Claim against Measurement: Statistical Artefacts in Quantum Error Mitigation Benchmarks

Dominik Köster
Technical University of Applied Science Regensburg
Regensburg, Germany
dominik.koester@othr.de

Wolfgang Mauerer
Technical University of Applied Science Regensburg
Siemens AG, Foundational Technology
Regensburg/Munich, Germany
wolfgang.mauerer@othr.de

***Abstract*—Quantum Error Mitigation (QEM) is widely regarded as a plausible bridge from Noisy Intermediate Scale Quantum (NISQ) devices to Fault Tolerant Quantum Computers (FTQC). Yet the empirical studies used to assess the effectiveness of QEM techniques on concrete problems have received comparatively little scrutiny with respect to the validity of their conclusions. We systematically review 81 recent QEM papers using an eight-criterion framework covering statistical rigour, reproducibility, and reporting quality. Among the 59 papers for which statistical evidence is applicable, only 15 (25%) use inferential methods, while 25 (42%) report uncertainty only descriptively, without testing whether the claimed effects are statistically supported.**

To demonstrate the consequences of these omissions, we use Zero-Noise Extrapolation (ZNE) as a representative and widely used case study and identify two compounding sources of artefacts in current QEM benchmarks. First, we observe *parameter sensitivity*: in a 132-configuration sweep, implicitly assumed choices such as scale factors, extrapolation method, and hardware calibration are not merely incidental but *active*, with variations changing conclusions from statistically significant improvement to statistically significant degradation. Second, we identify a *drift-induced effectiveness illusion*: in a 72-hour longitudinal study on real hardware, temporal drift alone can make the same ZNE configuration exhibit an effect size more than three times as large, depending solely on when it is executed, and also drastically reduces the effective number of independent observations. These findings do not imply that QEM methods are intrinsically unsound; rather, they show that current evaluation practice can make mitigation performance appear more robust than the evidence warrants. We therefore propose minimum reporting standards for QEM evaluations, including explicit parameter documentation, robustness checks, longitudinal drift assessment, and inferential statistical testing with effect-size reporting.

***Index Terms*—quantum error mitigation, benchmarking, statistical artefacts, zero-noise extrapolation, hypothesis testing, reproducibility, NISQ**

I. Introduction

We remain in the NISQ era of quantum computing [1]–[4], still some distance from FTQC architectures with sufficiently many qubits, low error rates, and operational Quantum Error Correction (QEC) [5], [6]. Although QEC is clearly the long-term route to reliable quantum computation, present implementations still incur substantial qubit overheads [7], [8]. In this intermediate regime, QEM has emerged (among hardware-efficient problem formulations [9]–[11] or target design automation techniques [12], [13]) as a pragmatic approach for reducing the impact of noise [14] on current NISQ devices without implementing full error correction [15]–[18]. Techniques such as ZNE [15], [16], Probabilistic Error Cancellation (PEC) [15], [19], and Clifford data regression [20] have shown promise in improving the accuracy of quantum computations. Beyond dedicated benchmarking studies, QEM techniques are increasingly adopted as standard components in broader quantum computing research – from variational quantum eigensolvers and quantum structure simulations [21], [22] via optimisation [23]–[27] and machine learning [28]–[31] to quantum simulation of many-body systems [32] and demonstrations of evidence for quantum utility [33] – where framework-default parameters are silently inherited and their influence is not examined. Artefacts arising from such implicitly assumed parameter choices therefore propagate beyond QEM evaluation into any higher-level result that relies on mitigation as a building block. However, the empirical evaluation of QEM techniques themselves frequently rests on experiments whose statistical foundations are not always commensurate with the strength of the conclusions drawn from them.

Previous work has already identified important statistical challenges in the evaluation of QEM techniques, including the need for explicit hypothesis testing [34]. In this work, however, our systematic review of 81 recent QEM papers (Section IV) shows that such concerns remain largely unaddressed in practice: most papers rely on descriptive reporting rather than inferential statistical evidence. This is not merely a matter of presentation, but one with direct experimental consequences. Reported QEM improvements are often small, since current NISQ hardware constrains experiments to shallow circuits and shot budgets limited by cost, placing many studies in regimes where the true effect is modest at best [35], [36]. In such settings, shot noise, hardware drift, and implicitly assumed parameter choices can each be sufficient to alter the experimental conclusion. Without the use of careful inferential testing, a comprehensive awareness of influence factors and boundary conditions, as well as the adequate availability of all ingredients required to perform faithful reproductions or replications, neither the original study nor any subsequent replication can reliably distinguish genuine mitigation effects from statistical or experimental artefacts.

In this paper, we identify two compounding sources of

artefacts in QEM evaluation, each independently capable of producing misleading results. First, a systematic replication study shows that documented parameters usually represent only a small portion of the full space of possible parameters for an experiment: the choices experiments do not explicitly specify – such as scale factors, folding strategy, or calibration snapshot – are *active*, meaning their variation can shift the interpretation of the outcome of an experiment from significant improvement to significant worsening compared to an unmitigated baseline. Second, a longitudinal hardware study reveals that temporal drift alone can produce large variation in apparent ZNE effectiveness on the same device at different access times; given the dominant use of cloud services with indeterministic batch access for many quantum experiments, this presents a substantial challenge for both reproducibility and interpretation of empirical results. We focus on ZNE with Richardson extrapolation as a representative case, as it is the most widely reported QEM technique in our corpus, and propose a minimum reporting checklist for QEM evaluations.

Concretely, our contributions are as follows:

- We provide a systematic eight-criterion review of 81 QEM papers, revealing that the majority of papers lack inferential statistics and drift control.
- We introduce a replication pipeline and case study demonstrating that implicitly assumed ZNE parameters are *active*: their variation shifts experimental outcomes across a large fraction of tested configurations.
- A longitudinal drift study shows that temporal hardware drift produces a *drift-induced effectiveness illusion*, where apparent ZNE effectiveness varies substantially across identical experiments at different times.
- We derive minimum reporting standards for QEM benchmarks to avoid incorrect claims and misinterpretations of QEM experiments.

II. Related Work

Cai *et al.* [18] provide a comprehensive review of QEM techniques, including their theoretical foundations, practical implementations, and performance analysis. Takagi *et al.* [37] derive fundamental sampling-overhead bounds under global depolarising noise, generalised by Quek *et al.* [38] for local noise. Krebsbacher *et al.* [39] derive variance bounds for Richardson extrapolation, guiding scale factor selection to minimise amplification. On the statistical side, Saki *et al.* [34] propose a hypothesis-testing framework for QEM evaluation, while Li *et al.* [40] survey the use of statistical methods in quantum software testing, finding that formal statistical methods are similarly rare in that domain. Moguel *et al.* [41] review quantum benchmarking methods, proposing a quantum experiment guideline extending established best practices from classical software benchmarking. On reproducibility, Senapati *et al.* [42], [43] highlight the reproducibility challenges in quantum machine learning, being device variability and temporal drift. Hirasaki *et al.* [44] demonstrate temporal fluctuations in superconducting qubits, yielding different measurements over time points.

III. Background

A. Zero-Noise Extrapolation (ZNE)

ZNE, first introduced by Li and Benjamin [16] and Temme *et al.* [15], is a widely used QEM technique that estimates the noise-free result of a quantum computation by extrapolating results obtained at amplified noise levels. Given the noise scale λ, the result of the smallest error rate on a circuit is given by the expectation value $E(\lambda_1)$ [18], [45]. By artificially increasing the noise to levels $\lambda_1 < \lambda_2 < \cdots < \lambda_K$ – called *scale factors* – we can fit a model to extrapolate the noise-free expectation value $E(0)$. Common amplification strategies include pulse stretching [15], [21], unitary folding [46], and gate-level folding [47]. To compute the extrapolated value, besides standard approaches like linear [16], polynomial and exponential extrapolation [17], [45], a technique called *Richardson extrapolation* is a commonly employed method [15], [17], [39], [46] that fits a polynomial of degree $K-1$ through K data points. The zero-noise limit is in this case estimated [18] as $\hat{E}(0) = \sum_{k=1}^{K} c_k E(\lambda_k)$, where $\sum_{k=1}^{K} c_k = 1$, and coefficients c_k are determined by Lagrange interpolation.

Since $\mathrm{Var}(\hat{E}_{\mathrm{ZNE}}) = \sum_{k=1}^{K} c_k^2 \,\mathrm{Var}(\hat{E}(\lambda_k))$, a convenient pre-experiment bound on variance amplification is $\sum_{k=1}^{K} |c_k|$ [18], [39] – computable from the scale factors alone, without running any circuits. For the widely used default set $\{1, 3, 5\}$ [45], [46], $\sum_{k=1}^{K} |c_k| = 3.5$, whereas $\{1, 1.1, 1.25, 1.5\}$ – motivated by arguments that finer spacing reduces extrapolation error [39] – yields $\sum_{k=1}^{K} |c_k| = 681$: a $194\times$ difference in the variance bound. Scale factor choice alone can therefore dominate the variance budget of a ZNE experiment – motivating its treatment as an active parameter in Section VI.

