ChatGPT

Experiment Plan for Validating QuartumSE's Shadow-Based Measurements on IBM Q

Current GHZ-Based Benchmark: Strengths and Limitations

The existing validation setup uses a 3-qubit Greenberger–Horne–Zeilinger (GHZ) state and measures ~6 Pauli-string observables (e.g. multi-qubit parity correlations) to evaluate QuartumSE's features. This provides a simple entangled state benchmark with known expected values. **Strengths:** The 3-qubit GHZ is straightforward to prepare (a single Hadamard and two CNOTs) and is *easy to verify* via a few parity measurements [1]. The small circuit depth means relatively low gate error, and the handful of observables can confirm entanglement (GHZ fidelity > 0.5 indicates genuine tripartite entanglement). This baseline gave an initial test of shadow-based estimation vs direct measurement under minimal conditions.

However, **limitations** of the current GHZ test are significant:

- **Limited Scope of Observables:** Measuring only 6 Pauli observables on 3 qubits barely taps into the advantage of classical shadows. Direct measurement can handle a few observables with minimal overhead (by grouping commuting Pauli terms into common measurement bases). With so few targets, classical shadows (randomized measurements) may not show a clear shot advantage – the scenario is too small-scale. In other words, the *sample efficiency gains* of the shadow approach become evident only when estimating many observables simultaneously [2], whereas 6 observables is too low to reveal a strong benefit. Indeed, prior experiments found that advanced shadow techniques outperform naive sampling **especially as the number of observables grows** [2]. The current setup doesn't stress-test that regime.

- **Small System Size:** A 3-qubit GHZ is a toy problem; it doesn't represent the complexity of larger quantum states or practical algorithms. GHZ states are **highly sensitive to noise**, and scaling them up is difficult [1] [3]. On only 3 qubits, high fidelity is achievable, but the *impact of noise mitigation* is modest. For example, readout errors on 3 qubits can be corrected, but the absolute improvement in accuracy might be minor given the state is short-lived anyway. Moreover, any single error collapses an N-qubit GHZ, so larger GHZ states quickly lose fidelity on real hardware [1]. Sticking to 3 qubits avoids this, but also avoids confronting the true noise challenges that QuartumSE's features aim to address.

- **Over-simplified Observables:** The 6 Pauli observables in use (likely including $Z\otimes Z$ correlations and a 3-qubit parity like $X\otimes X\otimes X$) are relatively low-weight and largely commuting. This means even the direct method can measure them in only a couple of distinct bases. The *shot efficiency* benefit of classical shadows is minimal here – one random measurement can indeed capture all Pauli expectations at once, but a deterministic grouping can do nearly the same since the observables are compatible. In essence, the GHZ scenario is **too favorable for direct measurement**, so it fails to showcase an improvement in efficiency. It also doesn't probe higher-weight Pauli strings (e.g. a 3-qubit GHZ has one weight-3 stabilizer $X_1X_2X_3$, but other measured terms are weight-2 or 1), leaving the question of how shadows handle *complex, non-commuting observables* unanswered.

- **Limited Statistical Robustness Check:** With a small number of observables and qubits, the experiment likely collected only modest statistics. It's hard to assess properties like confidence interval (CI) coverage or estimator bias thoroughly from such a limited dataset. For example, the GHZ state's true expectation values are either +1, –1, or 0; a few hundred shots per observable might suffice for a rough estimate, but we cannot gauge if the reported 95% CI actually captures the truth 95% of the time without repeated trials. The current setup does not include repeated experiments or a large sample of observables to evaluate **CI coverage or distribution tails**. Classical shadow estimators can have heavy-tailed error distributions, necessitating techniques like median-of-means for rigorous guarantees. In a small single-run test, these subtleties might be overlooked, risking **mis-calibrated error bars**.

- **Low Scientific Novelty:** GHZ state demonstrations are well-trodden ground in quantum computing. A 3-qubit GHZ in particular is routinely prepared on IBM hardware and has even been exceeded (IBM recently prepared GHZ states up to 120 qubits as a system benchmark [4] [3]). Using it solely to validate a new measurement technique, without pushing into new territory, may not yield publishable insights. The current results are likely seen as an internal sanity check. To support an academic publication, we need experiments that go beyond this simple case – either by **improving measurement efficiency** substantially or by exploring regimes of scientific interest (larger entanglement, complex observables, or relevant algorithmic tasks).

In summary, the 3-qubit GHZ benchmark established a baseline but suffers from limited scope. It neither demonstrates a clear improvement in shot efficiency (due to too few observables) nor addresses more challenging measurement scenarios. As such, it's imperative to **extend this experiment** to more demanding and insightful configurations that can fully exercise QuartumSE's shadow-based measurement and noise mitigation capabilities.

Enhanced GHZ Experiments: Higher Weights and More Qubits

To build on the GHZ benchmark, we first propose **improved variants of the GHZ experiment** that use more qubits and higher-weight correlations. The goal is to maintain the conceptual clarity of GHZ states (maximally entangled "cat" states) while pushing into a regime where measurement is more demanding and noise mitigation is crucial.

