

Finite-Shot Pauli-Measurement Protocol for Calibrated Multipartite-Entanglement Diagnosis

Wenlong Sun

Shenyang University of Technology

Yuanfeng Jin

yfkim@ybu.edu.cn

Yanbian University

Research Article

Keywords: quantum information processing; multipartite entanglement; finite-shot Pauli measurements; conformal prediction; uncertainty calibration

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10146310/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Finite-Shot Pauli-Measurement Protocol for Calibrated Multipartite-Entanglement Diagnosis

Wenlong Sun^{1,2,3} and Yuanfeng Jin^{2,3}

¹School of Science, Shenyang University of Technology, Shenyang 110870, China

²Department of Physics, Yanbian University, Yanji 133001, China

³Institute of Quantum Science and Technology, Yanbian University, Yanji, Jilin 133002, China

Correspondence: Yuanfeng Jin, yfkim@ybu.edu.cn

Abstract

Near-term quantum experiments often provide only finite-shot records of selected Pauli measurements, rather than tomographically complete data. We formulate multipartite-entanglement diagnosis as a calibrated decision problem from such records. The protocol combines binomial Pauli estimators, task-informed observable designs, score-based diagnostic models, class-conditional conformal prediction, and out-of-distribution abstention. It outputs a single label, a prediction set, or an abstention according to the available evidence. Synthetic benchmarks show genuine multipartite entanglement diagnosis up to 20 qubits and reveal an information hierarchy: global coherence probes support coarse diagnosis, whereas finer block-structure labels require richer local and few-body Pauli information.

Keywords: quantum information processing; multipartite entanglement; finite-shot Pauli measurements; conformal prediction; uncertainty calibration

1 Introduction

Reliable characterization of multipartite entanglement is a basic task in quantum information processing [1–3], near-term quantum computation and simulation [4], and quantum metrology [5, 6]. In practice, however, larger devices seldom come with informationally complete data. What is often available is a chosen list of Pauli measurements, each measured with a finite number of shots. The question is then not simply “what is the state?”, but rather “what entanglement-structure claim is justified by this measurement record?” This is the information-limited regime considered in this work.

We therefore focus on an uncertainty-aware diagnostic protocol for quantum-information experiments. Its input is a fixed Pauli measurement design and a finite shot budget; its output is a calibrated statement about multipartite entanglement structure. When the evidence is strong, the protocol may return a single structural label. When the evidence is ambiguous, it should return a set of plausible labels. When the sample falls outside the calibrated regime, it should abstain. In this form, multipartite-entanglement diagnosis becomes finite-data inference from physically accessible Pauli records.

Multipartite entanglement also comes with a natural hierarchy of questions. Beyond asking whether a state is entangled, one may ask whether it has genuine multipartite entanglement (GME), what its entanglement depth is, how many separable blocks it contains, or which block-size partition best matches the observed correlations. These labels are useful for judging the operational value of many-qubit states, but they are not

equally visible from the same limited data. A recurring point in this paper is this information hierarchy: a few global coherence probes can be enough for coarse GME diagnosis, while fine partition labels require much more local and few-body correlation information.

Full quantum state tomography provides complete information but scales unfavorably with system size [7]. Entanglement witnesses and separability criteria provide rigorous and physically interpretable tools [1–3, 5, 6, 8], but they are often optimized for specific target states, observables, or prior assumptions. Resource-efficient schemes based on local measurements, graph-state structure, metrological hierarchy information, randomized measurements, classical shadows, and derandomized Pauli measurements [9–15] reduce measurement cost, while a separate decision layer is still needed to translate finite-shot estimates into calibrated multipartite-structure statements. A measurement-to-decision protocol must therefore connect finite-shot estimates to explicit decision-level uncertainty.

Learning-based methods have shown strong potential for quantum-state characterization, entanglement detection, and multipartite-structure recognition [16–26]. Many such methods, however, are primarily accuracy-oriented and return a single label even when the input is ambiguous or lies outside the calibrated training distribution. In finite-shot experiments, a confident but unsupported structural assignment can be more misleading than an explicit uncertain output. This motivates calibrated diagnosis: the protocol should decide when a singleton label is supported, when a larger prediction set is more

appropriate, and when the sample should be rejected or marked as unknown.