B. Statistical Methods

Our approach, implemented in a reproducible pipeline (link in PDF) [48], relies on standard tools from frequentist statistics to quantify how much a QEM technique improves performance. Unlike domains such as clinical trials or psychology, QEM evaluation has no established domain-specific statistical standards: there is no agreed effect-size threshold for claiming practical mitigation, no canonical hypothesis test, and no prescribed power requirement. In the absence of such conventions, we adopt the well-validated general-purpose tools from empirical science [49]–[51], which are commonly used in related empirical fields. Despite frequentist hypothesis testing (e.g., paired t-tests and p-values) and effect-size estimation (e.g., Cohen's d) being standard, well-validated tools in empirical science, these methods remain rare in quantum computing. As they are textbook knowledge [50], [51], we only briefly review their essential properties below, especially to fix notation of the approaches in the context of quantum error mitigation.

a) Paired t-test: A paired t-test compares differences between two groups: Given n independent repetitions producing raw errors $|\epsilon_{\mathrm{raw},i}|$ and mitigated errors $|\epsilon_{\mathrm{mit},i}|$, the paired difference $\delta_i = |\epsilon_{\mathrm{raw},i}| - |\epsilon_{\mathrm{mit},i}|$ captures per-repetition improvement. The paired t-test evaluates the null hypothesis

$H_0 : \mu_\delta = 0$ via $t = \bar{\delta}/(s_\delta/\sqrt{n})$. We classify each configuration as *significantly better* ($p < 0.05$, $d > 0$), *not significant* ($p \geq 0.05$), or *significantly worse* ($p < 0.05$, $d < 0$) based on a 5% significance level (while using a prescribed explicit significance level is known to be exhibit a number of issues [52], we stick to this approach as it is common practice in the considered references, and conclusion stability can only be assessed when the same approach as in the original work is employed). As a non-parametric alternative that does not assume normality of the paired differences, we compute the Wilcoxon signed-rank test for every configuration.

b) Effect-Size Measures: We use two complementary effect-size quantities. **Cohen's** $d = \bar{\delta}/s_\delta$ [49] is the standard effect-size measure in empirical sciences such as clinical trials or psychology, with well-established conventions $|d| = 0.2/0.5/0.8$ (small/medium/large) [49] and $1.2/2.0$ (very large/huge) [53]; positive d means QEM reduces, negative d means it increases error. These thresholds were, however, calibrated for domains where effect sizes are bounded by natural variability rather than a controllable experimental parameter. In shot-based quantum experiments, $s_\delta \propto 1/\sqrt{n_{\text{shots}}}$, so d scales with shot count: at $n_{\text{shots}} = 4096$, a modest improvement of $\Delta = 0.18$ already yields $d \approx 6$, far above the established thresholds. Absolute d values are not comparable to those conventions and we use d primarily as a relative metric for cross-configuration comparison. Because no single effect-size measure is universally agreed on for QEM evaluation, we additionally report **Cliff's** $\delta = (N_> - N_<)/n$ – where $N_>$ and $N_<$ count paired differences $\delta_i > 0$ and $\delta_i < 0$ – as a non-parametric, distribution-free alternative in $[-1, +1]$ that is insensitive to this shot-count inflation and directly reflects the probability of a genuine directional improvement across repetitions.

IV. Systematic Review

To evaluate the current state of statistical reporting in QEM related papers and to identify common pitfalls and areas for improvement, we systematically reviewed 81 publications published over half a decade (2022–2026), collected via Google Scholar, arXiv, IEEE digital libraries, and forward/backward citation tracking from the review by Cai *et al.* [18]. We used queries like "quantum error mitigation", "zero-noise extrapolation", "probabilistic error cancellation" in combination with terms like "algorithm", "drift", or "experiment" to identify relevant papers (the full list is provided in the reproduction package). Each paper was evaluated based on eight criteria detailed in Table I that assesses the presence and quality of statistical reporting in the experimental evaluation of QEM techniques. Each criterion is rated as *adequate* (clear, specific evidence), *partial* (mentioned but incomplete), *missing*, or *not applicable*. The criteria themselves are derived from a priori expert knowledge, typical desiderata in the literature, and requirements documented in reviews [54]–[56] or existing work on quantum reproducibility [48], [57], [58].

a) Review process: All raters are among the paper authors; the process follows established patterns for lightweight semi-formal reviews [59], and aims at balancing required manual human effort and comprehensiveness/completeness of coverage. Two raters jointly reviewed all 81 papers against the criteria. Two additional raters each independently rated a (different) subset of 15 papers. Disagreements among the raters were then resolved by consensus discussion.

TABLE I
Eight-criterion analysis framework. Each paper is rated adequate, partial, missing, or N/A.

ID	Criterion	Key Question(s)
C1	Sample Size	How many shots/circuits? Size justified?
C2	Variance	Error bars, CIs, or variance reported?
C3	Stat. Evidence	Inferential or descriptive evidence for claims?
C4	Drift Control	Temporal hardware drift accounted for?
C5	Overhead	Classical/quantum overhead quantified?
C6	Noise Model	Noise model validated or discussed?
C7	Reproducibility	Code, data, or sufficient detail?
C8	Neg. Results	Failure cases or limitations reported?

Additionally, we established a comparison baseline using automatic text processing: A regular expression scanner matched textual evidence against patterns targeting key statistical terms – for example, p-value reporting phrases, mentions of confidence intervals, or software repository URLs – and an LLM (Claude Opus 4.6) filtered candidate ratings for known false positives such as physical noise probabilities misidentified as statistical p-values, or framework citations misidentified as reproduction packages. Automated ratings agreed with the human consensus in 77% of applicable paper-criterion pairs. Full scan logs, rating sets, and a per-paper LLM evidence report are provided in the reproduction package. Overall, we believe the approach provides a sound and sufficiently large-scale basis that leads to generalisable conclusions.

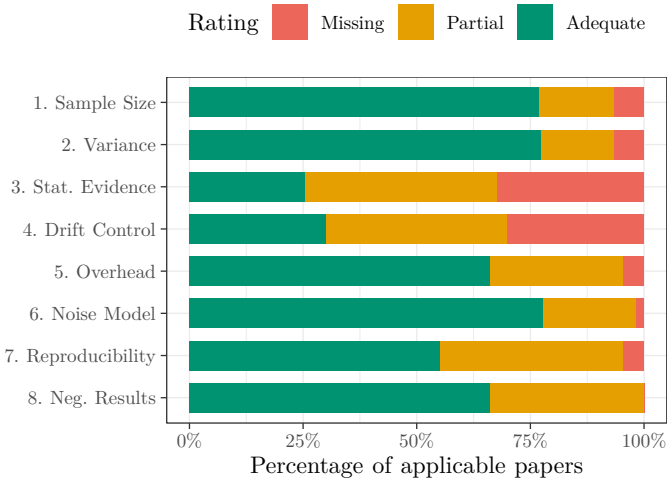

Fig. 1. Summary of the systematic review results across the eight criteria.

Figure 1 shows the observed compliance of the 81 papers

with each of the eight criteria. Five criteria are well addressed, with over 60% adequate reporting: Sample Size (C1, 77%), Variance (C2, 77%), Noise Model (C6, 78%), Overhead (C5, 66%), and Negative Results (C8, 66%). Reproducibility (C7, 55%) is in the middle, but plays a crucial role as it is often the only way to confirm and build upon reported results. The two criteria with the lowest compliance are Drift Control (C4, 30%) and Statistical Evidence (C3, 25%).

Of the 59 applicable papers for statistical evidence, only 15 employ inferential statistical methods: hypothesis tests, Bayesian inference, bootstrap-based comparisons, or scaling analysis with uncertainty quantification. The remaining 42% report uncertainty only descriptively (error bars, standard deviations, improvement factors), and 19 papers (32%) do not provide statistical evidence.

The "partial" rating on C3 deserves explanation and credit to the respective studies. These papers report variance, quantitative improvement metrics, accuracy thresholds, or bootstrap error bars, all forms of descriptive statistical evidence common in quantum experiments, but do not use them for inferential comparison.

QEM improvements are often small [35], [36]: in regimes where noise, drift, and parameter choice can each be sufficient to shift the outcome of an experiment, simple descriptive evidence alone is insufficient to confirm genuine effects. The two empirical analyses below demonstrate what practical consequences arise from a failure – as is omnipresent in the literature – to provide more complete descriptions and a more rigorous statistical analysis. Both artefact sources are independently capable of producing misleading evaluations, and the two factors compound when present simultaneously.

V. The Reproduction Parameter Space & Extraction Pipeline

Our analysis shows C3 and C4 have the smallest compliance, yet matter most when parameter choices and timing can influence or even determine the outcome of an experiment. Before we demonstrate this experimentally, let us formalise the structure of the problem. Figure 2 illustrates the relation between a quantum (software) experiment and an attempt to reproduce or replicate the findings. Given the overall variability and currently fast-paced change in quantum hardware, rapidly changing software environments [55], [58], [60], and other factors, an exact reproduction is rarely possible even if the original software artefacts are available, which leads to an effectively different set of empirical parameters. The space of relevant parameters for quantum experiments can be divided into three categories of outcomes for a given configuration: regions of significant improvement, regions of significant worsening, and regions with no statistically significant change. In typical experiments, only a specific subset of possible values is tested per parameter (highlighted as the orange circle), where some are explicitly specified (θ_{doc}) and others are left unspecified (θ_{undoc}). A parameter is *inert* if varying it does not change the experimental conclusion, and *active* if it does.

