

Falsification Criteria for Quantum-Advantage Claims: An Effective-Criterion Audit Framework

Falsification Criteria for Quantum-Advantage Claims: An Effective-Criterion Audit Framework

Abstract

Claims of quantum advantage are frequently stated in forms that resist empirical refutation: the advantage is qualified by hardware assumptions, sampling definitions, or problem encodings that can be revised after a classical rebuttal appears. This paper constructs a structured falsification framework for such claims, modeled on the pattern of *effective* criteria in algebra — criteria satisfied on a dense, open subset of the relevant class of objects — as found in tensor identifiability, critical-pair elimination for Gröbner bases, and finite-step decidability of quantum gate universality, and on the model-relative character of falsification debates in consciousness studies. We define a five-component falsification index $F \in [0, 1]$ covering specification completeness, inference–prediction independence, classical-baseline auditability, replication independence, and pre-registered failure thresholds, with fixed weights and explicit decision thresholds. We instantiate the framework against two existing audit artifacts: the QNFO full-spectrum audit of the Huang–Choi–McClean–Preskill five-keystone framework and the QNFO consilient gap synthesis, which catalogs 42 gaps across 5 structural categories with a phased roadmap spanning 8 calendar weeks. Quantitative summaries are derived with fully shown arithmetic: 8.4 gaps per category, 2.0 weeks per roadmap transition, 1.5×10^{11} tokens per language as an upper bound on the mean of a 30-trillion-token corpus, and worked index values $F = 0.675$ and $F = 0.10$. Failure modes, falsification conditions for the framework itself, and open questions are stated explicitly.

1. Introduction

A quantum-advantage claim asserts that a quantum device solves some computational task faster, cheaper, or with fewer resources than any classical method. The community's response to contested claims has been largely reactive: each contested experiment generates an ad hoc classical simulation, a dispute about resource counts, and a slow drift toward ambiguity. The claim survives not because it is true but because it has been made unfalsifiable — the task is redefined, the noise model adjusted, or the baseline weakened until the claim is no longer comparable to the rebuttal.

What is missing is a criterion that is *effective* in a precise sense borrowed from algebraic geometry and computational algebra. In the theory of tensor decompositions, a criterion for h -identifiability is called effective if it is satisfied in a dense, open subset of the set of rank- h tensors [1]. Genericity is the point: a criterion that fails on a measure-zero set of degenerate cases is still useful, because the degenerate cases can be recognized and handled separately. We argue that falsifiability criteria for quantum-advantage claims should be designed and evaluated the same way — while also flagging, in Section 6, that claim space has no canonical topology or measure, so the dense-open language can only be used heuristically.

The philosophical anchor is the debate over falsifiability in theories of consciousness. Kleiner and Hoel introduced a formal mathematical model of the process of generating observable data from experiments and using that data to generate inferences and predictions onto an experience space, and a reply to that work defended functionalism against the resulting substitution argument [3]. Whatever the merits of that dispute, it demonstrates that falsification

arguments are themselves model-relative and can be contested at the level of the model rather than the conclusion — exactly the situation of quantum-advantage claims, where the model connecting experiment to claim is where unfalsifiability hides.

The practical anchor is the QNFO audit corpus. A full-spectrum audit of Huang, Choi, McClean and Preskill's "The vast world of quantum advantage" (arXiv:2508.05720) subjects both that work's five-keystone framework and QNFO's own prior critiques to symmetric scrutiny, and identifies Theorem 1 of the audited work as the most sophisticated challenge to QNFO's existing critical apparatus [9]. A consilient gap synthesis of the QNFO/QWAV portfolio catalogs 42 gaps across 5 structural categories with a phased roadmap spanning 8 calendar weeks [10]. This paper is a re-entry from that synthesis: it converts the falsification-criteria gap into a concrete, computable instrument.

Our contributions are: (i) a five-criterion falsification framework with explicit scoring rules; (ii) a composite index F with fixed weights and full arithmetic; (iii) worked examples demonstrating decidability and termination; (iv) derived infrastructure statistics connecting the framework to open datasets and platforms [5], [6], [8]; and (v) an explicit list of failure modes and falsification conditions for the framework itself.

2. Background and Related Work

We review the supplied bibliography, which spans algebra, quantum information, philosophy of science, and open-infrastructure projects. Each entry is discussed only in terms of what its supplied summary states; several summaries are truncated, and we say so where that limits the analogy.