Here we develop such a finite-shot Pauli-measurement protocol for calibrated multipartite-entanglement diagnosis. The features are binomial finite-shot estimators of Pauli expectation values, so the method works directly with sampled observables rather than ideal density matrices. We use task-informed Pauli observables to reveal correlations relevant to the target structural label, and trainable score functions to summarize these features. The final decision layer combines class-conditional conformal prediction [27–29] with an OOD-calibrated abstention rule. It returns either a singleton label, a non-singleton prediction set, or an abstention. As in split conformal prediction, the coverage statement is tied to the calibration distribution and the corresponding exchangeability assumption. The protocol is meant to complement, not replace, witness-based certification: it supplies a measurement-efficient and uncertainty-aware diagnostic layer for Pauli-accessible experiments.

The measurement design is chosen with the physics of the task in mind. For scalable GME diagnosis, phase-aware global X/Y -type Pauli probes target the real and imaginary quadratures of GHZ-like coherences and help separate perturbed global GHZ-like states from large-block non-GME alternatives. For finer hierarchical tasks, we add identity-inclusive local and few-body probes, because block decompositions with similar global signatures can differ in their subsystem marginals and lower-order correlations. The paper makes three main contributions. First, it treats finite-shot Pauli measurements as a direct diagnostic input and asks when the measured evidence is sufficient for multipartite-structure decisions under native binomial shot noise. Second, it introduces task-informed Pauli designs for GME and hierarchical labels, pairing phase-sensitive global probes with GHZ coherence and identity-inclusive few-body probes with block-structure information; these measurements are then combined with calibrated prediction sets and OOD abstention for decision-level risk control. Third, the numerical benchmarks make the information hierarchy visible: GME diagnosis is possible up to $N = 20$ with a small set of physically motivated Pauli observables, whereas full block-size partition diagnosis needs substantially richer measurement information. As a first mixed-state robustness check, we also test white-noise admixtures in the Appendix. Table 1 places the protocol next to several representative measurement-based characterization approaches.

2 Finite-Shot Measurement Setting and Structure Labels

This section sets up the two objects used throughout the paper: the finite-shot Pauli feature vector and the multipartite-structure labels to be diagnosed. The labels should be read as calibrated diagnostic targets for the stated benchmark family and measurement protocol. We consider an N -qubit state ρ and infer multipartite entanglement information from a restricted set of Pauli observables. Let

$$P_j = \sigma_{j_1} \otimes \sigma_{j_2} \otimes \cdots \otimes \sigma_{j_N}$$

be the j -th selected observable. For global-correlation measurements, the local factors are chosen from $\{X, Y, Z\}$. For structure-sensitive measurements, identity factors are also allowed, so that the local factors are chosen from $\{I, X, Y, Z\}$, excluding the all-identity operator. The ideal expectation value is

$$m_j = \text{Tr}(\rho P_j).$$

In the proposed protocol, however, m_j is not used as an ideal input. Instead, each observable is estimated from finite-shot parity outcomes. If $n_{j,+}$ denotes the number of $+1$ parity outcomes in S shots, then

$$n_{j,+} \sim \text{Binomial}\left(S, \frac{1+m_j}{2}\right), \quad \hat{m}_j = \frac{2n_{j,+} - S}{S}.$$

For K selected Pauli observables, the diagnostic input is

$$x = (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_K) \in \mathbb{R}^K, \quad B = KS,$$

where B is the observable-level shot budget. Thus, the protocol is task-oriented: it uses a finite set of Pauli estimators to infer selected entanglement properties directly from the measurement record.

The decision targets are organized by multipartite entanglement structure. We describe these structures by integer partitions

$$\lambda = (\lambda_1, \dots, \lambda_h), \quad \lambda_1 \geq \cdots \geq \lambda_h, \quad \sum_{j=1}^h \lambda_j = N.$$

Each part λ_j represents the size of one entangled block in the benchmark construction. The entanglement depth and separability number are

$$w = \max_{1 \leq j \leq h} \lambda_j, \quad h = |\lambda|,$$

respectively, while the block-size partition label λ specifies the complete block-size composition. In this benchmark construction, these block-size labels define the entanglement-depth and separability-number targets used for supervised diagnosis. The diagnostic labels are $(y_{\text{GME}}, y_w, y_h, y_\lambda)$, corresponding to GME diagnosis, entanglement-depth diagnosis, separability-number diagnosis, and full block-size partition-label diagnosis. For the binary GME task, the single-block partition $\lambda = (N)$ is treated as the GME-positive class, and all other partitions are treated as non-GME structures.