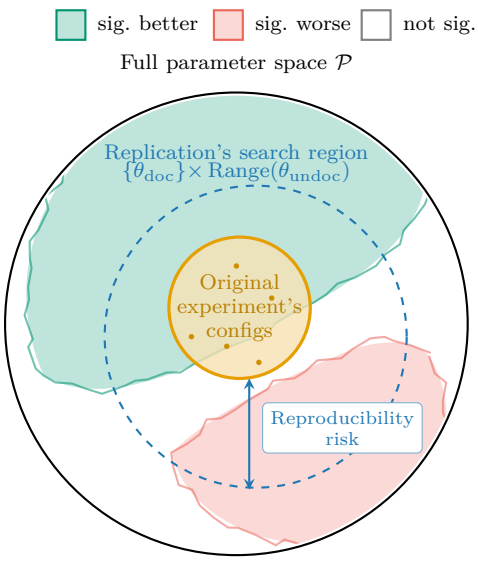

Fig. 2. The full parameter space $\mathcal{P}$ of a QEM experiment, with regions of improvement (green), no improvement (grey), and worsening (red). Studies typically test a small subset of parameter configurations (orange).

A. Parameter Space $\mathcal{P}$

A QEM experiment requires many choices beyond the circuit and observable. We formalise the *reproduction parameter space* for a ZNE experiment as: $\mathcal{P} = \mathcal{H} \times \mathcal{C} \times \mathcal{Q} \times \mathcal{F} \times \mathcal{E} \times \mathcal{S}$ As $|\mathcal{P}| > 10^4$, and the size of parameter space for reproductions can easily grow to hundreds of reasonable configurations.

TABLE II
Reproduction Parameter Space.

Axis	Name	Examples
$\mathcal{H}$	Hardware/Backend	QPU vendor/model, noise model, calibration snapshot
$\mathcal{C}$	Circuit	Qubit count, depth, transp. level
$\mathcal{Q}$	Shots & reps	Shot count, number of repetitions
$\mathcal{F}$	Folding	Local-left, local-right, global
$\mathcal{E}$	Extrapolation	Linear, polynomial, exponential, Richardson
$\mathcal{S}$	Scale factors	$\{1,3,5\}$, $\{1,2,3\}$, $\{1,1.5,\ldots,3\}$

B. Reproduction Pipeline

Following established terminology, we distinguish *reproduction*, where a different team re-runs the same experimental artefacts from *replication*, where a different team uses a different experimental setup, yet addresses the same question as an existing piece of research. Our case study addresses both aspects: we use the original circuit specification but systematically vary the unspecified parameters across a wider range than explored in the original study. To systematically explore the parameter space, we apply a four-stage pipeline to a target paper: (1) **Parameter Extraction**: extract all documented parameters θ_{doc} and identify unspecified ones θ_{undoc}. Also categorise any parameters θ_{amb} that are either ambiguously

mentioned or very likely given the context, but not explicitly stated. (2) **Parameter-Space Sampling**: sample θ_{undoc} and θ_{amb} across reasonable values. Reasonable is context-dependent, but in most cases, refers to the most commonly used values in literature. The result is a set of configurations $\{p_1, p_2, \ldots, p_n\} \subset \mathcal{P}$. (3) **Statistical Analysis**: run n_{reps} repetitions for each configuration and perform statistical tests e.g. conduct paired t-tests, and compute Cohen's d effect size. (4) **Artefact Classification**: Classify each configuration based on the statistical analysis. If the claimed improvement only holds for a subset of $\mathcal{P}$, it is not generalisable and relies on *active* parameters. This pipeline is designed to be extensible to any QEM by replacing the ZNE specific parameters ($\mathcal{F}, \mathcal{E}, \mathcal{S}$).

VI. The Parameter Space is Active

To test whether implicitly assumed parameters θ_{undoc} and the statistical test gap (C3) have practical consequences – that is, whether an experiment with seemingly the same prerequisites can yield different outcomes – we demonstrate how varying different axes of $\mathcal{P}$ can shift the experimental outcome.

A. Case Study: *Khan* et al. *(2024)*

1) Parameter Extraction: To evaluate the effectiveness of different QEM techniques for NISQ devices, Khan *et al.* [61] apply dynamic decoupling, twirled readout error extraction and ZNE to four-qubit Quantum Trotter Circuits (QTC) whose gate structure ([61], Algorithm 1) corresponds to the Trotterisation of a transverse-field Ising chain ($\mathcal{C}$). The paper executes these circuits on IBM Kyoto and Osaka machines (both now retired [62]), and compares them with an ideal QASM Simulation ($\mathcal{H}$) [63]. Error rates, qubits properties and architecture (at the time of execution) are documented, as well as pseudo-code for the QEM procedures. Unfortunately, parameters are neither explicitly nor implicitly (via a reproduction package) available. This leaves us with the unknown $\mathcal{F}$, $\mathcal{E}$, $\mathcal{S}$, and $\mathcal{Q}$.

2) Parameter-Space Sampling: For the missing parameters, we applied common default values: 4096 measurement shots and transpilation level 1 [60]. Additionally, for ZNE, folding-from-left strategy, Richardson extrapolation method, and scale factors $\{1, 3, 5\}$ [18], [45], [46] are employed. From this baseline, we sweep values for reasonable alternatives and execute the configuration on calibration snapshots for IBM Kyoto and Osaka. Inspecting the Kyoto snapshot, we find ECR gate errors of all 144 qubits are near hundred percent, producing a noise-floor output. Given this is not the exact environment the work conducted its simulations, we added a depolarising noise model matching the error rate of the original work. Additionally, we added the reported QASM simulator for ideal results (negative control, as QEM cannot improve ideal results).

Ten configurations arise for a one-at-a-time sweep of the five parameters plus the baseline. We vary each axis independently while keeping the others at their default values and tests them against three Trotter depths (TC1, TC3 , TC5) specified in the original paper. In Summary 11 configurations $\times$ 4 backends $\times$ 3 Trotter depths = 132 configurations are tested in total.

3) Statistical Analysis: For each configuration, we run $n_{\text{reps}} = 200$ independent repetitions, then conduct paired t-tests on the per-repetition absolute error $|\epsilon|$ relative to the ideal expectation value, and compute Cohen's d as an effect-size measure. We apply a significance threshold of $\alpha = 0.05$.

4) Artefact Classification: Figure 3 shows the results of the parameter-space sampling. As expected, the ideal simulator shows a significant worsening with a median Cohen's d of -0.95: this serves as a negative control, confirming that ZNE cannot improve results that are already ideal, since any introduced noise amplification only degrades an already noiseless expectation value. In the last three columns we see neutral and mostly non-significant results for the fake IBM Kyoto (median Cohen's d of -0.03), which is also expected given the fully faulty ECR gates – their errors drive the output to the totally mixed state, yielding near-zero expectation values regardless of the circuit. The other two backends show a more positive pattern. The depolarising simulation (Kyoto error rates from the paper) shows significant improvement in $29/33$ configurations with a median Cohen's d of $+5.95$. Meanwhile the Osaka snapshot has a less positive improvement of median $+2.23$. As expected, the improvement is more pronounced for the depolarising simulation given its more idealised noise model. In contrast, the Osaka snapshot is closer to real hardware and therefore shows a more mixed pattern of results with other error patterns. The non-improving results and smaller positive improvements are largely driven by the scale factors $\mathcal{S}$, due to variance amplification, and the exponential extrapolation method $\mathcal{E}$. We motivated the amplification of variance by the scale factors in Section III. Different extrapolation methods assume different functional forms of noise decay and have different sensitivities to the noise model [17], [46]. The exponential model assumes an exponential decay of the noise, motivated by global depolarising noise, where $E(\lambda) \propto (1-p)^{\lambda n}$ decays exponentially [17]. However, real hardware noise is often more complex and may not follow this idealised model, as in our case. The extrapolation method $\mathcal{E}$ is therefore an active parameter that can significantly affect the results, and its sensitivity to the noise model can lead to different conclusions about the effectiveness of our ZNE. Cliff's δ values (right half of each cell) closely mirror Cohen's d patterns, confirming directional results are not an artefact of the shot-count inflation: Wherever d is strongly positive, δ is near +1, and wherever d is negative, δ is near -1.

a) Multiple-comparisons perspective: Given that we apply repeated identical tests that lead to a near-certain probability for false positives, we applied the standard Bonferroni and Benjamini-Hochberg corrections to all paired t-test p-values. Of the 107 out of 132 uncorrected significant results (58 better, 49 worse), 103 survive the strict Bonferroni threshold (57 better, 46 worse) and 106 survive Benjamini-Hochberg. Only four results are dropped by Bonferroni, all corner cases near the noise floor. As improvements *and* degradations, are robust to multiplicity correction, this reinforces that experimental outcomes depend on implicitly assumed parameter values, and degradations observed on the noiseless and FakeKyoto

