extraction fail	Semantic fail
Gemini 3.8 Flash	258	236	2	20
Grok 4.6	258	233	0	25
Qwen3.8 Max	258	231	7	20
GLM 5.3	258	230	5	23
GPT 6 Astra	258	228	0	30
Fable 5.1	258	226	7	25
DeepSeek V4 Pro	258	221	5	32
Opus 5	258	220	3	35
Kimi K3	258	220	9	29
DeepSeek V4.1 Flash	258	213	13	32
Sonnet 4.6	258	202	5	51
Gemini 3 Flash	258	202	8	48
GPT OSS 120B	258	191	18	49
R1 Distill 70B	258	161	66	31
Qwen3 Coder	258	144	41	73

K Model Run Analysis

§6 reports the headline numbers. This appendix asks two questions a ranking cannot answer: what the models fail at, and how much of the benchmark still distinguishes them. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below, to within, exceeds or an inequality sign, and each end of a stated span, is rounded in its safe direction instead, and an equality sign joins only exact values.

All numbers below are computed over the September cohort: 15 models, 86 tasks and 3 samples, 3,870 generations, each against a pinned provider endpoint and each scored against the shipped dataset v1.2.0 rather than the verdict recorded at collection time. Nine of the 3,870 verdicts differ between the two. Of the 3,870, 3,158 pass.

K.1 The Ranking and Its Two Failure Modes

Fig. 5 gives the pass rate of each model with a Wilson 95% interval and splits the failures into the two kinds the runner reports: a reply that yields no circuit object (an extraction failure) and a circuit that fails verification (a semantic failure). The pass rates run from 91.5% (Gemini 3.8 Flash) to 55.8% (Qwen3 Coder).

The split differs strongly between models. Extraction accounts for none of the failures of Grok 4.6 and GPT 6 Astra and for 68% of those of R1 Distill 70B, 66 of its 97. 54 of these 66 replies end with the finish reason `length`, after 7,068 to 7,242 completion tokens (App. L), so the split also reflects where generation was stopped. A score that pools the two kinds hides that difference.

One further reading concerns the acceptance region rather than the models. On the 27 tasks graded by a single scalar fidelity, 93 generations fail semantically, and each records a fidelity value. Of those, 81 (87.1%) lie at or below $F = 0.5$, two lie in $[0.95, 0.99)$, and none reaches $F \geq 0.99$. These failures are not near misses: lowering the threshold to 0.95 would turn two of them into passes. The other 42 failing generations on these tasks are extraction failures and are left out; 24 of them carry a placeholder value of 0 and 18 no value. The restriction to scalar graded tasks matters: on a parametric test the `fidelity` field records the minimum over the evaluated test points, starting from 1, so one T4.12 record shows $F = 1.000000$ and fails because none of its test points could be evaluated.

K.2 How Much of the Benchmark Still Separates

Fig. 6 counts, for each task, how many of the 15 models solve it on a majority of their three samples. 38 of the 86 tasks are solved by all 15, three by none, and 45 by some but not all.

Those 45 tasks carry the comparison. The five models carried over from the earlier cohort reach 69.8% over all 86 tasks and the ten newer ones 87.5%, a gap of 17.8 points. On the 45 the same comparison is 50.4% against 83.2%, a gap of 32.8 points. On the 38 saturated tasks every model passes a majority of its samples, so they separate no pair under the paired test of App. I. That is a limitation with a known direction: a successor cohort needs replacements for them.

Do the tasks accept each other's answers? Overlapping acceptance regions would cost resolution as well, and this is measured. The release submits every canonical solution to every other task with the same qubit count, 1,532 ordered pairs, through the shipped runner, after a positive control in which all 86 canonical solutions pass their own task. Eleven submissions are accepted. Nine of them are not defects: five build the same unitary as the target's canonical solution up to a global phase, two the same truth table, and two differ only by a qubit relabeling the runner permits, so on those nine pairs the two tasks ask for the same behavior as far as their tests read it. Two are false positives: T5.10 accepts the canonical solution of T5.8, and T6.9 that of T5.6. Even under the most favorable qubit relabeling the runner permits, the circuits differ on 4 and on 12 of the 32 basis inputs, but the two tests check only 8 and 5 inputs, and on those the circuits agree. That is 0.13% of the pairs, and it concerns test case coverage on two tasks rather than the acceptance radius.