- **Larger GHZ States (4–5 Qubits):** We will prepare GHZ states on 4 or 5 qubits (depending on device availability and connectivity). This involves adding extra CNOT entangling gates in a chain to extend $|\text{GHZ}_3\rangle = (|000\rangle+|111\rangle)/\sqrt{2}$ to $|\text{GHZ}_4\rangle$ or $|\text{GHZ}_5\rangle$. The circuit depth increases linearly with qubit count, making the state more fragile – an ideal test for noise mitigation. We expect the unmitigated fidelity of a 5-qubit GHZ to be quite low on a free-tier IBM device (potentially $F<0.5$), since GHZ states "are extremely sensitive to imperfections in the experiment" [3]. **QuartumSE's measurement error mitigation (MEM) will be critical here:** readout errors can otherwise mask entanglement by flipping outcome parity. (Indeed, readout error can introduce ambiguity between genuine entanglement loss vs. mere measurement mistakes [5].) By calibrating and correcting the measurement results, we aim to **recover the true multi-qubit correlations**. A key success criterion will be demonstrating a *measurable GHZ fidelity above the classical threshold* of 50% once error-mitigated. For example, if raw data for a 4-qubit GHZ yields a fidelity ~0.4, but after applying MEM it rises to ~0.6, that conclusively shows entanglement was present and that error mitigation was necessary to observe it. This would directly showcase the value of QuartumSE's noise mitigation on real hardware. Additionally, a larger GHZ provides more

higher-weight Pauli observables to estimate: e.g. a 4-qubit GHZ has a weight-4 parity $X_1X_2X_3X_4$ as an important observable (along with multiple weight-2 and weight-3 correlators). Estimating these higher-weight operators efficiently is challenging for direct methods (they involve all qubits), but classical shadows can predict them with the same effort as any local observable. We will quantify how well the shadow-based estimator predicts, say, the 4-body XXXXXX parity (for 5 qubits) or XXXX (for 4 qubits) **relative to direct measurement**. If classical shadows achieve the same error with significantly fewer shots, that will reflect as a high **shot-saving ratio (SSR)** for the high-weight correlations.

- **Higher-Weight Observables on GHZ:** Even for the existing 3-qubit GHZ, we could extend the range of measured observables beyond the current 6, to include **all Pauli terms relevant to fidelity** and entanglement witnesses. For instance, a 3-qubit GHZ state's fidelity can be verified by measuring its stabilizer set: ${Z_1Z_2,\;Z_2Z_3,\;X_1X_2X_3}$ (each ideally +1) along with perhaps single-qubit Z_i expectations (ideally 0) [6]. The present setup uses some of these, but we can add *non-commuting basis combinations* like $Y_1Y_2X_3$, $Y_1X_2Y_3$, $X_1Y_2Y_3$ which appear in Mermin's inequality tests for GHZ. Although these are not all simultaneously needed for fidelity, including them would increase the number of unique Pauli terms to estimate (into the 8–10 range). This tests QuartumSE's ability to **handle a broader observable set** without additional cost. Direct measurement would require separate bases to get those Y-containing correlators, whereas a classical shadow (random Pauli measurements) naturally samples across X,Y,Z bases and can yield estimates for all terms. We anticipate that by expanding to a *full basis of GHZ correlations* (including some that are zero in theory but non-zero if there's, say, certain error types), the classical shadow approach will demonstrate superior **shot efficiency**. A simple metric is the total number of circuit shots required: e.g. measuring 10 observables one-by-one might take 10,000 shots (at 1000 shots each) whereas using 1000 random measurements could estimate all 10 with comparable precision. Achieving an SSR > 1 (i.e. fewer total shots for the shadow method to reach the same error) in this GHZ extension would validate the advantage beyond the trivial regime.

- **Improved Statistical Robustness:** Using a larger GHZ and more observables also allows more rigorous statistical checks. We can repeat the GHZ measurement experiment multiple times (e.g. prepare and measure the 4-qubit GHZ in 10 independent runs) to empirically assess **confidence interval coverage**. For each run, QuartumSE will produce estimates with error bars (e.g. 95% CIs) for each observable or the fidelity. By comparing these across runs, we can verify if the true values (as determined by a high-statistics reference or the ensemble mean) fall within the predicted CIs the expected ~95% of the time. Any significant under-coverage (say only 70% of runs' CIs contain the reference value) would signal that the estimator's error model (perhaps assuming normal variance) is unreliable – a point to address in analysis. Larger sample sizes and observables will reveal if **classical shadows exhibit heavy-tail outliers** (we might catch occasional runs with large error, affecting CI coverage). If so, strategies like median-of-means or derandomization could be noted as future improvements. In short, the extended GHZ experiments serve as a bridge between the simplistic initial test and more complex scenarios, ensuring our methods are statistically sound and *calibrating our error mitigation* on a known entangled state.

By pushing GHZ experiments to 4–5 qubits and a richer set of measurements, we expect to **demonstrate clear benefits of shadow-based estimation** (measuring high-weight entanglement correlations more efficiently) and to underscore the necessity of error mitigation to observe multi-qubit entanglement on real hardware. These findings, while still a controlled "toy" scenario, would strengthen any publication by moving beyond the trivial baseline and highlighting how QuartumSE improves measurement fidelity and efficiency for entangled states that are at the edge of the hardware's capability.