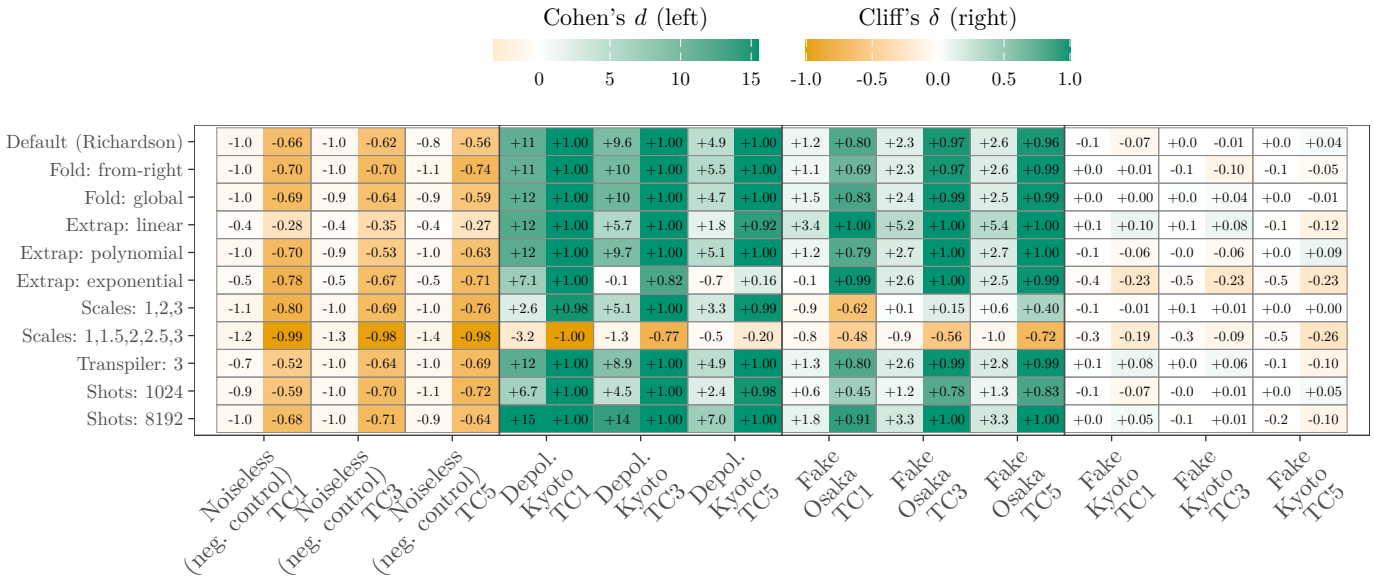

Fig. 3. Parameter space heatmap for 4 backends × 3 Trotter depths (separated by vertical lines), one-at-a-time sweep from defaults. Each cell is split: the left half shows Cohen's d (colour scale left), the right half shows Cliff's δ (colour scale right). Green = ZNE significantly reduces error, yellow = ZNE significantly increases error (α=0.05), white = not significant.

backends are genuine rather than statistical artefacts.

b) Hardware calibration and effect size: Khan *et al.* report expectation values and variances (Table III in [61]), but it is not specified whether these variances represent per-run estimator uncertainty, an average of shot-variance across runs, or a different measure. The number of independent repetitions per configuration is also not reported. This means we cannot directly verify the statistical significance of the reported improvements, nor recover paired differences needed to compute Cohen's d. We estimate $\hat{d}$ for the original paper by computing $\Delta = E_{\text{ZNE}} - E_{\text{raw}}$ from the reported values and divide by the simulated $\sigma_{\text{improvement}}$ as proxy.

Since Khan *et al.* used real IBM hardware, the IBM Osaka calibration snapshot is the closest available approximation for comparison. Our replication yields $d = +1.16$ at TC1 – a moderate but statistically significant improvement. For the original paper [61], $E_{\text{raw}} = 0.728$, $E_{\text{ZNE}} = 0.814$, $E_{\text{ideal}} = 0.828$; dividing the improvement $\Delta = 0.086$ by our $\sigma_{\text{improvement}} = 0.016$ gives $\hat{d} \approx +5.3$, more than four times our replication value. Under idealised depolarising noise (same circuit, same configuration), the same calculation yields $d = +11.3$. The original effect size therefore depends critically on the specific hardware calibration ($\mathcal{H}$) at the time of the experiment, which is no longer accessible. We therefore cannot distinguish whether $d \approx 1.16$, $d \approx 5.3$, or $d \approx 11.3$ is the better characterisation of the true effect, let alone assess its robustness across noise models [42], [64].

To check if simulation-based findings transfer to quantum hardware, we ran the default configuration on the 156-qubit IBM Marrakesh processor [65] for TC1 and TC3. The Heron-based machine has a median two-qubit error rate of 0.241%, roughly four times lower than the 0.947% reported for IBM Kyoto. Our experiment yields $d = -0.75$ for TC1, but $d = +0.55$ for TC3. The negative effect at TC1 arises as low error rates leaves the shallow circuit (six ECR gates) nearly ideal ($|\epsilon_{\text{raw}}| = 0.013$), so ZNE's variance amplification (2.7×) outweighs any potential correction, pushing the mitigated value *further* from ideal. At TC3 (18 ECR gates), sufficient error accumulates for ZNE to provide genuine improvement. This not only confirms hardware dependence of the outcome, but also reveals a regime where low-error hardware makes ZNE counterproductive for shallow circuits.

c) Shot-count variance and statistical power: Figure 4(a) confirms the expected $d \propto \sqrt{n_{\text{shots}}}$ scaling: on the depolarising model and FakeOsaka, d grows monotonically with shot count, while FakeKyoto stays near zero (no genuine signal) and the ideal simulator is fixed at $d \approx -0.95$. A bootstrap power analysis (see Figure 4(b)) shows that $n_{\text{reps}} \geq 20$ suffices for 80% power at moderate effects, while large depolarising effects need as few as $n_{\text{reps}} = 5$; the near-zero FakeKyoto effect never reaches 80% power even at $n_{\text{reps}} = 200$. Low shot counts ($\mathcal{Q}$) and few repetitions therefore risk both false negatives for genuine effects and an inability to confirm the absence of improvement where none exists.

B. Case Study: Desdentado et al. (2025)

So far, we have assumed that multiple runs of an identical configuration yield the (essentially) identical result. Desdentado *et al.* [66] provide a case where this assumption breaks: they observe a temporal confound whose structure is consistent with calibration drift, as studied below in Section VII. Their work proposes an algorithm to estimate the ideal shot count for a given quantum circuit to achieve the best possible

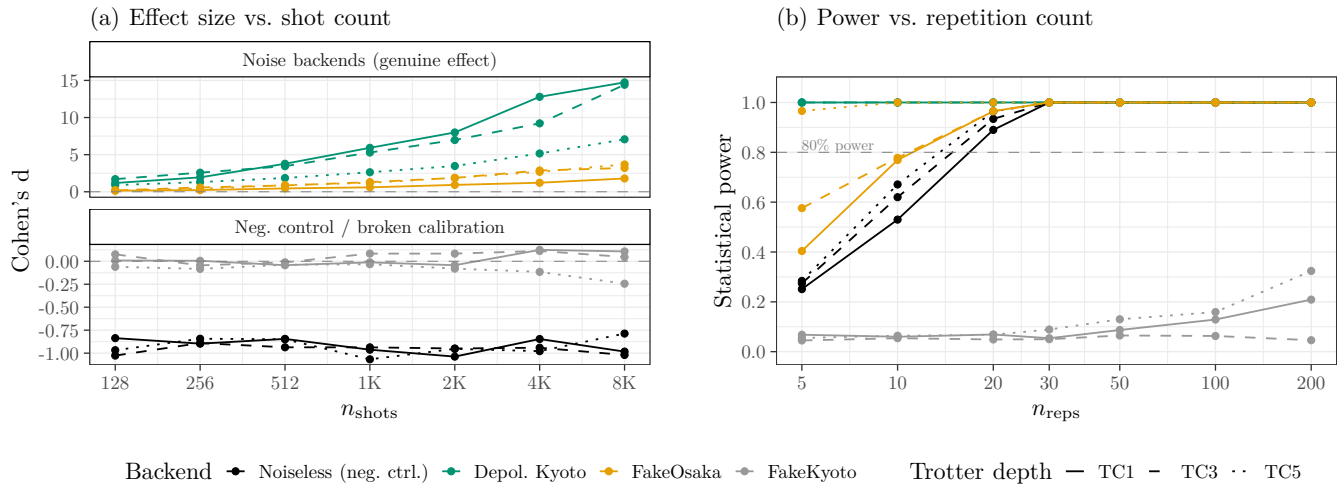

Fig. 4. Sensitivity analysis. (a) Cohen's d vs. shot count; d scales with $\sqrt{n_{\text{shots}}}$ for genuine effects. (b) Statistical power vs. repetition count (1,000 bootstrap resamples); 80% power requires $n_{\text{reps}} \geq 20$ for moderate effects.

result under *shot noise* – the statistical sampling variance that decreases as $1/\sqrt{n_{\text{shots}}}$ when more measurement shots are taken. Hardware noise (*e.g.*, gate errors, calibration drift, ...) is, in contrast, not reduced by increasing shot count.

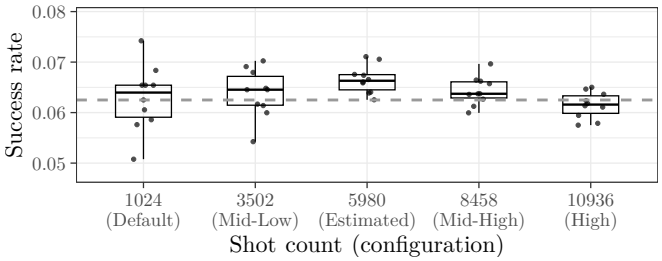

Fig. 5. Shot estimation results from Desdentado *et al.* [66] for a five qubit Grover circuit on the IBM Brisbane. The proposed shot estimation yields the best target state probability, while higher shot counts drift to the noise floor.