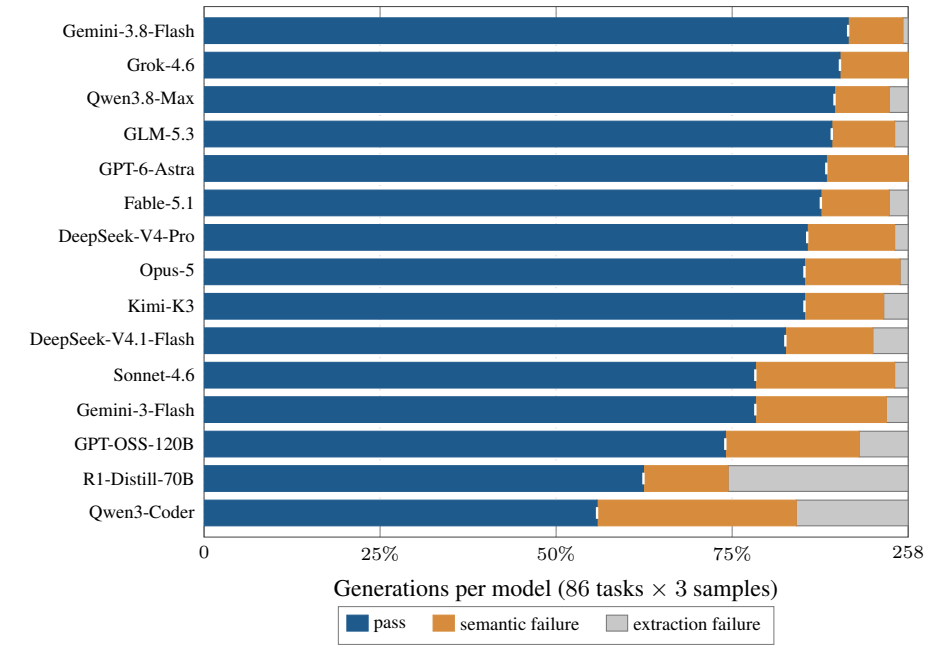


Figure 5: Pass rate of the 15 models over the 3,870 generations of the September cohort, 258 per model (86 tasks × 3 samples, provider pinned), scored against the shipped dataset v1.2.0. Bars are stacked counts, and the white whisker is the Wilson 95% interval on the pass rate. A failure is of one of two kinds: the reply yields no circuit object (extraction) or a circuit fails verification (semantic). The extraction share of a model's failures is 0% for Grok 4.6 and GPT 6 Astra and 68% for R1 Distill 70B.

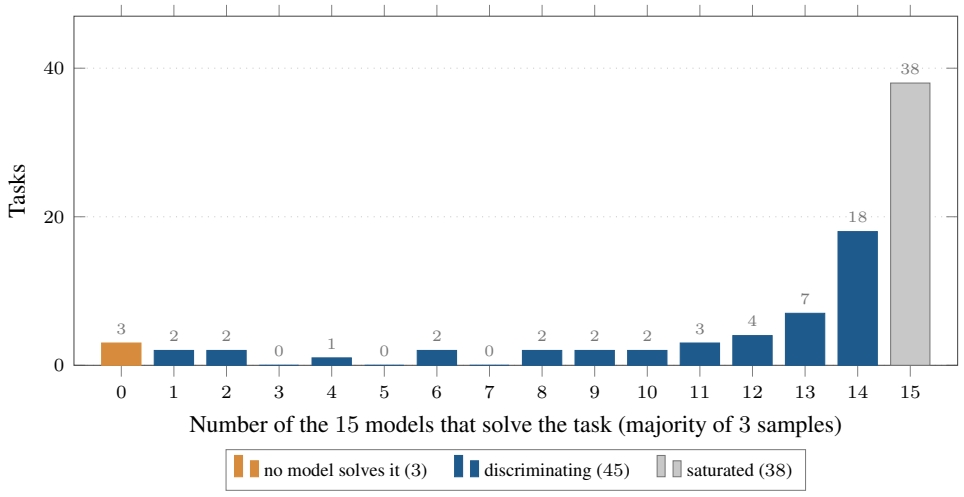


Figure 6: How much of the benchmark still separates models: for each task, the number of the 15 models that solve it on a majority of their three samples. Of the 86 tasks, 38 are solved by all 15 and 45 by some but not all; three, T4.11, T4.12 and T4.15, are solved by none (App. K.5). On the 45 the five models of the earlier cohort reach 50.4% and the ten newer ones 83.2%, against 69.8% and 87.5% over all 86. No task falls in the bins $k = 3, 5, 7$; they are drawn empty on purpose.