Alternate Circuit Families for Broader Validation

Beyond GHZ states, we propose experiments on a variety of **circuit families** to validate QuartumSE's features in diverse contexts. This will not only test the generality of the shadow-based measurement approach but also target scenarios of *higher scientific interest* (e.g. relevant to quantum algorithms or fundamental benchmarks). Each of the following families introduces different state characteristics and observable sets, going well beyond simple 3-qubit parity measurements:

- **Product of Bell Pairs (Disjoint Entanglement):** As an intermediate complexity test, we can prepare an n-qubit state that consists of multiple independent Bell pairs. For example, on 4 qubits prepare two Bell states $|\Phi^+\rangle_{12}\otimes|\Phi^+\rangle_{34}$ (each via one CNOT on a Hadamard-prepared qubit pair). This state has entanglement, but only locally within each pair, and no global n-qubit coherence. It serves as a useful validation case for **simultaneously measuring multiple subsystem observables**. We will task QuartumSE with estimating, for instance, the correlation $\langle Z_1Z_2 \rangle$ and $\langle X_1X_2 \rangle$ for the first pair, and similarly for the second pair, **all in one go**. Direct measurement could handle each pair separately (since observables on different pairs commute and can be grouped), but as the number of pairs grows, grouping becomes more complex. Classical shadows will naturally capture all these local correlations without any explicit grouping – every random measurement on all qubits yields data for both pairs. **Metrics:** We will check the accuracy of each pair's correlation (targeting near $+1$ on ZZ and -1 on XX for an ideal $|\Phi^+\rangle$) and use MEM to correct any bias from readout error (which often significantly depresses raw Bell state correlations). A simple but impactful result here is to demonstrate a **Bell inequality (CHSH) violation** under noise mitigation: measure the CHSH combination S for one Bell pair (involving correlations like XX, XZ, ZX, ZZ in different bases). Without mitigation, readout error might shrink S below the classical limit 2, obscuring the violation; with mitigation, we expect $S>2$ by a comfortable margin, showing quantum correlations are recovered. Achieving a CHSH value close to the ideal ~2.5–2.8 would be publishable evidence of how error mitigation unveils non-local correlations. While Bell pairs are standard, doing this *in the context of simultaneous multi-pair measurement* is novel. It stresses that QuartumSE can handle **multiple entangled subsystems at once**, yielding results as if each pair were measured individually. Furthermore, we can compute SSR here by comparing the shots needed when measuring both pairs together (with shadows or a single grouped circuit) vs measuring each pair separately. As n (number of qubits) increases, the number of disjoint pairs grows as $n/2$, and the benefit of a one-and-done shadow approach should increase accordingly. This experiment is a **benchmark for parallelized measurements** – if successful on 4 qubits, we can extend to 6 or 8 qubits (3 or 4 Bell pairs) within the free-tier limit to reinforce the point. It's a clean scenario to verify that adding more qubits (and thus more observables) doesn't degrade QuartumSE's performance, thanks to classical shadows' *basis-agnostic scalability* [7] [8].

- **Random Clifford Circuits (Generic States & Fidelity Estimates):** Random Clifford circuit states provide a pseudo-random ensemble of multi-qubit states that are **highly non-trivial, yet classically tractable** (because stabilizer states can be efficiently simulated). We will generate one or more random Clifford circuits on, say, 5 qubits (using random sequences of Clifford gates with modest depth) to produce random stabilizer states. These states typically have no simple product or GHZ structure; their Pauli expectation values are broadly distributed (most are 0, a few stabilizer generators are ±1). This experiment addresses two goals: (1) **Generic benchmarking of estimator performance** and (2) **Direct fidelity estimation (DFE)**. First, by treating the random state's many Pauli expectations as "targets", we can evaluate how well classical shadows predict a large number of them simultaneously. A random n-qubit stabilizer state has $2n$ non-trivial stabilizer observables fixed at ±1, and all other Pauli expectations ideally 0. We can ask QuartumSE to estimate, say, **dozens of Pauli observables** (both those that are true stabilizers and some that are not) for the given state. This is effectively a partial tomography test: did the measurement scheme identify the correct stabilizer group? In theory, Huang *et al.* showed that $\mathcal{O}(\log M)$ random measurements can estimate M observables [9], but in practice with noise, we will see some estimation error. The **mean absolute error (MAE)** across many observables and the **distribution of errors** will be measured. Since we know the ideal values (0 or ±1), we can directly compute how often the 95% CI contains the true value (giving CI coverage). A well-calibrated method should have ~95% of the observables' true values within their CIs; significant deviations might occur if, for example, a few observables have much higher variance than assumed (heavy-tail effects). By repeating this on a few random Clifford instances (or repeating the measurement on the same state multiple times), we gather statistics to assess **robustness**. This provides a *broad stress-test* of QuartumSE: instead of one special state, we cover a random sample of states.

Secondly, for **direct fidelity estimation**, note that stabilizer states like these are an ideal use-case: DFE requires only a constant number of measurements to estimate fidelity to a known stabilizer state [10]. We will leverage classical shadows to estimate the fidelity of the prepared state to the ideal (simulated) state. In practice, fidelity to a known pure state $|\psi\rangle$ can be computed by measuring the expectation of the projector $|\psi\rag\langle\psi|$, which equals the average of all Pauli terms in $|\psi\rangle$'s stabilizer decomposition [6]. For a Clifford state, this decomposition contains relatively few Pauli terms (equal to the number of stabilizers). For example, if the random Clifford state has stabilizers $g_1,...,g_{n}$, then fidelity can be obtained by measuring $\langle g_1 \rangle,\ldots,\langle g_n \rangle$. **Plan:** We will use QuartumSE to estimate each g_i's expectation. Directly, we could measure each stabilizer by picking the appropriate basis (each g_i is a Pauli string), but that's n separate measurements. Instead, classical shadows with random Pauli bases should capture all g_i values in one run, analogous to how IBM's 120-qubit GHZ experiment employed DFE to avoid separate experiments for each parity term [5] [6]. We'll compare the *shadow-based fidelity* estimate to a conventional approach (measuring each stabilizer with dedicated circuits) in terms of shots needed and accuracy. Success here would be a strong indicator that **QuartumSE can efficiently verify large state preparations** – crucial for scalable benchmarking. For instance, if a 5-qubit random Clifford state yields fidelity $F \approx 0.9$ with a small uncertainty using only, say, 500 measurement shots (shadows), whereas a traditional approach might require a few thousand shots distributed over multiple circuits, that's clear evidence of measurement efficiency (an SSR of several-fold). It also aligns with IBM's finding that DFE is a scalable method for stabilizer states [6]; we would demonstrate this on the free-tier hardware with our tool. These random-state trials are academically interesting because they test the *worst-case scenario* (no special structure) for our measurement schemes and ensure that our error mitigation and statistical analysis hold up generally, not just for highly structured states.