[1] Effective identifiability criteria for tensors and polynomials (arXiv:1703.02637v1). This work defines a tensor T in a given tensor space to be h -identifiable if it admits a unique decomposition as a sum of h rank-one tensors, and calls a criterion for h -identifiability effective if it is satisfied in a dense, open subset of the set of rank- h tensors. The paper gives effective h -identifiability criteria for a large class of tensors and then improves them for some cases (the supplied summary is truncated at this point and gives no further detail on the improvements). This is the direct template for our notion of an effective falsification criterion: genericity on a dense open set, with degenerate cases handled separately.

[2] Non-Commutative Gebauer-Moeller Criteria (arXiv:1302.3805v3). For an efficient implementation of Buchberger's Algorithm it is essential to avoid treating as many unnecessary critical pairs or obstructions as possible; in the commutative polynomial ring this is achieved by the Gebauer-Moeller criteria, and the paper adapts these criteria to non-commutative polynomial rings, i.e., free associative algebras (the summary is truncated thereafter and supplies no further technical detail). The audit-relevant lesson is pruning: an audit protocol that must check every conceivable objection is computationally infeasible, and structural criteria that eliminate unnecessary checks in advance are what make an audit tractable.

[3] No Substitute for Functionalism — A Reply to 'Falsification & Consciousness' (arXiv:2006.13664v3). Kleiner and Hoel, in 'Falsification and Consciousness', introduced a formal mathematical model of the process of generating observable data from experiments and using that data to generate inferences and predictions onto an experience space; the resulting substitution argument built on this framework was used to show that any theory of consciousness with independent inference and prediction components (the summary truncates here) is subject to a structural objection, and the reply defends functionalism against this argument. For our purposes the episode establishes two lessons: falsification arguments are model-relative and can be attacked at the level of the model; and a falsification framework must specify which components of a theory are independently inferable and predictable, or it will be vulnerable to the same substitution-style counterattack. We build this requirement into criterion C_2 .

[4] On the unraveling of open quantum dynamics (arXiv:2309.13408v2). The state operator of an open quantum system can generically be represented as the solution of a time-local quantum master equation; unraveling in quantum trajectories offers a dual picture, in which physical indicators are computed as Monte-Carlo averages over a stochastic process (the summary is truncated). This supplies the technical vocabulary for our stochastic-error requirement: many advantage claims depend on noise models whose observable consequences are computed exactly as such Monte-Carlo averages, so a falsification criterion must state which unraveling, or which master-equation equivalent, is part of the claim's specification.

[5] OpenPodcar: an Open Source Vehicle for Self-Driving Car Research (arXiv:2205.04454v2). OpenPodcar is a low-cost, open-source hardware and software autonomous-vehicle research platform built on an off-the-shelf, hard-canopy mobility-scooter donor vehicle, with hardware and software build instructions provided to convert the donor vehicle into a fully autonomous platform; the open platform consists of hardware components (CAD designs, bill of materials, and build instructions — summary truncated) and software. It exemplifies the replication standard our criterion C_4 demands: complete build instructions, not descriptions of builds.

[6] CoReCon: an open, community-powered collection of Reionization constraints (arXiv:2309.06480v1). Constraints on the Universe during and before cosmic reionization are rapidly growing in number but are scattered across inhomogeneous formats, unit systems, and sampling strategies; CoReCon is a Python package providing a growing set of constraints on key physical quantities related to the Epoch of Reionization and a platform for high reproducibility (summary truncated). This is the data-governance model for C_3 and C_4 : constraints — here, audit evidence and baseline data — scattered across formats are effectively unusable; a community-curated, format-normalized collection is the remedy.

[7] Criteria for universality of quantum gates (arXiv:1610.00547v6). This work addresses deciding whether a set of quantum one-qudit gates $\mathcal{S} = \{U_1, \dots, U_n\}$ is universal, providing compact-form criteria leading to a simple algorithm that decides universality of any given gate set in a finite number of steps; for a non-universal $\mathcal{S}$, the criteria indicate what type of gates could be added to make it universal (summary truncated). This is the strongest template in the bibliography for our program: a decision procedure that terminates in finitely many steps, and that is *constructive on failure* — it says not only "no" but "what is missing." Our repair-path design borrows this constructive-failure property.