For numerical benchmarking, each partition is represented by a block-product state

$$|\Psi_\lambda\rangle = |\psi_{\lambda_1}\rangle \otimes |\psi_{\lambda_2}\rangle \otimes \cdots \otimes |\psi_{\lambda_h}\rangle.$$

Blocks of size one are random single-qubit pure states. Blocks of size at least two are perturbed GHZ-like states,

$$|\text{GHZ}_{\lambda_j, a, \phi}\rangle = a|0\rangle^{\otimes \lambda_j} + \sqrt{1-a^2}e^{i\phi}|1\rangle^{\otimes \lambda_j}, \\ a \in (0, 1), \quad \phi \in [0, 2\pi).$$

After the block-product state is generated, the qubits are randomly permuted so that the model cannot rely on fixed qubit

Table 1: Conceptual comparison with representative measurement-based characterization approaches. The entry “uncertainty” refers to an explicit decision-level uncertainty output rather than only statistical error bars for estimated observables.

Method	Finite-shot input	Avoids full tomography	Decision-level uncertainty	OOD/unknown output	Multipartite-structure labels
Full tomography	Limited by scaling	No	Not intrinsic	No	Yes, after reconstruction
Entanglement witnesses	Yes	Yes	Target-dependent	Limited	Usually target-dependent
Classical shadows / randomized measurements	Yes	Yes	Observable-level guarantees	Not intrinsic	Indirect
This work	Yes	Yes	Calibrated prediction sets	Calibrated abstention	Direct diagnostic labels

positions. This controlled benchmark tests which structural decisions are supported by finite-shot Pauli measurements, from coarse GME diagnosis to finer entanglement-structure labels. The resulting block-size partition labels define the supervised diagnostic targets used in the numerical experiments. Accordingly, the reported labels are calibrated diagnostic labels relative to the specified benchmark family and measurement protocol.

3 Calibrated Diagnostic Method

This section describes the decision layer that converts finite-shot Pauli features into calibrated diagnostic outputs. The question is deliberately practical: after a fixed set of Pauli observables has been measured with finite shots, is the record strong enough to justify a single structural statement? The trainable model supplies diagnostic scores. Conformal prediction and abstention then decide whether those scores support one reliable label, a broader set of possible labels, or no decision within the calibrated regime.

3.1 Score-based diagnostic model and conformal prediction

Given the finite-shot feature vector

$$x = (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_K) \in \mathbb{R}^K,$$

we use a shared feature extractor with task-specific diagnostic heads. For each diagnostic task

$$r \in \{\text{GME}, w, h, \lambda\},$$

the model outputs the class-score distribution

$$q_\theta^r(y | x) = \text{softmax}(g_\theta^r(f_\theta(x))),$$

where the four heads correspond to GME diagnosis, entanglement-depth diagnosis, separability-number diagnosis, and full block-size partition-label diagnosis. For hierarchical diagnosis, the training loss is

$$\mathcal{L} = \mathcal{L}_{\text{GME}} + \mathcal{L}_w + \mathcal{L}_h + \mathcal{L}_\lambda,$$

whereas binary GME experiments use

$$\mathcal{L} = \mathcal{L}_{\text{GME}}.$$

Architecture and training details are given in the Appendix. The following conformal layer is independent of the particular score model, provided that the calibration and test data satisfy the stated exchangeability condition.