- **Trotterized Hamiltonian States (Physical Systems Simulation):** Finally, to target scientifically relevant use-cases, we will validate QuartumSE on states obtained from **simulated Hamiltonian dynamics or ground states**, using trotterization. This connects our experiments to quantum simulation and variational algorithm contexts. Two concrete examples are: (a) a **Quantum Spin Chain** (e.g. a 1D Ising model with transverse field) and (b) a **Molecular Ground State** (e.g. H_2 molecule in a minimal basis). Both involve Hamiltonians with multiple Pauli terms, where measuring the energy or other observables is non-trivial.

For (a) *spin chain dynamics*: we can start with an easy-to-prepare state (like all spins polarized $|000\dots0\rangle$ along Z) and then apply a short time evolution under a Hamiltonian such as

$H_{\text{Ising}} = \sum_{i} J Z_i Z_{i+1} + h \sum_i X_i$. Using a first-order Trotter circuit for a small time t, we prepare the state $|\psi(t)\rangle \approx e^{-i H t}|000...0\rangle$. Even a single Trotter step (applying $e^{-iJ Z_iZ_{i+1}\Delta t}$ and $e^{-i h X_i \Delta t}$ sequentially) on 4–6 qubits will create an entangled state with non-trivial correlations (e.g. nearest-neighbor spin correlations, domain formation, etc.). This is a **realistic dynamic state** one might encounter in NISQ algorithms. We will task QuartumSE with measuring key observables of this state, such as the **energy (expectation of H_{Ising})** and possibly quantities like magnetization $\langle X_i \rangle$ or two-point correlators $\langle Z_i Z_j \rangle$. The Hamiltonian includes terms that do not all commute (e.g. Z_iZ_{i+1} terms commute with each other but not with X_i terms). A *naïve measurement strategy* would require separate circuits for the $Z Z$ correlations (measure in the Z basis) and for the X terms (measure in the X basis). Even with optimal grouping, at least **2 groups of measurements** are needed (and more if the Hamiltonian had random orientations). Classical shadows, on the other hand, can estimate **all terms in one unified scheme** – each random Pauli measurement covers some Z-basis and some X-basis information across the qubits. We expect QuartumSE's shadows+MEM approach to estimate the energy **with fewer total shots** than a grouped direct approach. A concrete plan: allocate a fixed budget of shots (e.g. 5,000 shots) and compare the accuracy of the energy estimate from (i) two predetermined measurements (2,500 shots in Z-basis + 2,500 in X-basis, combined) vs. (ii) 5,000 shots of random Pauli measurements processed by the shadow estimator. We will evaluate **mean error and variance** of the energy estimates under both schemes. A successful outcome would be that the shadow-based method achieves equal or lower error in energy for the same total shots – demonstrating **variance reduction or higher shot efficiency** thanks to measuring all terms at once. Additionally, shadows will provide *extra data*: since they measure a complete basis each shot, we can simultaneously extract other observables (like multi-spin correlation functions) from the same dataset at no additional cost, something the direct method cannot do without extra measurements. This highlights the **breadth of data** captured by classical shadows: for example, we could compute not just $\langle H \rangle$ but also $\langle H^2 \rangle$ (energy variance) from the same measurement set, by estimating higher-order correlators – a task that would be very costly with separate measurements for each term [11] [12]. In essence, the trotterized spin chain experiment benchmarks QuartumSE in a scenario akin to variational algorithms or real-time dynamics, where many local terms must be measured quickly. If our approach can **reduce the measurement cost for estimating energy or correlation functions** in such a simulation, that is a notable result given that measurement overhead is a known bottleneck in variational quantum eigensolver (VQE) algorithms [7].

For (b) *molecular ground state (H_2)*: we will simulate the electronic ground state of the hydrogen molecule mapped onto qubits (4 qubits in a minimal STO-3G basis after Jordan–Wigner or Bravyi–Kitaev encoding). Instead of performing a full VQE, we can take a known approximate circuit for H_2's ground state at a certain bond length (for example, a two-qubit UCC ansatz repeated on two orbital pairs, resulting in a 4-qubit state) and run it on hardware. The Hamiltonian H_{H2} in qubit form consists of a sum of Pauli terms (e.g. Z_0Z_1, Z_2Z_3, Z_1Z_2, $X_0X_1X_2X_3$, etc., typically on the order of 10–15 terms) [13]. Measuring the ground state energy with classical methods requires grouping these terms into measurement bases – a non-trivial task because many terms don't commute. Prior research has shown this can take multiple groups and many samples, which is a limiting factor for quantum chemistry on NISQ devices [7]. We will use QuartumSE to perform **classical-shadow energy estimation** of H_2. In practice, this means using the *"shadows v1 + MEM"* method (potentially a locally biased shadow that gives more weight to important terms) to estimate all Pauli terms' expectations and summing them for the energy. Meanwhile, we will also do a baseline measurement using a near-optimal grouping strategy (e.g. grouping by qubit-wise commutativity [14] [15]) to estimate the energy with the same total shots. **Metrics:** We will compare the estimated energy to the known exact energy of H_2 (from classical calculation) to compute accuracy, and also compare the statistical uncertainty. A key target is to see if the **confidence interval for the energy** from QuartumSE's method is tighter for a given runtime. If, for example, the shadow+MEM method can achieve an energy within ± 0.02