The paper proposes an algorithm to estimate the ideal shot count and tests five configurations for a five-qubit Grover circuit on IBM Brisbane. If finds an improvement of 0.37% for two, and 0.53% for four target states. Their proposed shot estimation yields the best target state probability for all their circuits while lower and higher shot counts yield worse results. Uniquely, the paper provides a complete reproduction package that allows us to analyse exact hardware results from published data (see Figure 5). Circuit depth and hardware error rate are dominant challenges: IBM Brisbane (Eagle r3) reports a median two-qubit gate error rate of 0.77% error rate through the last IBM calibration snapshot. Alternate reported median ECR error rates ranging from 0.762% [67], 0.79% [68] to 0.832% [69]. The transpiled circuit uses 904 ECR gates, which leads to a total circuit fidelity of $\mathcal{F} = (1 - 0.00762)^{904} \approx 0.09\%$ with the most optimistic error rate.

The low circuit fidelity explains a near-random target state probability. Desdentado *et al.* (see Figure 5) report probability degradation after exceeding the estimated ideal shot count, with a non-monotonic pattern. We only show results for one circuit, but the pattern applies to all. This contradicts the shrinking variance of shot noise with increasing shot count, which should result in a monotonic improvement in target state probability as shot noise decreases. Job identifiers reveal that shot count groups were executed in sequence over a 20 minute time span, except for a gap of 13 minutes between parts of group 8458 and the entire group 10936. Because shot count is confounded with execution time and the group size is small ($n = 10$), inter-group differences may also reflect calibration drift. Particularly given the non-monotonic pattern (the last complete executed shot group, ideal shot estimation, coincides with lowest error), we cannot distinguish hardware drift, shot count or pure chance as cause.

This illustrates how temporal drift can masquerade as parameter effect with un-randomised execution order. To isolate drift as an independent factor, we conduct a controlled longitudinal experiment measuring its impact on ZNE effectiveness.

Cross-vendor fragility To test whether these findings transfer beyond IBM hardware, we executed the same circuit on the 54-qubit IQM Euro-Q-Exa [70] machine. At TC1 (6 CZ gates at λ_1), the circuit retains 27.7% of the ideal expectation value ($\bar{E}(\lambda_1) = 0.290$ vs. $E_{\text{ideal}} = 0.980$). At λ_3 and λ_5, the expectation value is near the noise floor ($E(\lambda_3) = 0.024$) or negative ($E(\lambda_5) = -0.105$). TC3 (18 CZ) retains only 10.9% of the ideal expectation value on Euro-Q-Exa, compared to 113% on IBM Marrakesh (coherent over-rotation). The circuit depth at which ZNE "works" is therefore a property of the circuit–hardware pair: cross-vendor studies are necessary for any generalisable claim. As the signal is near a fully mixed state at TC3 on the Euro-Q-Exa, TC1 is used for the longitudinal drift study in the next section.

VII. The Contingency of Time

Even with fixed parameters, replications can depend on *when* an experiment is run, given time-varying HW properties.

We find only 28% of papers address *hardware drift* (C4). Given that superconducting processors exhibit calibration fluctuations for minutes to hours [42], [44], this raises concerns.

A. Experimental Design

A longitudinal experiment on the 54-qubit IQM Euro-Q-Exa system available at LRZ [70] tests the effect of temporal drift using the Khan *et al.* [61] QTC at depth TC1 (6 CZ at λ_1). We chose the smallest circuit, because the expectation value $\bar{E}(\lambda_1)$ of TC3 is already dominated by noise on this hardware. To minimise parameter confounds, we use default configurations from Section VI-A2 and run an experiment every 30 minutes for different time periods. This allows us to observe temporal variability from hardware drift. Four days with two 12 hour periods and one 48 hour weekend period with a total of 147 time points with $n_{\text{reps}} = 30$ independent runs allow for observing short-term and long-term drift patterns.

We address three questions: Is drift reproducible across sessions (RQ1); do raw expectation values at λ_1 exhibit temporal autocorrelation inconsistent with the i.i.d. assumption of paired tests (RQ2); does drift cause per-time-point ZNE effect size to vary substantially such that *identical* experiments at different times yield different effectiveness *conclusions* (RQ3).

B. Results

Across all sessions, $\bar{E}(\lambda_1) \approx 0.28{-}0.30$, corresponding to approximately 29% of the ideal expectation value E_{ideal}, $E(\lambda_3) \approx 0.02$ (near noise floor), and $E(\lambda_5)$ is consistently *negative* (mean -0.10), which is inconsistent with depolarising noise decay and indicating coherent over-rotation past the zero-crossing. Figure 6 shows the time series across all three sessions alongside the associated ZNE effect size per time point. Table III summarises the drift severity metrics.

TABLE III
DRIFT SEVERITY ON EURO-Q-EXA. η^2: FRACTION OF TOTAL VARIANCE BETWEEN TIME POINTS. r_1: LAG-1 AUTOCORRELATION OF $\bar{E}(\lambda_1)$. n_{EFF}: EFFECTIVE INDEPENDENT REPETITIONS (NOMINAL n_{REPS}=30). d RANGE: COHEN'S d; ALL PER TIME-POINT.

Session	TPs	η^2	r_1	n_{eff}	d range
Day 1 (12 h)	25	0.35	0.55	3.5	6.1–10.9
Day 2 (12 h)	25	0.20	0.21	4.9	7.2–12.9
Weekend (48 h)	97	0.55	0.83	1.8	3.3–11.3

a) Three sessions, three drift patterns (RQ1): The top part of Fig. 6 reveals qualitatively different dynamics: Day 1 shows a slight upward drift with a discrete step-change at $t \approx 9.5$ h (that is, asymmetric across scale factors: λ_1 affected, $\lambda_{3,5}$ insensitive). Contrary to the first session, day 2 exhibits a gradual downward trend to the lowest measured expectation value. Session 3 was conducted over the weekend, and reveals what appears to be an overnight recalibration shift in the second night (between a Saturday and Sunday), which is absent in the first night. Between days 1 and 2, $E(\lambda_3)$ crosses zero ($+0.024 \rightarrow -0.011$), which the negative Richardson coefficients convert into a large change in the ZNE estimate at around $t = 18$ h. Different patterns across sessions indicate that drift is not a stable, reproducible process, but a non-stationary phenomenon that can yield different outcomes for ZNE effectiveness at different times. We extended this experiment to a full seven-day window (163 scheduled hours, Figure 7): the trace shows qualitatively different overnight behaviour, and the baseline level after a 43-hour outage (red gap) differs visibly from before, confirming that drift persists over days and is not resolved by recalibration.

b) Drift is pervasive and severe (RQ2): In the 48-hour weekend study, more than half of the total variance in the raw signal is attributable to the time of measurement ($\eta^2 = 0.55$ (see Table III)), and consecutive time points are strongly correlated ($r_1 = 0.83$). Each measurement carries information about the next, violating independence assumptions of standard paired tests. Additionally the autocorrelation is asymmetric: $r_1 = 0.17$ in the first 24 hours vs. 0.91 in the second, coinciding with a visible upward shift during the second night ($t = 39{-}43$ h) which results in a higher $E(\lambda_1)$ up to 0.346 compared to most of the other time points with expectation values below 0.300. The 12-hour sessions confirm drift at lower severity ($\eta^2 = 0.20{-}0.35$, $r_1 = 0.21{-}0.55$).

c) The drift-induced effectiveness illusion (RQ3): Figure 6 (b) shows per-time-point Cohen's d of ZNE vs. raw across all sessions. The d values vary substantially over time (3.3–12.9) (although inflated by the high measurement precision at $n_{\text{shots}} = 4096$). What matters is not effect size, but relative variation over time. For instance, in the weekend session, the experiment yields $d = 3.3$ at one time point and $d = 11.3$ twelve hours later: a $3.4\times$ difference in apparent ZNE effectiveness. For comparison, switching from the Osaka calibration snapshot to a fundamentally different depolarising noise model, produces only a ratio of 2.7. *Temporal drift on a single back-end can produce larger variation in apparent ZNE effectiveness than changing the entire noise model.*

In our experiment, the effect is significantly positive at every time point ($d > 3$ throughout), as expected: $\bar{E}(\lambda_1)$ is well above the noise floor and the high shot count inflates d (Section III-B). The illusion manifests not as a sign reversal, but as uncontrolled *magnitude variation*: for moderate effects ($d \approx$ 1–2) commonly reported in QEM hardware evaluations [33], this temporal variation alone is sufficient to determine the binary conclusion of statistical significance. Additionally, the within-time-point Intraclass Correlation Coefficient (ICC) reduces the n_{reps}=30 nominal repetitions to as few as $n_{\text{eff}} = 1.8$ effective independent observations (Table III), following Kish's design-effect formula $n_{\text{eff}} = n/(1+(n-1)\cdot\text{ICC})$ [71]. Note that this ICC is distinct from the between-time-point r_1: it is computed via one-way ANOVA *within* each time point. A genuinely moderate improvement that appears highly significant at n_{reps}=30 becomes non-significant once this effective sample size is accounted for. Two tests of the same ZNE configuration – one in the morning, one the next day – can give contradictory conclusions. Neither would be wrong given measured results, but neither would paint an accurate picture. This limits generalisability of typical quantum (software) experiments that are often based on