K.3 Tier, Verification Type, and a Confound We Do Not Resolve

Pooled over all 15 models the pass rate by tier is 90.6%, 89.0%, 89.3%, 56.7%, 93.2% and 78.1% for Tiers 1 to 6 (Tab. 10 in App. I gives every model). The profile does not fall with the tier number: the lowest rate is at Tier 4, and Tier 5 is above Tiers 1 to 4.

Fig. 7 puts that trough beside the pass rate by verification type, and the two panels show one observation. Every one of the 16 Tier 4 tasks carries a parametric test, and every parametric task is in Tier 4. "Tier 4 is the hardest stratum" and "parametric verification is the hardest kind" are one finding, and this design cannot separate them.

What the type axis adds is a comparison of the same kind of quantity checked without free parameters and checked at two to four parameter points: 93.1% for unitary fidelity against 52.2% for parametric unitary fidelity, and 81.1% for state fidelity against 58.1% for parametric state fidelity. Part of that gap need not be difficulty, for the reason given in App. K.5.

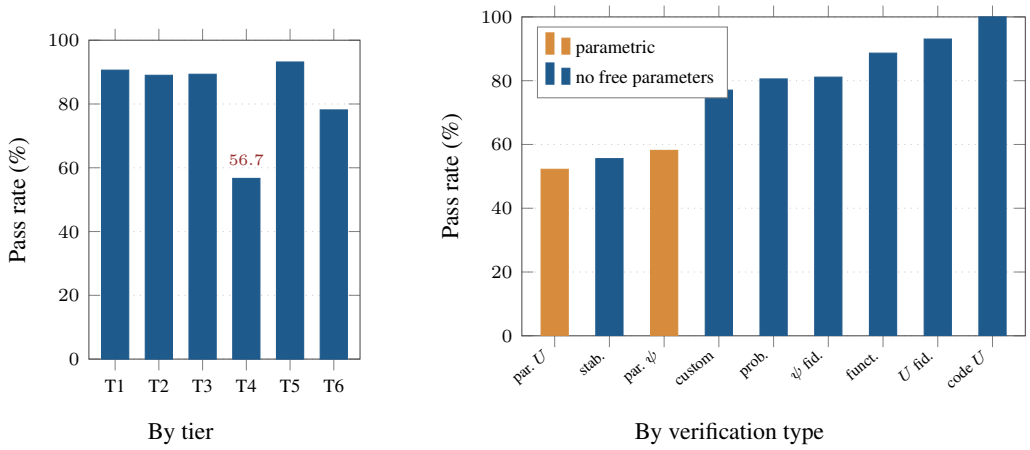


Figure 7: The Tier 4 trough and the parametric penalty are one observation. Every one of the 16 Tier 4 tasks carries a parametric test and every parametric task is in Tier 4, so the two readings cannot be separated in this design. The type axis adds a comparison of the same kind of quantity verified without free parameters and at two to four parameter points: unitary fidelity 93.1% against parametric unitary fidelity 52.2%, state fidelity 81.1% against parametric state fidelity 58.1%. Part of that gap need not be difficulty: the runner binds parameter values by name, and for the three unsolved tasks the prompt does not state every required name (App. K.5).

K.4 Where the Models Agree to Fail

Fig. 8 shows, for each pair of models, the Jaccard similarity of the sets of tasks they fail. The lowest value is 0.16 (GLM 5.3 and Qwen3 Coder), the highest 0.76 (Sonnet 4.6 and Gemini 3 Flash). Among the five models with the most passes it lies between 0.44 and 0.75. Counting a task as failed when fewer than two of its three samples pass, three tasks are failed by all 15 models, and 48 of the 86 by at least one. The shared core is small, and the comparison rests on where the models differ.