Hartree of the true value with 95% confidence in 10 minutes of data, whereas the direct grouped method yields ± 0.05 uncertainty in the same time, that's a clear improvement. Even more, we'll track SSR: e.g. to reach a fixed error threshold (say 0.02 Hartree standard error), how many shots does each method need? This can be extrapolated if needed by running multiple shot counts or by examining variance – yielding an SSR factor. A recent experimental study showed that advanced measurement schemes (like derandomized shadows) **outperform exhaustive grouping as Hamiltonians grow more complex** [8] [16]. We aim to corroborate this on real IBM hardware with H_2. Achieving chemical accuracy (1.6 mHartree) is likely beyond reach due to noise, but demonstrating any noticeable **variance reduction or shot savings** is noteworthy. Furthermore, this experiment has *publication merit*: it addresses the pressing issue of measurement cost in quantum chemistry. If QuartumSE can cut down the runtime or shots needed for energy estimation under free-tier constraints, it suggests a way forward for resource-efficient VQE-like studies – a result the community would find valuable.

In summary, these alternate circuit family experiments – Bell pair arrays, random Clifford states, and trotterized Hamiltonian states – ensure that we test QuartumSE's shadow-based measurement and error mitigation in a **wide range of scenarios**. Each scenario targets different metrics (Bell: fidelity & non-local correlation, Clifford: multi-observable accuracy & CI calibration, Hamiltonian: efficiency in estimating complex operators) and together they provide a comprehensive validation. Importantly, they move beyond contrived examples into regimes relevant for quantum network tests, random circuit benchmarks, and quantum simulation respectively, aligning our validation with academically interesting problems.

Experimental Configurations, Shot Allocation, and Success Criteria

To execute the above plan within the IBM Quantum free-tier limits (10 minutes of quantum processor time per month [17]), we must carefully optimize the shot budgets and define clear criteria for success for each experiment. Table 1 below summarizes the proposed experiment configurations, including the state/circuit to prepare, the measurement tasks, estimated shot allocations (ensuring the total fits in ~10 minutes runtime), and the success thresholds or metrics for each. These thresholds represent the targets we consider as validation "passes" or publishable outcomes for the experiment.

Table 1. Experiment plan overview with configurations, shot budget estimates, and success criteria (primary metrics in **bold**).

Experiment	**State / Circuit** (Qubits)	**Measurements & Targets**	**Shot Allocation** (approx.)	**Success Criteria** (Thresholds)
Extended GHZ Entanglement	4- or 5-qubit GHZ state circuit (Hadamard + CNOT chain)	- Estimate full GHZ **fidelity** from multi-qubit parity observables (e.g. $X!X...X$, $Z!Z$ pairs) - Evaluate entanglement witness (fidelity > 0.5) with/without MEM - Metrics: **Fidelity**, SSR, CI, MAE for parity expec.	~3,000 shots total: – 5 bases × 600 shots each for direct method (grouped) – 3,000 random shots for shadows v0/v1 *(fits ~1–2 min runtime)*	**Fidelity ≥ 0.5** (entangled) **with MEM** vs <0.5 without (showing mitigation effect) [5] . GHZ parity **MAE < 0.1** (absolute error) for all methods. Classical shadows achieves **SSR > 1** (e.g. uses 50% fewer shots) to reach same fidelity error as direct. 95% CI for fidelity includes ideal value (0.707 for 5-qubit GHZ) in ≥90% runs.
Multiple Bell Pairs	Two Bell pairs on 4 qubits (e.g. $	Φ^+\rangle_{12} \otimes	Φ^+\rangle_{34}$)	- Measure **Bell-state correlations**: $X!X$, $Z!Z$ for each pair - CHSH inequality test on one pair (S-value) - Metrics: Correlation values, CHSH S, fidelity per pair, runtime

Experiment	**State / Circuit** (Qubits)	**Measurements & Targets**	**Shot Allocation** (approx.)	**Success Criteria** (Thresholds)
Random Clifford State	Random 5-qubit Clifford circuit (depth ~10, yields random stabilizer state)	- **Predict 50+ Pauli observables** (stabilizers and random others) with shadows vs grouped direct - Compute state **fidelity** via stabilizer measurements (DFE) - Metrics: **SSR** for multi-observable estimation, distribution of errors, CI coverage, fidelity estimate	~5,000 shots total: – Direct: ≥5 distinct groups for 50 observables, 1000 shots each = 5,000 shots – Shadows: 5,000 random shots (could split into 5 runs of 1000 to gauge variance) *(≈2 min runtime)*	**High SSR**: Shadows use ~same shots to estimate 50 obs as direct needed for 1 obs (theoretical SSR ~10×; expect effective SSR > 2×) [2] . Average **CI coverage ~95%** for known expectation values (verify via repeated trials). Fidelity to ideal state estimated within ±0.05 (absolute) of true value, CI contains true fidelity in 95% of bootstrap trials. At least **80% of observables' MAE < 0.1** (accurate predictions overall).