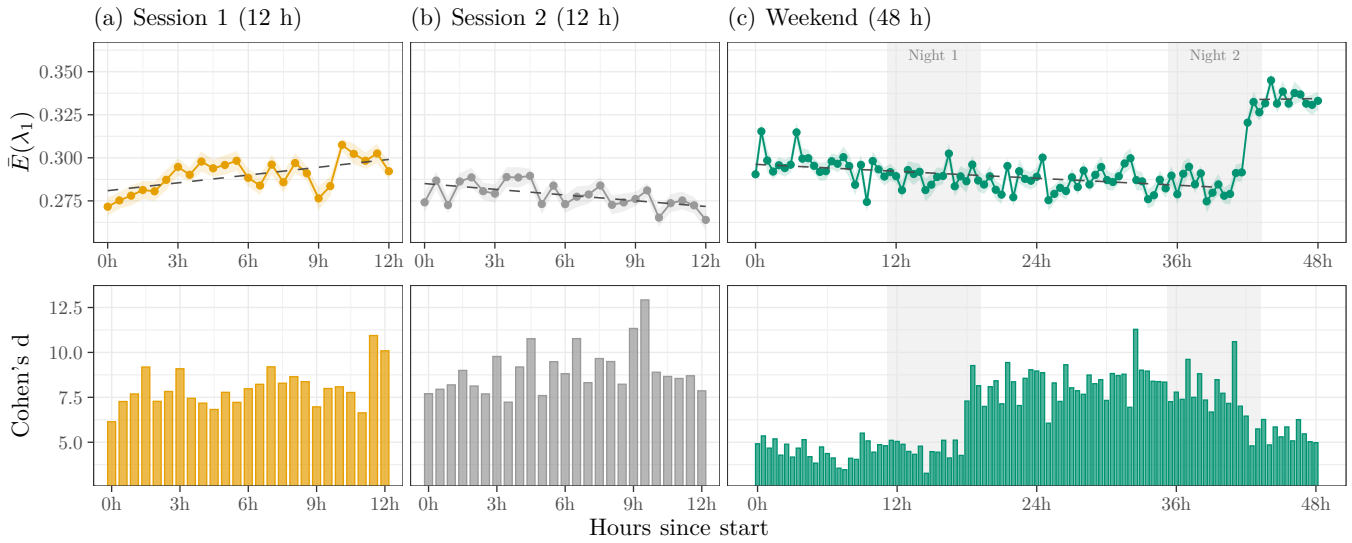

Fig. 6. Longitudinal drift study on IQM Euro-Q-Exa (147 time points across 72 hours in three independent sessions). Top: $\bar{E}(\lambda_1)$ averaged values with 95% CI vs. elapsed time; each session exhibits a qualitatively different drift pattern (step-change, gradual decline, overnight shift). Bottom: per time point Cohen's d (ZNE vs. raw): d varies around $3\times$ across the weekend, this is a *drift-induced effectiveness illusion*.

immediately consecutive hardware runs and limited repetition counts, especially given the financial commitment for such experiments. IBM `ibm_brussels` (TC3, 12-hour session, Figure 8) shows $\bar{E}(\lambda_1)$ drifts by approximately 0.14 within a single session, confirming that drift-induced effectiveness illusion is not specific to IQM.

VIII. Discussion

A. The compound nature of QEM artefacts

Our analysis reveals that a favourable ZNE outcome requires *both* conditions to hold: the chosen parameter configuration must fall in the improving region of $\mathcal{P}$ (Section VI), and the hardware calibration at the time of measurement must produce a stable, noise-discernible expectation value (Section VII).

Critically, these two artefact sources interact and amplify each other. A paper that reports a single configuration at one point in time risks confounding both: the observed outcome may be an artefact of a fortunate parameter choice or a favourable calibration window. The Khan *et al.* reproduction illustrates this compounding: even within the moderate Osaka noise model, d ranges from -0.75 on IBM Marrakesh to $+11.3$ under idealised depolarising noise (Section VI). Adding temporal drift introduces an additional $3.4\times$ variation in apparent ZNE effectiveness within 48 hours on the same device.

This compound structure explains why few papers survive both challenges: a result must be robust against parameter choices *and* stable under drift. The rarity of meeting all review criteria suggests that the current literature overestimates ZNE reliability not because of method flaws, but because evaluation does not control for confounders.

B. Scope and limitations

We focus on ZNE with Richardson extrapolation, the most widely used QEM method in our corpus. Artefact sources – parameter sensitivity and temporal drift – are method-agnostic and apply to any quantum experiment and hardware platform, but patterns may differ. Our design does not capture interaction effects; a full factorial design could reveal additional effects.

Internal validity: Paired t-tests assume approximate normality of differences. While robust to moderate non-normality, smaller configurations may violate this assumption. We therefore computed the Wilcoxon signed-rank test for every configuration, where the tests agree on 129 of 132 outcomes, (all on the FakeKyoto backend at marginal p-values). Bonferroni and Benjamini–Hochberg corrections on 132 parameter-space configurations leave qualitative conclusions unchanged.

External validity: Drift patterns may differ on other devices, architectures, or time scales. The Khan et al. reproduction uses noise-model snapshots rather than live hardware for the parameter sweep, which may not capture all device effects.

Construct validity: Our eight-criterion review framework is a pragmatic operationalisation of statistical rigour. Consequently, other framings could yield different compliance rates.

C. Recommendations

Based on our analysis, we propose a reporting checklist, ordered by implementation effort:

1) Document all **Active Parameters**, including calibration snapshot ($\mathcal{H}$), shot count and repetition count ($\mathcal{Q}$), transpilation seed, or method-specific hyper-parameters.
2) Report **Inferential Statistics**: *At least* pair claimed improvements with a hypothesis test and effect-size measure.
3) Provide a **Reproduction Package** containing all code, data, transpiled circuits, calibration snapshots, etc. as explicit baseline for independent verification.
4) Ensure **Result Robustness** by evaluating at least a small grid of configurations (*e.g.*, multiple scale factors, backends, ...) to avoid misleading results.

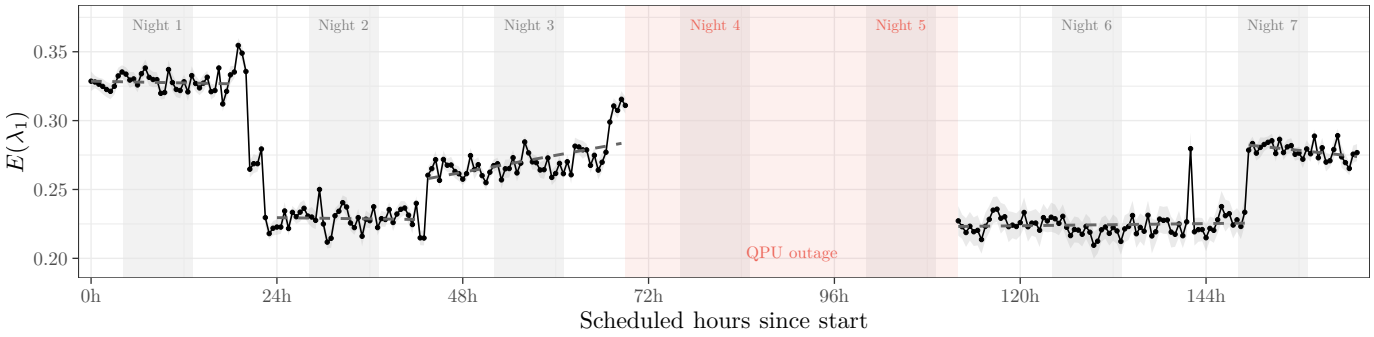

Fig. 7. Seven-day longitudinal drift study on IQM Euro-Q-Exa (163 scheduled hours, 2026-04-08 to 2026-04-15). $\bar{E}(\lambda_1)$ per time point with 95% CI; dashed lines show piecewise-linear interpolation through the mean of each night-bounded daytime cluster (pre- and post-outage separately); grey bands mark local night (21:00–06:00 CEST) and the red gap a 43-hour QPU maintenance outage. $\bar{E}(\lambda_1)$ post-outage shows HW recalibration does not restore a stable baseline.

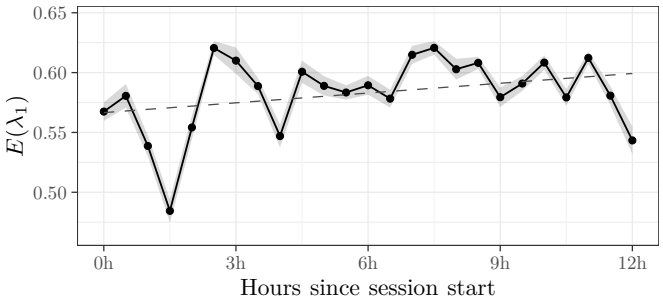

Fig. 8. Twelve-hour drift session on IBM `ibm_brussels` (TC3, 25 time points). $\bar{E}(\lambda_1)$ with 95% CI ribbon and linear trend (dashed); the dotted line marks the ideal value ($E_{\text{ideal}} = 0.846$). Despite a fixed hardware and parameter configuration, $\bar{E}(\lambda_1)$ drifts across the session.

5) Quantify **Temporal Stability** by distributing HW experiments over in time, or randomise execution order to deconfound drift from parameter effects.

IX. Conclusion

Our systematic review of 81 QEM papers reveals predominantly descriptive reporting: only 25% of the papers employ inferential statistics, and only 30% address hardware drift. A two-stage analysis shows this methodological gap has substantial consequences: Prevailing practices yield different interpretations of the same technique when uncontrolled factors (*e.g.*, parameter choice, execution time) vary.