K.5 Three Tasks No Model Solves

T4.11, T4.12 and T4.15 are solved by no model on a majority of its samples. Of their 135 generations one passes, a T4.15 generation of DeepSeek V4 Pro at seed 42. The positive control of the cross submission above passes the canonical solution of all 86 tasks, and the three canonical solutions pass again over the path used below, so each of the three tests accepts a correct circuit. Whether a model can infer that circuit and its parameter names from the prompt is a separate question, and the three prompts answer it differently. The counts of built circuits, qubit widths, matching names and the passes after renaming below come from one run of the release script that measures them, `baseline/unsolved_tasks.py`.

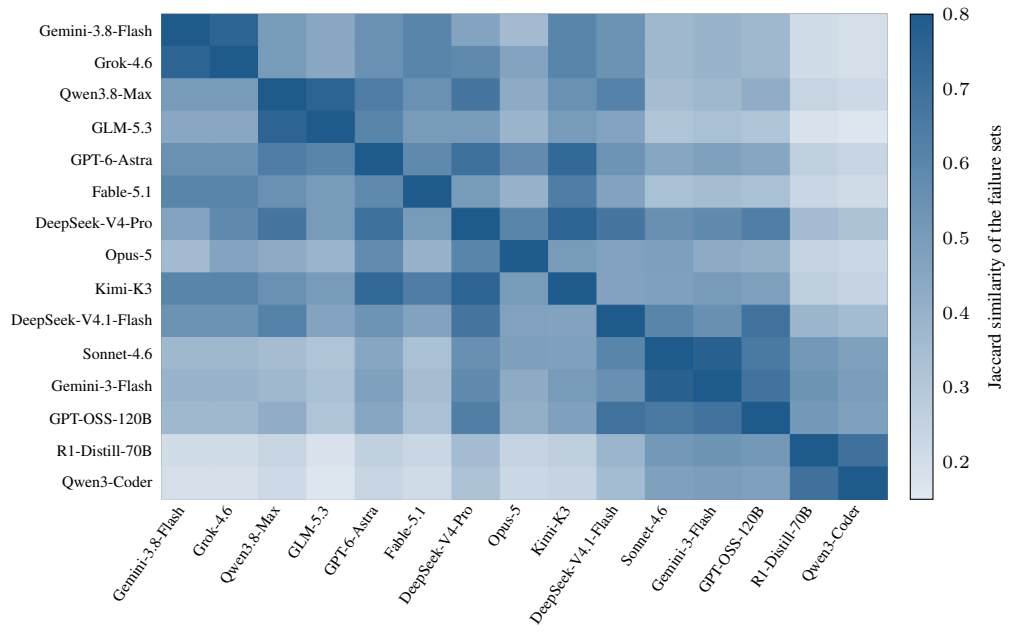


Figure 8: Do the models fail on the same tasks? For each pair, the Jaccard similarity of the sets of tasks they fail, where a task counts as failed when fewer than two of its three samples pass. Models are ordered as in Fig. 5. The lowest value is 0.16 (GLM 5.3 and Qwen3 Coder), the highest 0.76 (Sonnet 4.6 and Gemini 3 Flash). Three tasks are failed by all 15 models and 48 of the 86 by at least one.

The mechanism. The runner binds the test values to circuit parameters by name. Every parameter whose name does not occur in the test case is bound to 0.0. Only if no name matches at all and the counts agree does the runner instead bind in sorted name order. A circuit that matches some but not all of the required names is therefore evaluated with its remaining angles at zero, and the runner reports only a low fidelity.

T4.11 and T4.12 are under specified. Their prompts have 97 and 91 characters, the two shortest of the benchmark, against a median of 418.5 over the 86 tasks. The whole of T4.11 reads: *"Implement the Quantum Convolutional Layer circuit. Return a Qiskit QuantumCircuit object as 'qc'."* The circuit it is graded against, three overlapping pairs on four qubits with an R_y on each member, a CZ between them and two rotation angles shared across the pairs, is fixed only by the canonical solution, which the model does not see. Of the 45 T4.11 generations, 43 build a circuit; 32 of these have four qubits and 10 have eight. None of the 43 uses the names the test requires, `phi` and `theta`. The prompt of T4.12 names no parameter either; 40 of its 45 generations build, 20 of them on four qubits, and one uses the required name.