Experiment	**State / Circuit** (Qubits)	**Measurements & Targets**	**Shot Allocation** (approx.)	**Success Criteria** (Thresholds)
Ising Chain Simulation	6-qubit Transverse Ising circuit (1–2 Trotter steps for small t)	- Estimate **energy $\langle H \rangle$** (sum of Z_iZ_{i+1} and X_i terms) using shadows vs two-basis direct - Also extract secondary observables: magnetization $ \langle X_i \rangle$, two-point $\langle Z_iZ_j \rangle$ from same data - Metrics: Energy estimate bias and uncertainty, **variance reduction**, additional info gained per run	~6,000 shots total: – Direct: 2 bases (Z, X) × 1500 shots each = 3,000 – Shadows: 3,000 shots (random Pauli) for v0 + 3,000 for v1+MEM *(~3–4 min runtime due to 6-qubit circuit execution time)*	**Energy error < 5%** of full scale (compare to exact energy). Shadows method achieves **≥2× lower variance** in energy estimate than single-basis sampling (or, halves the shots needed for same error). All nearest-neighbor $ \langle Z_iZ_{i+1}\rangle$ correlations within ±0.1 of exact value (with MEM). Successful extraction of *all* term expectations from one dataset (demonstrating parallel measurement of non-commuting terms).

Experiment	**State / Circuit** (Qubits)	**Measurements & Targets**	**Shot Allocation** (approx.)	**Success Criteria** (Thresholds)
H_2 Ground State Energy	4-qubit H_2 VQE ansatz state (approximate ground state at fixed bond length)	- Measure **Hamiltonian terms** (~10–15 Pauli terms) and sum for ground-state energy - Compare shadows (possibly biased to important terms) vs optimal grouping direct method - Metrics: **Energy estimation error**, shot cost (runtime), SSR, improvement with MEM	~8,000 shots total: – Grouped direct: ~4 groups × 1000 shots each = 4,000 – Shadows v1+MEM: ~4,000 shots (random with bias toward Z terms that dominate Hamiltonian) *(~4 min runtime)*	**Energy estimate within 0.02–0.05 Ha** of exact value (chemical accuracy not expected, but aim for few % error). Shadow+MEM method achieves **≳30% lower uncertainty** (std. error) on energy than direct with equal shots (SSR ≈ 1.3+). If biasing used: no significant bias in result (check energy estimate consistent with unbiased method within error). Demonstrate viable energy measurement under free-tier constraints, supporting publishable VQE benchmark.

Notes on runtime and shot feasibility: The shot counts in Table 1 are chosen to sum to roughly 24,000–25,000 shots distributed across experiments, which we anticipate can fit in a 10-minute quantum compute budget. IBM's 10-minute free allowance translates to on the order of 10^5 single-shot executions on a small device, given fast gate times (~100–300 ns) and modest readout times, though overheads (job queuing, result retrieval) also consume wall-clock time. By batching shots into as few circuits as possible (e.g. each "basis" execution uses ~500–1000 shots) and minimizing circuit depth (most circuits are short), we maximize the useful shots per minute. Calibration overhead for MEM (e.g. running bit-string calibration circuits for each qubit) will be performed once per experiment (costing maybe 32 circuits for 5 qubits, negligible in this budget). We will prioritize experiments as needed to remain within runtime – for instance, the H_2 energy estimation (4 min) and Ising simulation (3–4 min) are the costliest runs; the GHZ, Bell, and Clifford tests are shorter and can be run in one session (~1–2 min each). If needed, experiments can be split across two monthly cycles (e.g. run half in Month 1, half in Month 2) without issue, since each is independent.

Each experiment's **success criteria** include primary metrics that must be met or exceeded for the validation to be considered successful:

- **Fidelity/Entanglement criteria (for GHZ, Bell):** We require that error mitigation enables detection of entanglement where it would otherwise be missed. For GHZ, a *fidelity* $\geq$*0.5* with confidence is the threshold for genuine multipartite entanglement [4]. For Bell pairs, a CHSH violation ($S>2$) is the threshold for non-locality. These binary criteria (entangled or not, violation or not) are powerful qualitative outcomes. Meeting them decisively (with statistically significant margins) makes for compelling evidence of QuartumSE's effectiveness.

- **Accuracy of estimates (MAE, absolute error):** For known target values (like GHZ ideal parity = ±1 or 0, Bell correlations = ±1, etc.), we set a tolerance (e.g. MAE < 0.1 or correlation > 0.9) that the methods should achieve. These thresholds ensure the measurements are in a regime that's meaningful (e.g. a two-qubit correlation of 0.95 indicates an excellent state, whereas 0.5 would indicate heavy decoherence). The error mitigation should help reach these thresholds by correcting biases.

- **Shot Efficiency and SSR:** Perhaps the most critical quantitative metric for publication is SSR (Shot Saving Ratio). We have defined SSR in context for each experiment as the factor by which the shadow-based approach reduces the required shots compared to a direct measurement strategy (for a given precision). For example, in the Clifford test, theoretically one shadow measurement can replace many separate observable-specific measurements [9]; we target an effective SSR > 2 (meaning, say, 5000 shots with shadows yields the same accuracy as ~10,000 shots distributed in the direct scheme). In the Hamiltonian cases, even a 30–50% reduction in shot count for the same error (SSR ~1.3–1.5) is significant given how measurement costs scale combinatorially with system size [7]. Each experiment has an SSR or variance-reduction goal (implicitly >1) in the table. Achieving these would validate that **QuartumSE's shadow and mitigation features genuinely improve efficiency**, not just accuracy.

- **Confidence Interval Coverage:** For any method intended for rigorous usage, having trustworthy error bars is important. Thus, we will declare success if our empirical CI coverage is close to the nominal level (e.g. ~95% of true values falling in 95% CIs). Small deviations are acceptable given limited trials, but a major shortfall would indicate underestimation of uncertainties (perhaps due to ignoring heavy tails). We expect that with sufficient samples, and possibly using ensemble or bootstrapped estimates, we can get near the correct coverage. This will bolster the credibility of any reported QuartumSE measurement — readers can trust the error estimates.