[8] HPLT 3.0: Very Large-Scale Multilingual Resources for LLMs and MT (arXiv:2511.01066v3). The HPLT initiative provides open, very large, high-quality, richly annotated textual datasets for almost 200 languages; at 30 trillion tokens it is likely the largest generally available multilingual collection of LLM pre-training data, derived from web crawls and accompanied by a complete open-source pipeline for document selection (summary truncated). We use its scale as a concrete calibration point for what "open at scale" means in quantitative terms (Section 4).

[9] QNFO: The Vast World of Quantum Advantage — A Full-Spectrum Audit. This audit subjects both the Huang, Choi, McClean & Preskill (2025) five-keystone framework (arXiv:2508.05720) and QNFO's own prior critiques to symmetric scrutiny. Its key finding, as stated in the supplied summary, is that Theorem 1 is the most sophisticated challenge to QNFO's existing critical apparatus; the summary truncates mid-word, and we draw no further conclusion from it. The methodological commitment — symmetric scrutiny of a target framework *and* of the auditor's own prior critiques — is the epistemic posture this paper adopts.

[10] QNFO: A Consilient Gap Synthesis of the QNFO/QWAV Research Portfolio. This synthesis catalogs 42 gaps across 5 structural categories (Infrastructure, Content, Physics Validation, Governance, Dissemination) with a phased version roadmap v2.0–v2.4 spanning 8 calendar weeks. It supplies the quantitative substrate for Section 4.

3. Methods

3.1 Claim formalization

An advantage claim $\mathcal{K}$ is a tuple

$$\mathcal{K} = (T, D, B, M, \Theta),$$

where T is the computational task, D the quantum device with noise model Θ , B the classical baseline, and M the comparison metric. The claim is *specification-complete* if each component is fixed in advance and cannot be revised after classical rebuttals appear.

3.2 The five falsification criteria

- C_1 (Specification completeness): every element of (T, D, B, M, Θ) is stated with enough precision that two independent laboratories would agree on its instantiation.
- C_2 (Inference–prediction independence): following the structural lesson of [3], the claim must state which quantities are inferred from experiment and which are predicted by theory, and these must be independently checkable; no component may be simultaneously inferred and predicted in a way that shields it from refutation.
- C_3 (Baseline auditability): the classical baseline B is specified with a resource bound, and the best-known classical algorithm's cost is either stated or bounded by a formula a third party can evaluate; where indicators are Monte-Carlo averages over stochastic processes, as in quantum-trajectory unravelings [4], the claimed indicator must carry an explicit sampling-error budget, and the classical comparison must meet the same budget.
- C_4 (Replication independence): the full experimental recipe — hardware, calibration, and analysis pipeline — is open in the sense of [5] and [8]: build instructions, data, and selection pipeline all public, with evidence deposited in a community-curated, format-normalized collection in the spirit of CoReCon [6].
- C_5 (Pre-registered failure threshold): a number $\epsilon \in (0, 1)$ is fixed in advance such that if the observed advantage ratio falls below threshold, the claim is declared refuted, with no post-hoc revision permitted.

3.3 The falsification index

Each criterion C_i receives a score $s_i \in \{0, 0.5, 1\}$: 1 if fully satisfied, 0.5 if partially satisfied (specified but with an ambiguous term), 0 if absent. The weights are fixed ex ante:

$$w_1 = 0.30, \quad w_2 = 0.25, \quad w_3 = 0.20, \quad w_4 = 0.15, \quad w_5 = 0.10,$$

with $\sum_{i=1}^5 w_i = 1$. The falsification index is

$$F(\mathcal{K}) = \sum_{i=1}^5 w_i s_i.$$

Decision rule: a claim with $F \geq 0.8$ is *falsifiable-ready*; $0.5 \leq F < 0.8$ is *conditionally falsifiable* (repairs required); $F < 0.5$ is *structurally unfalsifiable* and should not be published as an advantage claim.

3.4 Termination

Mirroring the finite-step decision procedure of [7], the audit terminates: each s_i is decided by inspecting a finite document set (the claim, its supplement, and the baseline code), so the total procedure is a finite composition of finite checks. If $F < 0.8$, the framework outputs the ordered list of criteria with $s_i < 1$, weighted by w_i , as a

constructive repair path — the analogue of [7]'s indication of which gates to add to a non-universal set.