We apply class-conditional conformal prediction separately to each task. For a calibration sample (x_i, y_i) , the nonconformity score is

$$s_i^r = 1 - q_\theta^r(y_i | x_i).$$

For each class c , let n_c be the number of calibration samples with true label c , and define

$$k_c = \lceil (n_c + 1)(1 - \alpha) \rceil.$$

The class-wise threshold $\hat{q}_{\alpha,c}^r$ is the k_c -th order statistic of the class- c calibration scores when $k_c \leq n_c$, and $+\infty$ otherwise. The prediction set is

$$\mathcal{C}_\alpha^r(x) = \{c \in \mathcal{Y}_r : 1 - q_\theta^r(c | x) \leq \hat{q}_{\alpha,c}^r\}.$$

A singleton set is reported as a single-label diagnosis. A non-singleton or empty set is treated as uncertain in single-label evaluation. The guarantee is therefore at the decision level: it calibrates the reported label set under the stated sampling assumptions.

Proposition 1 (Class-conditional conformal coverage). *Fix a diagnostic task r and a class c . Suppose that the calibration samples with label c and the test sample conditional on $y_{\text{test}} = c$ are exchangeable. Then the split-conformal prediction set defined above satisfies*

$$\mathbb{P}[y_{\text{test}} \in \mathcal{C}_\alpha^r(x_{\text{test}}) | y_{\text{test}} = c] \geq 1 - \alpha,$$

up to the standard finite-sample conformal correction determined by the order statistic $k_c = \lceil (n_c + 1)(1 - \alpha) \rceil$.

A self-contained rank-based proof is given in the Appendix. The proposition calibrates the reported prediction sets for the distribution represented by the calibration data. If formal certification is needed, these outputs can be paired with witnesses or device-specific validation criteria. In all experiments, we set

$$\alpha = 0.1, \quad 1 - \alpha = 0.9.$$

3.2 Risk metrics and OOD-calibrated abstention

We summarize point-label performance by balanced accuracy and macro-F1, and risk-control performance by coverage, reject rate, accepted risk, and overconfident error. The point prediction used for balanced accuracy and macro-F1 is $\hat{y}^r(x) = \arg \max_{c \in \mathcal{Y}_r} q_\theta^r(c | x)$; balanced accuracy is the mean class-wise recall. For task r , the empirical coverage and reject rate are

$$\text{Coverage}^r = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{y_i \in \mathcal{C}_\alpha^r(x_i)\},$$

$$\text{Reject}^r = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|\mathcal{C}_\alpha^r(x_i)| \neq 1\}.$$

The overconfident error is

$$\text{OE}^r = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|\mathcal{C}_\alpha^r(x_i)| = 1, y_i \notin \mathcal{C}_\alpha^r(x_i)\},$$

which counts singleton predictions that exclude the true label. For the accepted set

$$\mathcal{A}^r = \{i : |\mathcal{C}_\alpha^r(x_i)| = 1\},$$

the accepted risk is

$$\text{Risk}_{\text{acc}}^r = 1 - \frac{1}{|\mathcal{A}^r|} \sum_{i \in \mathcal{A}^r} \mathbf{1}\{y_i \in \mathcal{C}_\alpha^r(x_i)\},$$

when $|\mathcal{A}^r| > 0$. Thus, rejection is treated as an uncertainty output, not as a failed classification.

To handle distribution shifts represented during calibration, we add an OOD-calibrated abstention layer. Let $R(x)$ be a novelty or risk score constructed from model confidence, conformal-set information, and distances in Pauli-feature or embedding space. A sample is marked as unknown when

$$R(x) > T^*.$$

The threshold is selected from in-distribution and OOD calibration samples by

$$T^* = \arg \max_T \text{Reject}_{\text{OOD-cal}}(T),$$

subject to

$$\text{Retain}_{\text{ID-cal}}(T) \geq \gamma.$$

In the main experiments, we use $\gamma = 0.90$. Samples with $R(x) \leq T^*$ are passed to the conformal layer, while samples with $R(x) > T^*$ are assigned to the abstention output. With this rule, samples that look like the calibrated OOD families are routed to an explicit unknown output. Its scope is limited to the OOD families represented in calibration; broader shift classes would require the corresponding calibration data and, when needed, additional risk scores.

4 Measurement Design

The risk-control layer can only work with the information present in the finite-shot feature vector. Measurement design is therefore the physical link between the quantum device and the statistical decision rule: it decides which aspects of multipartite structure are visible before learning or calibration enters. This section lists the Pauli observable families used in the experiments and explains why they are matched to the corresponding structure labels.