- **Secondary metrics:** We also track *variance reduction*, *runtime*, and *cost*. Variance reduction is tied to SSR (lower variance means fewer shots needed). Runtime we keep within 10 minutes by design. Cost (in terms of cloud credits or similar) isn't directly an issue in free tier, but our results could be extrapolated to paid usage (demonstrating fewer shots = lower cost experiments). As a soft success criterion, we consider it a win if we demonstrate that certain experiments **previously thought infeasible under the free plan can be done in 10 minutes** thanks to our optimizations. For instance, measuring a molecular energy to reasonable accuracy under free-tier constraints would be a strong message of efficiency.

In executing these configurations, we will also remain flexible: if one experiment is running longer than expected, we may reduce shot counts or skip a lower-priority measurement to stay within budget. The table's allocations intentionally leave a bit of headroom (summing to ~10 min) and focus on critical data.

We will use classical simulation or prior literature values as references for “true” values whenever possible (e.g. exact diagonalization for the Ising model energy, known exact H_2 energy). Where that’s not possible (e.g. the exact prepared state on hardware is unknown due to noise), we will use a high-shot direct measurement as the ground truth for comparison [18]. This approach was used in the prior classical shadows experiment to define errors [18].

Overall, this experimental design ensures that **each scenario has a clear goal and benchmark for success**. Meeting these success criteria across the board will provide a compelling validation of QuartumSE: we will have shown improved measurement accuracy (via MEM), improved efficiency (via classical shadows) and reliable statistical performance, all under realistic use conditions on IBM quantum hardware.

High-Impact Results and Publication Potential

Not all the above experiments are equal in terms of scientific impact. We conclude by highlighting which results are most likely to **yield publishable, high-impact outcomes** under the given constraints, and why:

- **Efficient Hamiltonian Measurement (H_2 Energy or Spin Chain):** Demonstrating a reduction in measurement cost for computing a Hamiltonian expectation is perhaps the most significant result for the broader community. Quantum algorithm researchers recognize that measuring an n-qubit Hamiltonian with many terms is a key bottleneck [7]. If our experiment shows, for example, that using classical shadows and error mitigation can **cut the required shots by ~50% or more for the same energy precision**, this is a publishable finding. It would directly cite that we achieved a higher efficiency than prior approaches on real hardware. The fact that we do this within a 10-minute free allocation makes it even more striking – it implies that even small-scale quantum computers can be pushed to do meaningful chemistry/physics calculations with smart measurement schemes. We expect the H_2 energy experiment in particular to draw interest: it connects to the extensive literature on reducing VQE measurement overhead (e.g. grouping, importance sampling, shadow tomography) and provides an experimental validation of those ideas [11] [13]. A figure in a paper could show the energy estimate convergence versus number of shots for direct vs QuartumSE’s method, highlighting the lower effort or higher accuracy of the latter. This would be a clear win for our approach and a strong justification for publication.

- **Improved Entanglement Verification via MEM:** Another high-impact outcome is showing that **without noise mitigation one would draw the wrong conclusion about a quantum state, but with mitigation the truth is revealed**. For instance, if our 5-qubit GHZ experiment finds raw fidelity ~40% (which would suggest merely classical mixture), but after QuartumSE’s error mitigation the fidelity is ~60% (indicating genuine entanglement), this is a compelling story. It exemplifies the importance of error mitigation on NISQ devices – essentially *enabling the detection of quantum correlations that hardware noise was hiding*. This kind of result is publishable as a case study in quantum benchmarking: it echoes the IBM 120-qubit GHZ work where readout errors had to be carefully mitigated to certify entanglement [5] [6]. While our scale is smaller, the free-tier context (with far less sophisticated calibration than IBM’s internal labs) makes it relatable to many researchers. We can generalize the message: **“Using classical shadows with measurement error mitigation, we accurately verified multipartite entanglement on hardware where naive methods failed.”** That underscores QuartumSE’s practical value. A possible high-impact data point is the CHSH violation in the Bell pair test: achieving an S value well above 2 (and close to the quantum maximum) on an IBM device, within a few minutes and with only error mitigation (not post-selecting or encoding), would be a notable demonstration that even free-tier devices can exhibit non-locality clearly. Past experiments have shown Bell violations on IBM Q, but doing so while comparing raw vs mitigated results (showing raw might not violate due to errors, mitigated does) provides a neat narrative of "noise obscured a quantum property, our tool recovered it."

- **Many-Observables Estimation and Statistical Consistency:** From a more theoretical perspective, if our random Clifford experiments show that we can **simultaneously estimate dozens of observables with high accuracy and proper confidence calibration**, that is a strong validation of the classical shadows approach in practice. The original theory promised an exponential compression in measurement effort for multiple observables [9]; confirming this experimentally (even on 5 qubits) will attract interest from quantum information scientists. Additionally, reporting on the distribution of errors and CI coverage in a real experiment can highlight any gaps between theory and practice (e.g. the need for median-of-means or derandomization to handle outliers). If we find that naive averaging sometimes underestimates uncertainties, that insight is publishable as guidance for future experiments – it adds to the understanding of how classical shadows behave with finite samples and device noise. Conversely, if coverage is good, we gain confidence in applying these methods to larger systems. Either way, documenting the *statistical reliability* of shadow tomography in a lab setting contributes to the literature (which so far has very few experimental points, the notable one being Zhang *et al.* 2021 on a photonic system [18] [8]). Our work would complement that by using superconducting qubits and a very limited resource budget, making it uniquely positioned.