3.5 Genericity caveat and pruning requirement

Following [1], we would like criteria that hold on a dense, open subset of the space of claims. However, "dense open" is a topological notion with no canonical measure on claim space; we therefore use it only heuristically and require the criteria to hold for *all* well-formed claims, not merely generically. This is a deliberate strengthening over [1]'s standard. Additionally, following the Gebauer-Moeller strategy of eliminating unnecessary critical pairs before they arise [2], the protocol must specify which checks are *not* required, so that audit cost remains finite. A fully universal audit that checks everything is the computational analogue of Buchberger's algorithm without the criteria: correct but unusable.

4. Analysis

All inputs below come from the supplied bibliography entries or are stated design choices of this paper; each is stated with its source, and every arithmetic step is shown.

Input 1 (weights, this paper, Section 3.3). $w_1 = 0.30$, $w_2 = 0.25$, $w_3 = 0.20$, $w_4 = 0.15$, $w_5 = 0.10$. Normalization check:

$$\sum_{i=1}^5 w_i = 0.30 + 0.25 + 0.20 + 0.15 + 0.10 = 1.00.$$

Input 2 (gap counts, from [10]). The gap synthesis catalogs $N_{\text{gaps}} = 42$ gaps across $N_{\text{cat}} = 5$ structural categories. Mean gap density per category:

$$\bar{g} = \frac{N_{\text{gaps}}}{N_{\text{cat}}} = \frac{42}{5} = 8.4 \text{ gaps per category.}$$

If gaps were distributed uniformly, each category would hold 8.4 gaps; since counts are integers, a uniform distribution is impossible, so at least one category deviates from the mean by at least 0.4 gaps. The categories are Infrastructure, Content, Physics Validation, Governance, and Dissemination [10]; the synthesis does not supply per-category counts, so no finer distribution is claimed here.

Input 3 (roadmap duration, from [10]). The roadmap spans versions v2.0 through v2.4 over $W = 8$ calendar weeks, i.e., $N_{\text{ver}} = 5$ version labels. We adopt the *transition* convention: five labels define $N_{\text{ver}} - 1 = 4$ transitions, so the cadence is

$$c = \frac{W}{N_{\text{ver}} - 1} = \frac{8}{4} = 2.0 \text{ weeks per version step.}$$

An alternative convention — dividing by the number of versions rather than transitions — would give $8/5 = 1.6$ weeks per version; we adopt the transition convention as primary because cadence is naturally an interval quantity, and record the alternative explicitly (Appendix A). Equivalently, the mean version-label density is

$$d = \frac{N_{\text{ver}}}{W} = \frac{5}{8} = 0.625 \text{ versions per week,}$$

which is convention-independent. Under the further assumption of uniform pacing across all 42 gaps, the mean time per gap is

$$t_{\text{gap}} = \frac{8 \text{ weeks}}{42 \text{ gaps}} = \frac{8}{42} = \frac{4}{21} \approx 0.1905 \text{ weeks per gap},$$

and converting to days (1 week = 7 days):

$$t_{\text{gap}} = \frac{4}{21} \times 7 = \frac{4}{3} \approx 1.333 \text{ days per gap}.$$

Input 4 (corpus scale, from [8]). The HPLT 3.0 collection contains $T_{\text{tot}} = 30$ trillion tokens, i.e., $T_{\text{tot}} = 30 \times 10^{12} = 3.0 \times 10^{13}$ tokens, covering "almost 200 languages." We take $L = 200$ as an upper-bound language count; since the true count is slightly below 200, per-language averages computed with $L = 200$ are slight *upper bounds* on the true per-language mean:

$$\bar{t} = \frac{T_{\text{tot}}}{L} = \frac{3.0 \times 10^{13}}{200} = 1.5 \times 10^{11} \text{ tokens per language}.$$

That is, 150 billion tokens per language on average. This is a *mean*, not a minimum: the HPLT summary states the data derive from web crawls [8], and web-crawl coverage is language-dependent, so the per-language minimum is certainly far below $\bar{t}$; the summary supplies no per-language counts, so no minimum is computed here.