On T4.15 the names decide part of the failures. Its prompt fixes the structure: two qubits, $R_x(x)$ on each, $R_y(\theta)$ on each as trainable parameters, then CNOT. All 45 generations build a two qubit circuit. The test requires three names, `x`, `theta_0` and `theta_1`, with two independent trainable angles; the prompt gives `x` and writes both angles as θ. 22 generations use at least one of the required names. The release then maps the parameter names of every generation whose parameter count matches onto the required names in sorted order, changes nothing else, and runs the same pinned test. Of these 13 generations, 10 pass; one of them had passed already, so 9 of the 44 failures turn into passes, the 9 of §7. The other three, all from Grok 4.6, still fail the fidelity check. The remaining failures of T4.15 are not traced to a cause here.

What does not follow. It does not follow that a task fails whenever its prompt omits the required names. Among the 16 Tier 4 tasks, the ten whose prompts name every required parameter are solved by 12.1 models on average, the five that name none by 3.6, and the one that names some, T4.15,

by none. But T4.13 names none of its required parameters and is solved by all 15: 38 of its 45 generations call Qiskit's `RealAmplitudes` class, whose default parameter names are the required $\theta[0]$ to $\theta[11]$. In the other direction T4.5 names both of its parameters and is solved by two models. The names are one factor among others; where the test binds by name, a circuit can fail on a name the prompt does not fix, and a benchmark should not score that.

We report this rather than repair it. The shipped dataset stays at v1.2.0: all 3,870 generations were produced against the present prompts, and changing them would invalidate the measurements rather than improve them. The correction belongs to a successor release, together with the replacements the saturated tasks call for.

K.6 The Chance Floor

34 of the 86 tasks are escalated: their first bound lies above 10^{-2}, and they carry a sharpened bound (§4.2). All 86 tasks also carry an empirical Haar calibration (App. D), whose 95% upper bound concerns a predicate of its own, in general neither the modeled event nor the shipped test. Fig. 9 puts that bound for the 34 escalated tasks next to what the models achieve.

The reading is a distance in orders of magnitude, not a probability that chance passes the shipped test. On 28 of the 34 tasks the calibrated bound is at most 3.9×10^{-9}, because the calibration found no hit in 10^9 draws, and the median pass rate over the 34 is 93.3%, more than eight orders of magnitude above that bound.

Two cautions belong with the figure. The apparent downward trend comes from the three tasks with the loosest bounds, T2.17, T2.18 and T6.6 (Pearson $r \approx -0.34$ over 34 points on the log scale, $r \approx +0.18$ without the three), and we do not read it as an effect. And the escalation criterion concerns the first bound on the modeled event, not how hard a task is for a model: the 34 escalated tasks are solved at 87.6% and the other 52 at 77.6%.

Five of the 34 calibrations, those of T2.9, T2.21, T2.24, T2.26 and T6.16, were completed only for this release. The five are not a neutral sample: their pass rates are 77.8%, 88.9%, 57.8%, 88.9% and 37.8%, the last the lowest point in the figure.