The basic principle is simple: spend the measurement budget on the correlations needed by the decision. Global, phase-sensitive Pauli strings are useful for GHZ-type GME diagnosis, while identity-inclusive local and few-body strings are better suited to entanglement depth, separability number, and block-size partitions. With this separation, the numerical comparisons are easier to interpret: changes in performance can be traced not only to the diagnostic model, but also to the information content of the measured observables.

4.1 Task-informed Pauli observable selection

The selected Pauli observables should reveal the correlations needed by each diagnostic task. For this reason, we compare random baselines with physically motivated designs, which lets us separate the role of the measured observables from the role of the learning model.

For GME-oriented experiments, the random strategy R samples full-weight Pauli strings from $\{X, Y, Z\}^{\otimes N}$. The global-plus-random strategy $G + R$ augments random strings with

$$X^{\otimes N}, \quad Y^{\otimes N}, \quad Z^{\otimes N},$$

and the hybrid global-depth strategy HGD further includes probes designed to improve coarse entanglement-depth diagnosis. For hierarchical structure diagnosis, where local and subsystem correlations are needed, we also use identity-inclusive Pauli strings from $\{I, X, Y, Z\}^{\otimes N}$, excluding the all-identity operator. The corresponding random baseline is denoted by RI .

The key design for the hard large- N GME benchmark is the phase-aware global-plus-random strategy $PG + R$. For

$$|\text{GHZ}_{N,a,\phi}\rangle = a|0\rangle^{\otimes N} + be^{i\phi}|1\rangle^{\otimes N}, \quad b = \sqrt{1-a^2},$$

a full-weight X/Y -type Pauli string P satisfies

$$\langle P \rangle = 2ab \cos\left(\phi - \frac{n_Y \pi}{2}\right),$$

where n_Y is the number of local Y operators in P . Different values of n_Y probe different quadratures of the unknown phase ϕ . The short derivation, based only on the action of X and Y on the two GHZ basis components, is given in the Appendix. The $PG + R$ strategy therefore includes several global X/Y -type probes, together with $Z^{\otimes N}$ and additional random full-weight strings. This design asks whether a small set of phase-sensitive global probes can recover coherence information that small random or phase-insensitive budgets may miss, and thereby

distinguish perturbed global GHZ-like states from large-block non-GME alternatives.

For full block-size partition-label diagnosis, global probes alone are insufficient because different block decompositions may have similar global signatures but different local and few-body correlations. We therefore use a structure-aware strategy SA, which combines global phase-aware probes, single-qubit local probes, pairwise probes, triple probes, and additional random identity-inclusive strings. This design mirrors the hierarchy of tasks: global coherence is most relevant for GME diagnosis, while local and few-body correlations become increasingly important for entanglement-depth, separability-number, and full block-size partition-label diagnosis. Thus, SA is not just a larger random feature set; it builds in the physical distinction between global coherence and subsystem-resolved structure.

For all strategies, the cost is quantified by the number K of selected Pauli observables and the number S of shots per observable. In the large- N GME experiments, we use as few as $K = 40$ observables with $S = 512$ shots per observable. For hierarchical structure diagnosis, we use $K = 80$ or $K = 160$, again with $S = 512$. In all cases, K is much smaller than the number of possible Pauli strings, 3^N for full-weight strings and 4^N when identity factors are allowed.

4.2 Pauli-measurement protocol and experimental workflow

The protocol does not require an offline tomographic reconstruction. It can be run directly as a Pauli-measurement-to-decision procedure. After state preparation, a prescribed set of Pauli observables is measured with finite shots. The empirical expectation values are then passed to a calibrated diagnostic rule, which either identifies a single entanglement-structure label, returns a small set of compatible labels, or abstains when the data are not informative enough. In this experimentally oriented workflow, the measured Pauli record is the resource-limited input and the reported label set is the calibrated diagnostic output.