Input 5 (storage projection assumptions, this paper). Assuming a stated assumption of this paper — not of [8] — of b bytes per token (UTF-8 encodings of multilingual text vary, so we bound it):

- Lower bound at $b = 1$: $3.0 \times 10^{13} \times 1 = 3.0 \times 10^{13}$ bytes = 30 TB.
- Central estimate at $b = 4$: $3.0 \times 10^{13} \times 4 = 1.2 \times 10^{14}$ bytes = 120 TB.
- Upper bound at $b = 4$ plus 25% annotation overhead: $1.2 \times 10^{14} \times 1.25 = 1.5 \times 10^{14}$ bytes = 150 TB.

These are projections under stated byte-per-token assumptions, not measurements of the corpus.

Derivation 6 (pruning benefit, illustrative model). To illustrate why pruning criteria matter, model an unpruned audit as checking all pairwise "objection interactions" among n audit checks:

$$P_{\text{unpruned}} = \frac{n(n-1)}{2}.$$

For $n = 20$ checks: $P_{\text{unpruned}} = \frac{20 \cdot 19}{2} = 190$ pairs. If Gebauer-Moeller-style criteria eliminate a fraction $p = 0.9$ of unnecessary pairs (an illustrative assumption, not a measured value), the pruned cost is

$$P_{\text{pruned}} = 190 \times (1 - 0.9) = 19 \text{ pairs},$$

a factor-10 reduction. The only computed facts are 190 and the proportional reduction $190 \rightarrow 19$ given the assumption $p = 0.9$.

Derivation 7 (worked example of the index). Consider a hypothetical advantage claim $\mathcal{K}_A$ with scores $s_1 = 1$, $s_2 = 0.5$, $s_3 = 1$, $s_4 = 0$, $s_5 = 0.5$ (fully specified task and device; inference–prediction partially independent; baseline stated with bound; no open replication recipe; failure threshold ambiguous). Then:

$$\begin{aligned} F(\mathcal{K}_A) &= (0.30)(1) + (0.25)(0.5) + (0.20)(1) + (0.15)(0) + (0.10)(0.5) \\ &= 0.30 + 0.125 + 0.20 + 0 + 0.05 = 0.675. \end{aligned}$$

Since $0.5 \leq 0.675 < 0.8$, $\mathcal{K}_A$ is conditionally falsifiable; the repair path prioritizes C_4 (weight 0.15, score 0, contribution gap 0.15), then C_2 (gap 0.125), then C_5 (gap 0.05). Total recoverable index if all repairs succeed: $1.00 - 0.675 = 0.325$.

Derivation 8 (second worked example). A claim $\mathcal{K}_B$ with $s_1 = 0, s_2 = 0, s_3 = 0.5, s_4 = 0, s_5 = 0$:

$$F(\mathcal{K}_B) = (0.30)(0) + (0.25)(0) + (0.20)(0.5) + (0.15)(0) + (0.10)(0) = 0.10.$$

Since $0.10 < 0.5$, $\mathcal{K}_B$ is structurally unfalsifiable under our decision rule.

Derivation 9 (necessity of C_1). The largest single weight is $w_1 = 0.30$; a claim scoring $s_1 = 0$ can reach at most

$$F_{\max}^{(s_1=0)} = 1 - 0.30 = 0.70 < 0.8,$$

so specification completeness is *necessary* for falsifiable-ready status. This is a designed property, verified by the arithmetic above.

All quantities in this section are exact or rounded from exact fractions ($8/42 = 4/21$; $\frac{4}{21} \times 7 = \frac{4}{3} \approx 1.333$); none are empirical measurements.

5. Results

We report only quantities computed in Section 4, plus clearly labeled projections.