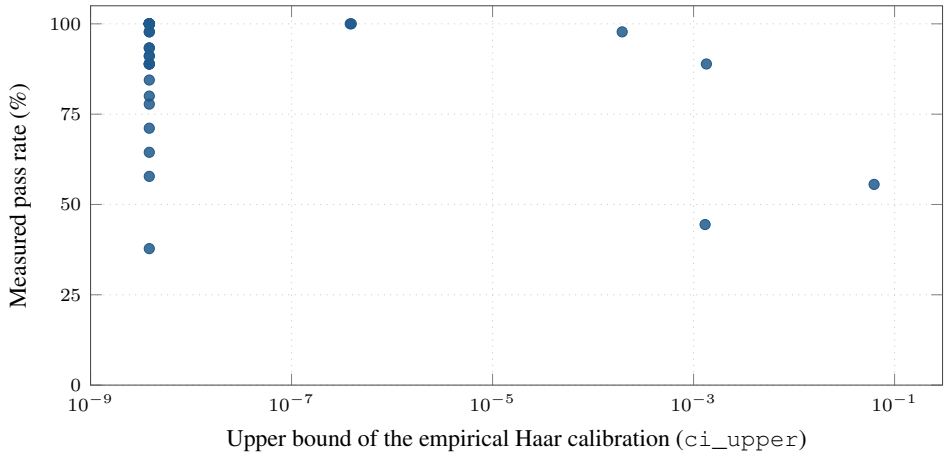


Figure 9: Calibrated bound against measured pass rate, one point per escalated task (34). The horizontal axis, on a log scale, is the 95% upper bound of the empirical Haar calibration (`ci_upper`), which samples a predicate of its own, in general neither the modeled event nor the shipped test (App. D); the vertical axis is the pass rate over all 15 models and three samples. 28 tasks share the same bound, at most 3.9×10^{-9}, because their calibration found no hit in 10^9 draws, so they are to be read vertically. The apparent downward trend comes from the three tasks with the loosest bounds (Pearson $r \approx -0.34$ over the 34 points on the log scale, $r \approx +0.18$ without those three), and we do not read it as an effect. The escalated tasks are solved at 87.6%, the other 52 at 77.6%.

K.7 PRICE

Fig. 10 reports what each ranking position costs. The cheapest full sweep costs $0.04 (GPT OSS 120B, 74.0%) and the most expensive $8.79 (GLM 5.3, 89.1%), a factor of 225 in price for 15 percentage points in rate. Price and rate are positively related: their rank correlation over the 15 models is 0.56 (Spearman), and the leading model costs $1.18, less than ten of the other fourteen. The recorded costs of the cohort sum to $55.01 over 3,867 of the 3,870 generations; three Opus 5 records carry no cost value.

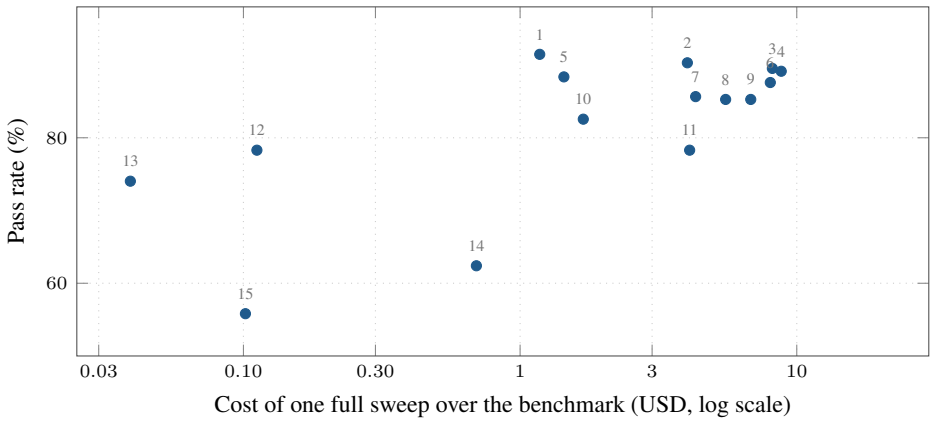


Figure 10: What the ranking costs. One point per model: the recorded price of its 258 generations against the pinned endpoint, on a log scale, against its pass rate; numbers are the ranks of Fig. 5. The cheapest sweep costs $0.04 (GPT OSS 120B, 74.0%) and the most expensive $8.79 (GLM 5.3, 89.1%), a factor of 225 in price for 15 percentage points in rate. The rank correlation of price and pass rate is 0.56 (Spearman), and Gemini 3.8 Flash leads the field at $1.18, less than ten of the other fourteen. The recorded costs sum to $55.01 over 3,867 of the 3,870 generations: three Opus 5 records carry no cost value, so that model's sum is a lower bound.

K.8 SAMPLE TO SAMPLE STABILITY AND OUTPUT LENGTH

Tab. 16 collects two per model quantities that read better as numbers than as figures.

The left half counts, for each model, how its 86 tasks divide by the number of samples passed. Across the cohort, 144 of the 1,290 model and task cells (11.2%) are mixed, so on these cells three samples do not settle the result, and a table built on a single sample could differ from run to run. The models differ here as well: GPT 6 Astra has no mixed cell, every task passing three times or not at all, while GLM 5.3 has 14 and R1 Distill 70B 17.

The right half compares the median completion length of passing and failing generations. For all 15 models the failures are longer, by factors from 1.30 to 7.10. Across models the same quantity is a weak guide to ability: GPT 6 Astra places fifth with a median of 118 tokens on its passes, and GLM 5.3 fourth with 1,313.5.