A task-specific list of Pauli observables is first selected. Each observable is measured for S shots, and the raw bit-string outcomes are converted into parity estimators $\hat{m}_j$, forming the finite-shot feature vector $x = (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_K)$. The calibrated diagnostic model then maps x to conformal prediction sets. If the set is a singleton and the OOD-abstention score is below threshold, the protocol reports a single diagnostic label; otherwise, it reports a prediction set or an unknown output. This workflow is compatible with platforms where local Pauli measurements are available, including superconducting qubits, trapped ions, Rydberg-atom arrays, and photonic systems.

Here K denotes the number of Pauli observables used as diagnostic features. For identity-inclusive observables, several compatible correlators may in principle be estimated from the same local-basis bit-string data. Thus, $B = KS$ should be understood as a conservative observable-level shot accounting unless stated otherwise. Possible grouping of compatible observables can further reduce the number of distinct experimental basis

settings. The output has a direct operational meaning: it says whether the collected finite-shot Pauli data support a reliable label within the calibrated diagnostic regime. It can also guide later witness measurements, adaptive Pauli selection, or device-specific verification when stronger validation is required.

5 Numerical Experiments

We test the framework through four experiments, each built around a concrete decision question under limited measurement information. Can finite-shot Pauli data support reliable in-distribution GME diagnosis under noise? Can OOD-calibrated abstention reduce overconfident decisions on unseen entanglement families? Can a small number of physically designed Pauli observables support large- N GME decisions? Finally, how quickly does the measurement demand grow when the target changes from coarse GME labels to fine-grained multipartite structure? Together, these experiments follow the full path from finite-shot Pauli estimators to calibrated prediction sets, abstention, and structure-level diagnostic outputs.

5.1 General experimental setting

Unless otherwise stated, each selected Pauli observable is measured with $S = 512$ shots. The data are divided into training, calibration, and test subsets. The training subset fits the score-based diagnostic model, the in-distribution calibration subset determines conformal thresholds, and the held-out test subset is reserved for final evaluation. For OOD abstention, a separate OOD calibration subset sets the abstention threshold. Additional implementation details are provided in the Appendix.

We use class-conditional conformal prediction with $\alpha = 0.1$, corresponding to target coverage $1 - \alpha = 0.9$. We report balanced accuracy, coverage, reject rate, and overconfident error; accepted risk is used as an auxiliary diagnostic summary when needed. For multi-class structure diagnosis, macro-F1 is also reported when relevant. Rejection is reported as calibrated uncertainty, rather than counted as an incorrect label.

The measurement strategies R, G + R, HGD, PG + R, RI, and SA denote random, global-plus-random, hybrid global-depth, phase-aware global-plus-random, random identity-inclusive, and structure-aware measurements, respectively. The OOD abbreviations LC, RC, and RG denote linear-cluster, random-circuit, and ring-graph states.

5.2 In-distribution diagnosis and noise robustness

We begin with a basic question: does the finite-shot record contain enough information for reliable coarse in-distribution diagnosis? The test states are generated from product states, Bell-block states, GHZ-block states, and global GHZ states. For binary GME diagnosis,

$$y_{\text{GME}} = 1 \iff \lambda = (N),$$

while all other block-product structures are treated as non-GME. The experiment is performed for $N = 4, 5, 6$ under R, G + R, and HGD. Representative results are shown in Table 2.

In this calibrated in-distribution setting, Table 2 shows that global correlations already give strong evidence for GME decisions. With G + R and HGD, the GME balanced accuracy is close to or equal to 1.000, and the empirical coverage stays near the target level within finite-sample fluctuations. Depth diagnosis is harder because it needs block-size information, but HGD still improves on purely random measurements.

We also test three expectation-level noise models. Depolarizing noise maps $m_j \mapsto (1 - p)m_j$. Dephasing noise maps

$$m_j \mapsto (1 - 2p)^{n_{XY}(P_j)} m_j,$$

where $n_{XY}(P_j)$ is the number of local X or Y factors. Readout error attenuates global parity approximately as $m_j \mapsto (1 - 2p)^N m_j$. The robustness experiment uses HGD, $K = 40$, and $S = 512$.