1. **Index normalization:** $\sum_{i=1}^5 w_i = 1.00$ exactly.
2. **Gap density:** the QNFO/QWAV portfolio's 42 gaps across 5 categories [10] yield $\bar{g} = 8.4$ gaps per category; the distribution is necessarily non-uniform, with at least one category deviating from the mean by at least 0.4 gaps.
3. **Roadmap pacing:** the 8-week v2.0–v2.4 roadmap [10] implies $c = 2.0$ weeks per version transition under the adopted transition convention, or $d = 0.625$ version labels per week (convention-independent); under the alternative per-version convention the figure would be 1.6 weeks per version. Under uniform pacing across all gaps, $t_{\text{gap}} = 4/21 \approx 0.1905$ weeks ≈ 1.333 days per gap.
4. **Open-data scale calibration:** the HPLT 3.0 corpus [8], at 3.0×10^{13} tokens over almost 200 languages, implies a mean per-language volume of at most $\bar{t} = 1.5 \times 10^{11}$ tokens (an upper bound on the mean because the true language count is below 200). As a labeled projection, its storage footprint lies between 30 TB (1 byte/token) and 150 TB (4 bytes/token plus 25% annotation overhead), central estimate 120 TB.
5. **Pruning benefit (projected):** under the stated illustrative assumption $p = 0.9$, a 20-check audit drops from 190 to 19 pairwise checks, a factor-10 reduction; the unpruned count of 190 is exact for $n = 20$.
6. **Worked index values:** $F(\mathcal{K}_A) = 0.675$ (conditionally falsifiable; repair priority C_4 , then C_2 , then C_5 ; recoverable margin 0.325) and $F(\mathcal{K}_B) = 0.10$ (structurally unfalsifiable).
7. **Necessity of C_1 :** any claim with $s_1 = 0$ has $F \leq 0.70 < 0.8$ and therefore cannot be falsifiable-ready.
8. **Protocol status:** the framework is a *design proposal with a genericity goal*, not a validated genericity result. No claim is made that it resolves any measured fraction of real claims; measuring that fraction is an open task (Section 6).

No empirical, simulated, or measured quantities are reported in this paper.

6. Discussion

Limitations. The scoring function is deliberately crude: $s_i \in \{0, 0.5, 1\}$ discards gradations that matter in practice, and the weights w_i are a design choice, not derived from data. A different weight vector could reorder repair priorities; the sensitivity check shows C_1 's necessity is guaranteed whenever $w_1 > 0.20$, and we have not proven robustness for all admissible weights. The decision thresholds 0.5 and 0.8 are likewise conventional. Derivation 2 shows the gap distribution must be non-uniform, but the synthesis [10] supplies no per-category counts, so we cannot identify which category is gap-dense — Governance, for instance, might hold 2 gaps or 12; any per-category claim beyond the mean would be unsupported. Similarly, the per-language mean of Derivation 5 masks the known unevenness of crawl-derived corpora; using it as a per-language *guarantee* would be a serious error, and we do not.

The cadence convention. Derivation 3 exposed a genuine convention choice: $8/4 = 2.0$ weeks per transition versus $8/5 = 1.6$ weeks per version. We chose the transition convention because intervals, not labels, carry scheduling information; a reader who needs per-label density should use $d = 0.625$ versions per week, which is convention-independent. Convention choices embedded silently in audit metrics are exactly the failure mode C_1 is designed to prevent.

Failure modes. Four are foreseeable. First, *gaming*: a claim author can maximize F by writing verbose specifications that satisfy the letter of C_1 while remaining operationally ambiguous; the 0.5 score for "specified but ambiguous" is a judgment call vulnerable to auditor–author disagreement. Second, *genericity leakage*: the dense-open template of [1] does not transfer cleanly, because claim space has no canonical topology or measure; our strengthening to universal applicability may make the criteria too strict for legitimate claims with inherently statistical components. Third, *baseline unboundedness*: if the best classical algorithm's cost is unknown, C_3 can only be scored 0.5 at best, and a genuinely hard baseline problem could keep a valid advantage claim permanently below the ready threshold — a false negative the framework does not currently detect. Fourth, *model-relativity of falsification itself*: the episode documented in [3] shows that a falsification argument can be attacked at the level of its underlying model; an adversary could similarly contest the inference–prediction split demanded by C_2 , and our framework has no meta-criterion to adjudicate such contests. A further failure mode concerns pre-registration: F5-style refutation conditions are only as good as the generosity of the named refutation families, and an author can name a family so narrow that falsification becomes practically impossible. Likewise, C_4 's normalization requirement needs a curator, and curators can be captured by the audited parties, degrading CoReCon-style neutrality [6]; and C_2 's independence demand is representable in the Kleiner–Hoel pipeline formalism [3] but representability is not enforcement.