Table 16: Per model stability across the three samples and completion length by outcome, over the 3,870 generations of the September cohort. The four stability columns count a model's 86 tasks by how many of its three samples pass; they sum to 86 in every row, and their totals are 174, 46, 98 and 972. Lengths are medians of the completion token count over the generations that record one (3 Opus 5 records have none), printed exactly, and the factor is the ratio of the two. Models are ordered as in Fig. 5.

	Tasks by samples passed				Completion tokens (median)		
Model	0/3	1/3	2/3	3/3	pass	fail	factor
Gemini 3.8 Flash	6	1	2	77	940.5	2944	3.13
Grok 4.6	7	0	4	75	880	2642	3.00
Qwen3.8 Max	7	1	4	74	1226	5078	4.14
GLM 5.3	4	2	12	68	1313.5	7688	5.85
GPT 6 Astra	10	0	0	76	118	232	1.97
Fable 5.1	6	3	8	69	217	282.5	1.30
DeepSeek V4 Pro	8	4	5	69	1187	6409	5.40
Opus 5	10	2	4	70	257.5	469	1.82
Kimi K3	9	0	11	66	471	2262	4.80
DeepSeek V4.1 Flash	10	3	9	64	893	3502	3.92
Sonnet 4.6	14	5	4	63	238.5	689.5	2.89
Gemini 3 Flash	15	3	5	63	191.5	335.5	1.75
GPT OSS 120B	15	4	14	53	420	918	2.19
R1 Distill 70B	23	11	6	46	997	7075	7.10
Qwen3 Coder	30	7	10	39	151	390.5	2.59
Total	174	46	98	972			

L Qualitative Failure Patterns

Beyond the aggregate numbers of §6, the runner's messages show the shape of the failures. Numbers printed in this appendix are rounded to the nearest value at the digits shown; a bound stated with at most, at least, below, to within, exceeds or an inequality sign, and each end of a stated span, is rounded in its safe direction instead, and an equality sign joins only exact values.

Of the 3,870 generations, 712 fail, and each failure is of one of two kinds. In an extraction failure the reply yields no circuit object; in a semantic failure a circuit reaches the runner and fails verification. Every semantic failure carries extracted code. The extraction failures include every reply that the provider ended other than normally: 62 with the finish reason `length`, 18 with a provider error and 10 with a content filter. The label records what the runner reported, not where the fault lies. Nine semantic failures, all on T4.15, come from the runner's parameter binding rather than from the circuit (§L.2).

Tab. 17 crosses the two kinds with the six tiers, pooled over all 15 models. The Tier 4 trough is semantic: of the 312 Tier 4 failures, 275 are semantic and 37 extraction, consistent with the 56.7% Tier 4 pass rate of Tab. 10.

Table 17: Failure kinds by tier, pooled over all 15 models and the three samples (seeds 42, 123 and 777). Each row gives the generations, passes, extraction failures (no circuit object) and semantic failures (a circuit that fails verification) of that tier; the columns after the first sum to it. Tier 4 is the trough, and its failures are mostly semantic.

Tier	Generations	Pass	Extraction fail	Semantic fail
T1	405	367	18	20
T2	855	761	23	71
T3	450	402	17	31
T4	720	408	37	275
T5	630	587	36	7
T6	810	633	58	119
Total	3,870	3,158	189	523

L.1 Extraction Failures

189 of the 712 failures are extraction failures. The runner gives one of two reasons for them. `Failed to extract circuit from code`, in 122 cases, means that the reply holds code which raises an error when it is executed; code that runs but defines no circuit object would receive the same message with a diagnosis appended, and none of the 122 carries one. `no code extracted`, in the other 67, means that the reply is empty. By the provider's finish reason, 99 of the 189 replies end normally, 62 with the finish reason `length`, 18 with a provider error and 10 with a content filter; all 67 empty replies are among the last three groups.

Every model contributes 258 generations, so the counts read directly as shares. R1 Distill 70B has 66 extraction failures (25.6% of its generations) and Qwen3 Coder 41 (15.9%); the other thirteen models lie between 0% and 7.0%, with none for Grok 4.6 and GPT 6 Astra. R1 Distill 70B and Qwen3 Coder fail differently. 54 of the 66 replies of R1 Distill 70B end with the finish reason `length`, after 7,068 to 7,242 completion tokens, and 53 are empty; all 41 of Qwen3 Coder hold code that yields no circuit object, 13 of them ended by a provider error.