First, we show implicitly assumed ZNE parameters, including scale factors, extrapolation method, and hardware calibration, are *active*: on two physics-based noise models, variations shift the conclusion from significant improvement to significant degradation in 8 of 66 configurations (12%), while 55 (83%) show significant improvement – depending solely on which parameters were implicitly assumed. We found that on IBM Marrakesh QPUs, ZNE can even be counterproductive for shallow circuits whose unmitigated output is close to ideal. Second, our 72-hour longitudinal study on IQM Euro-Q-Exa shows temporal drift alone induces a 3.4-fold variation in apparent ZNE effectiveness, exceeding the 2.7-fold variation observed when changing the entire noise model. The same drift reduces the n_{reps}=30 nominal repetitions to as few as n_{eff}=1.8 effective independent observations, substantially weakening the evidential basis of nominally repeated measurements.

Parameter sensitivity and temporal drift compound on real hardware. Their interaction challenges the validity of QEM benchmarks that do not include inferential testing, robustness analysis, and drift control. An apparent QEM improvement may reflect a favourable point in parameter space, a favourable calibration window, or both. This is not an issue of the methods, but relates to their *use*. We believe this is partly because especially use-case-centric empirical measurements are often carried out by domain specialists who may lack deeper training in quantum computing. It is reasonable for them to rely on standard settings and avoid involvement with low-level details. We believe the core of the problem is a software challenge: Providing good mechanisms, abstractions and reference patterns would alleviate "users" from having to deal with such details. We hope our reproduction pipeline, together with the proposed reporting standards, will support more robust QEM evaluation (and results with improved practical credibility, as well as scientific soundness) as the field progresses towards practical quantum advantage.

Data Availability

Code, data, and plotting scripts are available in our reproduction package that can build the paper including analysis result. HW calibration snapshots and logs allow for analysing parameters and drift without machine access.

Acknowledgments The authors gratefully acknowledge the use of the quantum system Euro-Q-Exa, co-funded by the EuroHPC JU, BMFTR (grant 13N16690), and the Bavarian State Ministry of Science and the Arts, operated by the Leibniz Supercomputing Centre (LRZ) in Garching, Germany, for providing the computational resources for this work. We acknowledge partial support by the German Research Foundation, grant MA 9739/1-1, and the High-Tech Agenda of the Free State of Bavaria. We also acknowledge partial support by the European Union (Project Reference 101083427), the European Funds for Regional Development (EFRE) (Project Reference 20-3092.10-THD-105), by the European Regional Development Fund (ERDF) and by the Free State of Bavaria as part of the project AIM-SMEs (Grant No. 2506-014-3.2), co-funded by the European Union.

References

[1] J. Preskill, "Quantum Computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018, arXiv:1801.00862 [quant-ph]. [Online]. Available: http://arxiv.org/abs/1801.00862

[2] F. Greiwe, T. Krüger, and W. Mauerer, "Effects of imperfections on quantum algorithms: A software engineering perspective," in *IEEE International Conference on Quantum Software (QSW)*. IEEE, 2023, pp. 31–42. [Online]. Available: https://doi.org/10.1109/QSW59989.2023.00014

[3] S. Thelen, H. Safi, and W. Mauerer, "Approximating under the influence of quantum noise and compute power," in *IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2024, pp. 274–279. [Online]. Available: https://doi.org/10.1109/QCE60285.2024.10291

[4] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea *et al.*, "Noisy intermediate-scale quantum algorithms," *Rev. Mod. Phys.*, vol. 94, p. 015004, Feb 2022. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.94.015004

[5] J. Preskill, "Fault-tolerant quantum computation," Dec. 1997, arXiv:quant-ph/9712048. [Online]. Available: http://arxiv.org/abs/quant-ph/9712048

[6] ——, "Beyond NISQ: The Megaquop Machine," *ACM Transactions on Quantum Computing*, vol. 6, no. 3, pp. 18:1–18:7, Apr. 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3723153

[7] M. Beverland, V. Kliuchnikov, and E. Schoute, "Surface code compilation via edge-disjoint paths," *PRX Quantum*, vol. 3, no. 2, p. 020342, May 2022, arXiv:2110.11493 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2110.11493

[8] C. Gidney, M. Newman, P. Brooks, and C. Jones, "Yoked surface codes," Dec. 2023, arXiv:2312.04522 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2312.04522

[9] L. Schmidbauer and W. Mauerer, "SAT strikes back: Parameter and path relations in quantum toolchains," in *Proceedings of the IEEE International Conference on Quantum Software (QSW)*. IEEE, 2025, pp. 1–12. [Online]. Available: https://doi.org/10.1109/QSW67625.2025.00021

[10] L. Schmidbauer, K. Wintersperger, E. Lobe, and W. Mauerer, "Polynomial reduction methods and their impact on QAOA circuits," in *IEEE International Conference on Quantum Software (QSW)*, 2024, pp. 35–45. [Online]. Available: https://doi.org/10.1109/QSW62656.2024.00018

[11] L. Schmidbauer, E. Lobe, I. Schaefer, and W. Mauerer, "Its quick to be square: Fast quadratisation for quantum toolchains," *ACM Transactions on Quantum Computing*, Mar. 2026, just Accepted. [Online]. Available: https://doi.org/10.1145/3800943

[12] S. Thelen and W. Mauerer, "Predict and conquer: Navigating algorithm trade-offs with quantum design automation," in *IEEE International Conference on Quantum Computing and Engineering (QCE)*. Los Alamitos, CA, USA: IEEE Computer Society, 2025, pp. 591–602. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/QCE65121.2025.00071

[13] R. Wille, L. Berent, T. Forster, J. Kunasaikaran, K. Mato *et al.*, "The mqt handbook : A summary of design automation tools and software for quantum computing," in *2024 IEEE International Conference on Quantum Software (QSW)*, 2024, pp. 1–8.

[14] S. R. Maschek, J. Schwittalla, M. Franz, and W. Mauerer, "Make some noise! measuring noise model quality in real-world quantum software," in *Proceedings of the IEEE International Conference on Quantum Software (QSW)*. IEEE, 2025, pp. 1–11. [Online]. Available: https://doi.org/10.1109/QSW67625.2025.00010

[15] K. Temme, S. Bravyi, and J. M. Gambetta, "Error mitigation for short-depth quantum circuits," *Physical Review Letters*, vol. 119, no. 18, p. 180509, Nov. 2017, arXiv:1612.02058 [quant-ph]. [Online]. Available: http://arxiv.org/abs/1612.02058

[16] Y. Li and S. C. Benjamin, "Efficient variational quantum simulator incorporating active error minimization," *Physical Review X*, vol. 7, p. 021050, 2017.

[17] S. Endo, S. C. Benjamin, and Y. Li, "Practical quantum error mitigation for near-future applications," *Physical Review X*, vol. 8, p. 031027, 2018.

[18] Z. Cai, "Quantum error mitigation," *Reviews of Modern Physics*, vol. 95, no. 4, 2023.

[19] E. v. d. Berg, Z. K. Minev, A. Kandala, and K. Temme, "Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors," *Nature Physics*, vol. 19, no. 8, pp. 1116–1121, Aug. 2023, arXiv:2201.09866 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2201.09866

[20] P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, "Error mitigation with Clifford quantum-circuit data," *Quantum*, vol. 5, p. 592, Nov. 2021, arXiv:2005.10189 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2005.10189

[21] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow *et al.*, "Error mitigation extends the computational reach of a noisy quantum processor," *Nature*, vol. 567, pp. 491–495, 2019.

[22] E. F. Dumitrescu, A. J. McCaskey, G. Hagen, G. R. Jansen, T. D. Morris *et al.*, "Cloud quantum computing of an atomic nucleus," *Physical review letters*, vol. 120, no. 21, p. 210501, 2018.

[23] L. Schmidbauer, E. Lobe, I. Schaefer, and W. Mauerer, "It's quick to be square: Fast quadratisation for quantum toolchains," *ACM Transactions on Quantum Computing*, vol. 7, no. 2, p. 46, 2026. [Online]. Available: https://doi.org/10.1145/3800943

[24] A. Lucas, "Ising formulations of many np problems," *Frontiers in Physics*, vol. 2, 2014. [Online]. Available: http://dx.doi.org/10.3389/fphy.2014.00005

[25] T. Krüger and W. Mauerer, "Out of the Loop: Structural Approximation of Optimisation Landscapes and non-Iterative Quantum Optimisation," *Quantum*, vol. 9, p. 1903, Nov. 2025. [Online]. Available: https://doi.org/10.22331/q-2025-11-06-1903

[26] L. Schmidbauer, C. A. Riofrío, F. Heinrich, V. Junk, U. Schwenk *et al.*, "Path matters: Industrial data meet quantum optimization," in *IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2025, pp. 2101–2111. [Online]. Available: https://doi.org/10.1109/QCE65121.2025.00230

[27] M. Schönberger, I. Trummer, and W. Mauerer, "Quantum-inspired digital annealing for join ordering," *Proc. VLDB Endow.*, vol. 17, no. 3, p. 511524, Nov. 2023. [Online]. Available: https://doi.org/10.14778/3632093.3632112

[28] M. Schuld, I. Sinayskiy, and F. Petruccione, "An introduction to quantum machine learning," *Contemporary Physics*, vol. 56, no. 2, pp. 172–185, 2015.

[29] M. Franz, T. Winker, S. Groppe, and W. Mauerer, "Hype or heuristic? quantum reinforcement learning for join order optimisation," in *IEEE International Conference on Quantum Computing and Engineering (QCE)*, vol. 01, 2024, pp. 409–420.

[30] M. Schuld and F. Petruccione, *Machine Learning with Quantum Computers*, ser. Quantum Science and Technology. Springer Cham, 2021.