What would falsify the framework. The framework is falsified if (a) a claim with $F \geq 0.8$ is nevertheless revised post-hoc in a way that evades a pre-registered refutation, showing the criteria are insufficient; or (b) a large fraction of widely accepted advantage claims score $F < 0.5$, showing the thresholds are miscalibrated against community practice. A deeper falsifier of the genericity thesis: if a substantial class of important advantage claims turned out to be *degenerate* in the technical sense — isolated, one-off configurations whose audit resists all generic criteria — then genericity-based design fails and case-by-case adjudication is unavoidable. Both are empirical conditions that the audit infrastructure of [9] and [10] could in principle test: if the 42-gap portfolio [10] were re-audited under symmetric scrutiny (the QNFO method of [9]) and found to contain gaps that no C_i check addresses, the protocol's coverage claim would be refuted at the specific points of omission.

Open questions. (i) Can the weights be learned from a corpus of historical advantage claims and their outcomes, rather than fixed ex ante? (ii) Can a genericity theorem for the framework be proved over a defined space of advantage claims, analogous to the dense-open-set results of [1]? (iii) What is the empirical value of the pruning

fraction p for real audits, replacing the illustrative 0.9? (iv) Can the repair-path output be made constructive in the strong sense of [7], where the repair is not merely indicated but algorithmically generated? (v) Does the Monte-Carlo indicator structure of unravelings [4] introduce irreducible statistical uncertainty into C_1 's noise-model specification? (vi) Can the baseline registry modeled on CoReCon's format unification [6] achieve community adoption in a field with competitive incentives? (vii) Does the symmetric-scrutiny method of [9] extend to auditing the auditor, and at what recursion depth does it stabilize? We note also that the supplied summaries of [1], [2], [3], [4], [5], [6], [7], and [8] are truncated, which limits the depth of the analogies we may legitimately draw from them; obtaining full texts is a prerequisite for any deeper engagement.

7. Conclusion

We have constructed a five-criterion falsification framework for quantum-advantage claims, with a composite index F , fixed weights, explicit decision thresholds, guaranteed termination, and a constructive repair path. The framework draws its structure from effective identifiability criteria [1], critical-pair elimination [2], finite-step universality decision with constructive failure [7], and the model-relativity lessons of the falsification debate in consciousness studies [3], and it is grounded operationally in the stochastic-indicator vocabulary of open quantum dynamics [4] and the open-replication standards exemplified by open hardware platforms [5], constraint registries [6], and open large-scale data pipelines [8]. It is instantiated against two existing audit artifacts [9], [10]. The quantitative spine of the paper — 8.4 gaps per category, 2.0 weeks per roadmap transition (or 0.625 version labels per week), ≈ 1.333 days per gap under uniform pacing, 1.5×10^{11} tokens per language as an upper bound on the corpus mean, and worked index values $F = 0.675$ and $F = 0.10$ — was derived with fully shown arithmetic from stated inputs. The framework's own falsification conditions are stated, and its principal weakness — the conventionality of its weights and thresholds — is flagged as the first target for empirical calibration.

References

- [1] Effective identifiability criteria for tensors and polynomials. arXiv:1703.02637v1. <https://arxiv.org/abs/1703.02637v1> [2] Non-Commutative Gebauer-Moeller Criteria. arXiv:1302.3805v3. <https://arxiv.org/abs/1302.3805v3> [3] No Substitute for Functionalism -- A Reply to 'Falsification & Consciousness'. arXiv:2006.13664v3. <https://arxiv.org/abs/2006.13664v3> [4] On the unraveling of open quantum dynamics. arXiv:2309.13408v2. <https://arxiv.org/abs/2309.13408v2> [5] OpenPodcar: an Open Source Vehicle for Self-Driving Car Research. arXiv:2205.04454v2. <https://arxiv.org/abs/2205.04454v2> [6] CoReCon: an open, community-powered collection of Reionization constraints. arXiv:2309.06480v1. <https://arxiv.org/abs/2309.06480v1> [7] Criteria for universality of quantum gates. arXiv:1610.00547v6. <https://arxiv.org/abs/1610.00547v6> [8] HPLT 3.0: Very Large-Scale Multilingual Resources for LLMs and MT. Mono- and Bi-lingual Data, Multilingual Evaluation, and Pre-Trained Models. arXiv:2511.01066v3. <https://arxiv.org/abs/2511.01066v3> [9] QNFO: The Vast World of Quantum Advantage — A Full-Spectrum Audit [10] QNFO: A Consilient Gap Synthesis of the QNFO/QWAV Research Portfolio