The Appendix gives the complete noise-robustness values. Depolarizing noise has a relatively mild effect, while dephasing and readout noise suppress global X/Y -type and parity correlations more strongly. Even when balanced accuracy decreases, the conformal layer keeps GME coverage near the target level within finite-sample fluctuations, and the depth-task reject rate rises under stronger noise. The Appendix also includes a white-noise mixed-state check, where GME performance remains stable over the tested admixture strengths.

5.3 OOD families and calibrated abstention

We next test what happens under distribution shift. The main diagnostic model is trained on product states, Bell-block states, GHZ-block states, and global GHZ states. The OOD families are W, Dicke, linear-cluster, ring-graph, and random-circuit states; these families are not used in the main training stage.

We first use a multi-score OOD baseline based on Pauli-space distances, embedding-space distances, kNN variants, softmax confidence, and conformal set size. The baseline removes some high-confidence mistakes, but it still leaves substantial overconfident errors for W and random-circuit states. For example,

$$\text{OE}_W = 0.790, \quad \text{OE}_{RC} = 0.676.$$

OOD-calibrated abstention then sets a rejection threshold from the designated OOD calibration subset. The main operating point is $\gamma = 0.90$.

Table 3: OOD-calibrated abstention results with $\gamma = 0.90$. AR, FA, FR, and OE denote abstention reject rate, final accept rate, final reject rate, and final GME overconfident error, respectively.

Family	AR	FA	FR	OE
ID	0.335	0.587	0.413	0.000
Dicke	1.000	0.000	1.000	0.000
LC	1.000	0.000	1.000	0.000
RC	0.948	0.048	0.952	0.048
RG	1.000	0.000	1.000	0.000
W	1.000	0.000	1.000	0.000

Here, LC, RC, and RG denote linear-cluster, random-circuit, and ring-graph states, respectively.

Table 3 shows the effect of adding OOD-calibrated abstention when the test OOD families are represented in calibration. For W states, the overconfident error drops from 0.790 to 0; for random-circuit states, it drops from 0.676 to 0.048. This improvement comes with a lower in-distribution accept rate: at $\gamma = 0.90$, the final in-distribution accept rate is 0.587. Here, γ controls the ID-retention constraint used to set the abstention threshold, whereas the final ID accept rate is computed after both OOD abstention and conformal-set rejection; therefore, the final accept rate can be lower than γ .

Within the calibrated OOD setting, the abstention layer therefore works as a tunable safeguard against distribution shift. The Appendix gives additional OOD baselines, γ -sensitivity results, and leave-one-family-out validation; there, random-circuit states and several $N = 6$ cases are the hardest transfer tests.

5.4 Large- N scalability of the GME module

We then test whether the decision-centered GME module can move beyond the small-system setting, where the number of possible Pauli observables grows exponentially but the measured subset remains small. For full-weight Pauli strings, the number of possible observables grows as 3^N , reaching

$$3^{20} = 3486784401$$

at $N = 20$. The goal is to determine whether a small, physically designed Pauli set contains enough information to support a reliable GME decision against large-block non-GME alternatives.

As a favorable baseline, ideal GHZ states are correctly diagnosed for $N = 8, 10, 12, 16, 20$ using G + R. We then consider a harder benchmark. The positive class consists of perturbed GHZ-like states

$$|\text{GHZ}_{N,a,\phi}\rangle = a|0\rangle^{\otimes N} + \sqrt{1-a^2}e^{i\phi}|1\rangle^{\otimes N},$$

with $a \in [0.60, 0.80]$ and $\phi \in [0, 2\pi)$. The negative class includes large-block non-GME states such as

$$\text{GHZ}_{N-1} \otimes |\psi\rangle, \quad \text{GHZ}_{N-2} \otimes \text{Bell},$$

and

$$\text{GHZ}_{\lfloor N/2 \rfloor} \otimes \text{GHZ}_{\lceil N/2 \rceil}.$$

This controlled benchmark isolates GHZ-coherence discrimination against strong non-GME multipartite correlations.

For this task, we use PG + R. A full-weight X/Y -type Pauli string satisfies

$$\langle P \rangle = 2ab \cos\left(\phi - \frac{n_Y \pi}{2}\right), \quad b = \sqrt{1-a^2},$$

so different values of n_Y probe different quadratures of the unknown phase.