[31] P. Wittek, *Quantum Machine Learning: What Quantum Computing Means to Data Mining*. Boston: Academic Press, 2014.

[32] I.-C. Chen, B. Burdick, Y. Yao, P. P. Orth, and T. Iadecola, "Error-mitigated simulation of quantum many-body scars on quantum computers with pulse-level control," *Physical Review Research*, vol. 4, no. 4, p. 043027, 2022.

[33] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg *et al.*, "Evidence for the utility of quantum computing before fault tolerance," *Nature*, vol. 618, no. 7965, pp. 500–505, Jun. 2023. [Online]. Available: https://www.nature.com/articles/s41586-023-06096-3

[34] A. A. Saki, A. Katabarwa, S. Resch, and G. Umbrarescu, "Hypothesis Testing for Error Mitigation: How to Evaluate Error Mitigation," Jan. 2023, arXiv:2301.02690 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2301.02690

[35] V. Russo, A. Mari, N. Shammah, R. LaRose, and W. J. Zeng, "Testing Platform-Independent Quantum Error Mitigation on Noisy Quantum Computers," *IEEE Transactions on Quantum Engineering*, vol. 4, pp. 1–18, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10219054/

[36] O. G. Maupin, A. D. Burch, B. Ruzic, C. G. Yale, A. Russo *et al.*, "Error mitigation, optimization, and extrapolation on a trapped ion testbed," *Physical Review A*, vol. 110, no. 3, p. 032416, Sep. 2024, arXiv:2307.07027 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2307.07027

[37] R. Takagi, S. Endo, S. Minagawa, and M. Gu, "Fundamental limits of quantum error mitigation," *npj Quantum Information*, vol. 8, no. 1, p. 114, Sep. 2022. [Online]. Available: https://www.nature.com/articles/s41534-022-00618-z

[38] Y. Quek, D. Stilck França, S. Khatri, J. J. Meyer, and J. Eisert, "Exponentially tighter bounds on limitations of quantum error mitigation," *Nature Physics*, vol. 20, no. 10, pp. 1648–1658, Oct. 2024. [Online]. Available: https://www.nature.com/articles/s41567-024-02536-7

[39] M. Krebsbach, B. Trauzettel, and A. Calzona, "Optimization of Richardson extrapolation for quantum error mitigation," *Physical Review A*, vol. 106, no. 6, p. 062436, Dec. 2022. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.106.062436

[40] Y. Li, M. Shao, J. Zhao, and Q. Wang, "A methodological analysis of empirical studies in quantum software testing," 2026, arXiv:2601.08367 [quant-ph].

[41] E. Moguel, J. A. Parejo, A. Ruiz-Cortés, J. Garcia-Alonso, and J. M. Murillo, "Quantum software experiments: A reporting and laboratory package structure guidelines," May 2024, arXiv:2405.04192 [cs]. [Online]. Available: http://arxiv.org/abs/2405.04192

[42] P. Senapati, Z. Wang, W. Jiang, T. S. Humble, B. Fang *et al.*, "Towards Redefining the Reproducibility in Quantum Computing: A Data Analysis Approach on NISQ Devices," in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, vol. 01, Sep. 2023, pp. 468–474. [Online]. Available: https://ieeexplore.ieee.org/document/10313593/

[43] P. Senapati, S. Y.-C. Chen, B. Fang, T. M. Athawale, A. Li *et al.*, "PQML: Enabling the Predictive Reproducibility on NISQ Machines for Quantum ML Applications," in *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, vol. 01, Sep. 2024, pp. 1413–1424. [Online]. Available: https://ieeexplore.ieee.org/document/10821454/

[44] Y. Hirasaki, S. Daimon, T. Itoko, N. Kanazawa, and E. Saitoh, "Detection of temporal fluctuation in superconducting qubits for quantum error mitigation," *Applied Physics Letters*, vol. 123, no. 18, p. 184002, Nov. 2023. [Online]. Available: https://doi.org/10.1063/5.0166739

[45] R. Majumdar, P. Rivero, F. Metz, A. Hasan, and D. S. Wang, "Best practices for quantum error mitigation with digital zero-noise extrapolation," Jul. 2023, arXiv:2307.05203 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2307.05203

[46] T. Giurgica-Tiron, Y. Hindy, R. LaRose, A. Mari, and W. J. Zeng, "Digital zero noise extrapolation for quantum error mitigation," *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 306–316, 2020.

[47] L. Hour, M. Go, and Y. Han, "Improving Zero-noise Extrapolation for Quantum-gate Error Mitigation using a Noise-aware Folding Method," May 2024, arXiv:2401.12495 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2401.12495

[48] W. Mauerer and S. Scherzinger, "1-2-3 reproducibility for quantum software experiments," in *IEEE International Conference on Software Analysis, Evolution and Reengineering*, 2022, pp. 1247–1248.

[49] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[50] L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, and G. Tutz, *Statistik: Der Weg zur Datenanalyse*. Berlin, Heidelberg: Springer, 2023. [Online]. Available: https://link.springer.com/10.1007/978-3-662-67526-7

[51] J. L. Devore, "Probability and statistics for engineering and the sciences," 2008.

[52] R. L. Wasserstein and N. A. Lazar, "The asa statement on p-values: context, process, and purpose," pp. 129–133, 2016.

[53] S. S. Sawilowsky, "New effect size rules of thumb," *Journal of Modern Applied Statistical Methods*, vol. 8, no. 2, pp. 597–599, 2009.

[54] J. M. Murillo, J. Garcia-Alonso, E. Moguel, J. Barzen, F. Leymann *et al.*, "Quantum software engineering: Roadmap and challenges ahead," *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 5, May 2025. [Online]. Available: https://doi.org/10.1145/3712002

[55] C. Carbonelli, M. Felderer, M. Jung, E. Lobe, M. Lochau *et al.*, *Challenges for Quantum Software Engineering: An Industrial Application Scenario Perspective*. Springer Nature Switzerland, 2024, p. 311335. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-64136-7_12

[56] T. Yue, W. Mauerer, S. Ali, and D. Taibi, *Challenges and Opportunities in Quantum Software Architecture*. Springer Nature Switzerland, 2023, p. 123. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-36847-9_1

[57] I. M. Veiga and E. Hänggi, "Reproducible builds for quantum computing," 2025. [Online]. Available: https://arxiv.org/abs/2510.02251

[58] V. Gierisch and W. Mauerer, "Qef: Reproducible and exploratory quantum software experiments," 1 2026. [Online]. Available: https://arxiv.org/pdf/2511.04563

[59] B. Kitchenham, S. Charters *et al.*, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[60] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman *et al.*, "Quantum computing with Qiskit," 2024.

[61] M. U. Khan, M. A. Kamran, W. R. Khan, M. M. Ibrahim, M. U. Ali *et al.*, "Error Mitigation in the NISQ Era: Applying Measurement Error Mitigation Techniques to Enhance Quantum Circuit Performance," *Mathematics*, vol. 12, no. 14, p. 2235, Jan. 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/14/2235

[62] IBM Quantum, "Retired QPUs." [Online]. Available: https://quantum.cloud.ibm.com/docs/en/guides/quantum.cloud.ibm.com/docs/en/guides/processor-types

[63] A. W. Cross, A. Javadi-Abhari, T. Alexander, N. d. Beaudrap, L. S. Bishop *et al.*, "OpenQASM 3: A broader and deeper quantum assembly language," *ACM Transactions on Quantum Computing*, vol. 3, no. 3, pp. 1–50, Sep. 2022, arXiv:2104.14722 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2104.14722

[64] M. Zheng, A. Li, T. Terlaky, and X. Yang, "A Bayesian Approach for Characterizing and Mitigating Gate and Measurement Errors," *ACM Transactions on Quantum Computing*, vol. 4, no. 2, pp. 11:1–11:21, Feb. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3563397

[65] IBM Quantum, "Processor types." [Online]. Available: https://eu-de.quantum.cloud.ibm.com/docs/en/guides/eu-de.quantum.cloud.ibm.com/docs/en/guides/processor-types

[66] E. Desdentado, M. Polo, and C. Calero, "Estimating the number of shots to improve results accuracy," 2025, preprint. [Online]. Available: https://github.com/GreenTeamAlarcos/Estimating-The-Number-Of-Shots-To-Improve-Results-Accuracy

[67] R. Robertson, E. Doucet, E. Spicer, and S. Deffner, "Simon's algorithm in the NISQ cloud," *Entropy*, vol. 27, no. 7, p. 658, Jun. 2025, arXiv:2406.11771 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2406.11771

[68] C. Benito, E. López, B. Peropadre, and A. Bermudez, "Comparative study of quantum error correction strategies for the heavy-hexagonal lattice," *Quantum*, vol. 9, p. 1623, Feb. 2025, arXiv:2402.02185 [quant-ph]. [Online]. Available: http://arxiv.org/abs/2402.02185

[69] M. AbuGhanem, "Practical Fidelity Limits of Toffoli Gates in Superconducting Quantum Processors," Sep. 2025, arXiv:2509.05395 [quant-ph] version: 1. [Online]. Available: http://arxiv.org/abs/2509.05395

[70] Leibniz Supercomputing Centre, "First European quantum computer for Germany: Euro-Q-Exa starts operation at LRZ - Leibniz-Rechenzentrum." [Online]. Available: https://www.lrz.de/en/news/detail/first-european-quantum-computer-for-germany-euro-q-exa-starts-operation-at-lrz

[71] L. Kish, *Survey Sampling*. New York: John Wiley & Sons, 1965.