This kind therefore has more than one cause: code that raises an error when it is executed, and replies that the provider ended other than normally. In no case did a circuit reach the verifier. Whether a stricter output contract or a larger completion budget would remove it is untested here: all 3,870 generations were collected under the single prompt of App. H and one endpoint per model, so the corpus holds no variation in either to measure against.

L.2 Semantic Failures

The other 523 failures are semantic. By tier they number 20, 71, 31, 275, 7 and 119 for Tiers 1 to 6, so Tiers 2 and 4 together hold 346 of the 523. Read by the kind of check the runner reports, 234 are parametric fidelity checks (197 state, 37 unitary), 90 functional checks, 62 state fidelity and 31 unitary fidelity checks, 44 structural checks on qubit count, parameter count or gate use, 27 Simon checks, 22 probability checks and 13 stabilizer checks.

The first shape is a wrong output state, a circuit that runs but whose state is far from the target. On T6.5, the Steane code encoding, the runner records `State fidelity: 0.250000 (threshold: 0.999)` in 23 generations.

The second shape is wrong functional behavior, where an arithmetic, oracle or walk circuit computes the wrong map. On T2.24, the quantum walk on a four position cycle, the runner records `Functional: 0/8 test cases passed` in 14 generations. On T2.18, Simon's algorithm for the hidden string 110, it records `Simon s=110: valid_prob=0.5000, invalid_y=[3, 4], missing_valid_outputs=[1, 6]` in 14 generations: half of the measured probability falls on outputs that are invalid for that string. On T6.6 the stabilizer check fails with a message that begins `Stabilizers: 0/4 with correct eigenvalues` in 12 generations.

The third shape is a parametric ansatz mismatch, a Tier 4 circuit that misses the target across the tested angles. Of the 234 parametric failures, 231 pass none of the tested parameter settings; examples are messages that begin `Parametric state fidelity: 0/3 passed (min F=0.437500)` on the T4.5 QAOA task, in 34 generations, and `Parametric unitary fidelity: 0/3 passed (min F_U=0.140625)` on the T4.6 single excitation ansatz. A structural error can take its place: on T4.11 the runner records `Qubits: 8 (expected 4)` in 10 generations, and on T4.12 a message containing `Parameters: 0 (min 1)` in 11. T4.11 and T4.12 fail in all 45 of their generations. Parametric failures make up 234 of the 275 semantic failures of Tier 4.

The shape is not in every case the model's. Renaming the parameters of a recorded generation to the names the test cases bind, without touching a gate, turns 9 failures into passes, all on T4.15. Of the 312 failing Tier 4 generations, 275 carry a readable circuit and could be rerun; 11 have no code and 26 no readable circuit. On no other Tier 4 task does the renaming turn a failure into a pass. App. K.5 gives the mechanism.

L.3 Cross Cutting Clustering

Failures do not spread evenly over the tasks. Of the 86 tasks, 58 ever fail. The ten most failed tasks already hold more than half of the 712 failing generations; the nine most failed hold 355 and the seventeen most failed, a fifth of the suite, hold 504. The worst are T4.11, the quantum convolutional layer, and T4.12, the quantum pooling layer, with 45 of 45 each, T4.15, the data reuploading layer, with 44 of 45, and T6.5, the Steane code encoding, with 42 of 45. The first three are the tasks no model solves on a majority of its samples, and App. K.5 separates their causes: T4.11 and T4.12 are under specified, and on T4.15 the naming of its parameters decides 9 of the 44 failures.

The seventeen most failed tasks are not spread over the tiers either. Eight are variational (Tier 4) and five error correction or algorithm tasks of Tier 6, among them T6.19, where 13 generations fail because their circuits use control flow operations the check cannot convert; the remaining four are two of Tier 2 and one each of Tiers 1 and 3.

Per task failure counts and run identifiers are not printed here. The released `baseline/baseline_dataset.csv` holds one row per generation, with model, seed, task, tier, the verdict under v1.2.0 and the runner's message cut at 160 characters, and the release ships the raw generations. The runner side caveats that bear on this ranking are collected in §7.