- **Effective Use of Limited Quantum Resources:** A more general high-impact aspect is the demonstration of **research-grade experiments on a strict resource budget**. By explicitly targeting the 10-minute monthly runtime and still achieving meaningful results, we showcase a methodology for others who have limited quantum hardware access (which is a common situation). This meta-level message – *that clever experimental design and classical post-processing can compensate for hardware limitations* – is certainly publication-worthy in venues focusing on experimental quantum computing or quantum software. For example, our results could be framed as *"Shadow-based methods enable significant data extraction in minimal quantum time"*. We will have essentially turned the free-tier device (with all its constraints) into a testbed for advanced measurement strategies that produce academically valuable data. This is encouraging for democratizing quantum research: it suggests one doesn't need unlimited premium access to do cutting-edge experiments. We anticipate reviewers would appreciate that we fit a benchmarking suite into a small time window, and yet gathered enough data to verify theoretical advantages (SSR > 1, proper CI's, etc.).

In prioritizing the experiments, those that maximize insight per minute are key. The **H_2 energy measurement and Ising simulation** are top priorities for their relevance to quantum algorithms. The **extended GHZ/Bell tests** are next, as they vividly illustrate noise mitigation's impact on entanglement verification. The **random Clifford test** is somewhat more diagnostic in nature; it solidifies confidence in the methods but might be seen as supplementary unless it reveals something unexpected (e.g. a breakdown of classical confidence predictions). It's still worth doing and reporting, perhaps in an appendix or supplementary section, to show comprehensive validation. If time is tight, we would ensure the Hamiltonian and GHZ/Bell experiments are executed first since they offer the clearest "story" for a paper.

Overall, by executing this carefully crafted set of experiments, we expect to **produce multiple noteworthy findings**: improved measurement efficiency, successful noise mitigation enabling entanglement detection, and confirmation of theoretical properties of classical shadows on real hardware. Each of these aligns with current research interests in the quantum computing community, making our results suitable for high-quality conference presentations or journal publications. Crucially, all of it is achieved within the modest limits of IBM's free tier – a testament to optimization and smart experimental design. This will strengthen the case for QuartumSE as a practical tool and underscore the message that **better measurement protocols are a key to unlocking more quantum science with today's devices** [7] [16] .

With clear positive outcomes against our success criteria, we will have a compelling, comprehensive validation of QuartumSE's shadow-based measurement and error mitigation features, ready to share with the academic community. The combination of benchmarking data, efficiency metrics, and real-hardware entanglement evidence should form the core of a strong publication, demonstrating both the *principles* and the *pragmatic advantages* of our approach.

Sources:

1. Huang, H.-Y. et al., *Nature Phys.* **16**, 1050 (2020) – Introduced classical shadow tomography, showing logarithmic scaling in measurements for many observables [9] .

2. Zhang, T. et al., *Phys. Rev. Lett.* **127**, 200501 (2021) – Experimental demonstration of classical shadows on a 4-qubit photonic GHZ state; compared uniform, biased, derandomized shadows vs grouped measurements [18] [8] .

3. IBM Quantum Platform Documentation – Free tier provides **10 minutes/month** of quantum processor runtime [17] .

4. Javadi-Abhari, A. et al., "Big cats: entanglement in 120 qubits...", *arXiv:2510.09520* (2025) – Prepared 120-qubit GHZ on IBM; noted GHZ states are *"easy to verify, but difficult to prepare due to high sensitivity to noise"* [1] and required efficient verification via parity tests and direct fidelity estimation (constant measurements independent of qubit count) [10] with careful readout error mitigation [5] .

5. IBM Quantum, *"Dynamic circuits enable efficient long-range entanglement"* (2023) – (Background on GHZ and Bell states on IBM hardware, not directly cited above but provides context on preparing entangled states on current devices).

6. McClean, J. et al., *Phys. Rev. A* **95**, 042308 (2017) – Identified the measurement bottleneck in VQE, where Hamiltonians with many Pauli terms require prohibitively many samples for accurate energy estimation [7] .

7. Verteletskyi, V. et al., *Phys. Rev. A* **101**, 012342 (2020) – Proposed heuristic grouping of Pauli terms to reduce measurements (used here as a baseline for direct strategies) [14] .

8. Torlai, G. et al., *Nature Phys.* **14**, 447 (2018) – Demonstrated learning quantum states from measurements; highlights importance of efficient state tomography (related to classical shadows concept).

9. Kandala, A. et al., *Nature* **549**, 242 (2017) – Performed VQE for H_2 on hardware; had to group measurements, illustrating the need for improved techniques (context for our H_2 experiment).

10. Sun, S.-N. et al., *Phys. Rev. A* **101**, 022307 (2020) – Studied measurement error mitigation on IBM devices; relevant to our MEM implementation and its expected impact on observables (not explicitly cited above, but underpins the MEM strategy we use).

---

1 3 4 5 6 10 Big cats: entanglement in 120 qubits and beyond
https://arxiv.org/html/2510.09520v1

2 7 8 11 12 13 14 15 16 18 [2106.10190] Experimental quantum state measurement with classical shadows
https://ar5iv.labs.arxiv.org/html/2106.10190

9 [2106.10190] Experimental quantum state measurement with classical shadows
https://arxiv.org/abs/2106.10190

17 Overview of plans | IBM Quantum Documentation
https://quantum.cloud.ibm.com/docs/guides/plans-overview