Fig. 1 shows that the GME module remains effective up to $N = 20$. Complete tabulated values for this hard benchmark

Table 2: In-distribution finite-shot diagnosis. BAcc, Cov., and Rej. denote balanced accuracy, conformal coverage, and reject rate, respectively.

N	Str.	GME task			Depth task
		BAcc	Cov.	Rej.	BAcc / Rej.
4	R	0.925	0.915	0.180	0.759 / 0.442
4	G+R	1.000	0.918	0.082	0.899 / 0.173
4	HGD	1.000	0.922	0.078	0.899 / 0.158
5	R	0.886	0.911	0.228	0.597 / 0.689
5	G+R	0.998	0.914	0.085	0.794 / 0.431
5	HGD	0.999	0.912	0.088	0.816 / 0.382
6	R	0.831	0.909	0.318	0.433 / 0.856
6	G+R	1.000	0.907	0.093	0.650 / 0.679
6	HGD	1.000	0.910	0.090	0.680 / 0.634

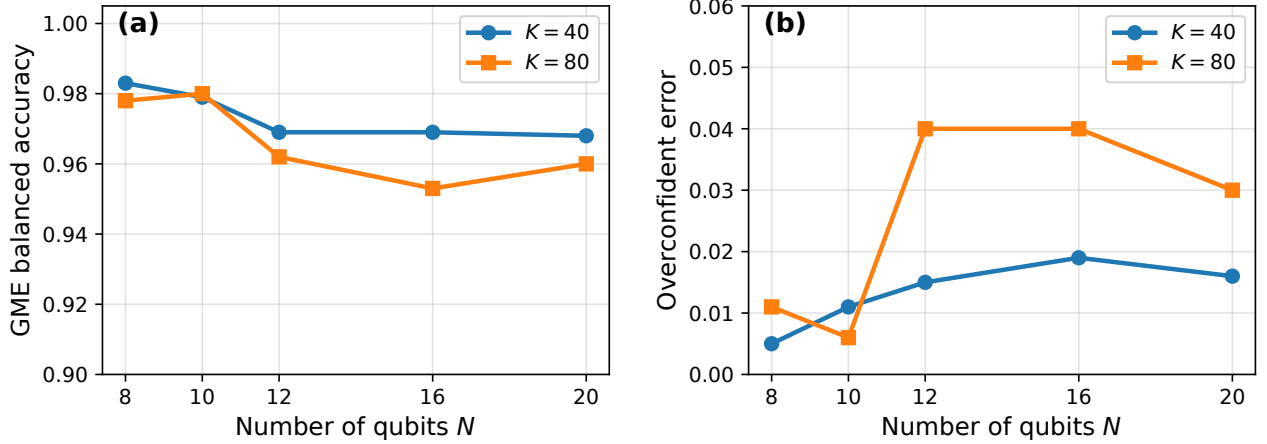


Fig. 1: Large- N GME performance on the hard benchmark using the phase-aware global-plus-random strategy.

are provided in the Appendix. With $K = 40$ and $S = 512$, the model achieves

$$\text{BAcc} = 0.968, \quad \text{OE} = 0.016$$

at $N = 20$, with empirical coverage fluctuating around the target level. Within this benchmark, the method separates GME-positive states from large-block non-GME alternatives under a finite measurement budget. The $K = 40$ and $K = 80$ observable sets are generated independently and are not necessarily nested, and the reported values include finite-sample training and measurement-design variability; therefore, monotonic improvement with K is not expected for every entry.

Table 4: Phase-aware measurement ablation on the hard benchmark. R, G+R, and PG+R denote random Pauli, global-plus-random, and phase-aware global-plus-random measurements, respectively. This table is computed from an independent ablation run, so the PG+R entry can differ slightly from the large- N sweep reported in Fig. 1 and the Appendix.

N	Str.	BAcc	Cov.	Rej.	OE
12	R	0.517	0.892	0.757	0.108
12	G+R	0.686	0.896	0.465	0.104
12	PG+R	0.975	0.908	0.082	0.010
20	R	0.503	0.907	0.816	0.093
20	G+R	0.672	0.906	0.521	0.094
20	PG+R	0.965	0.896	0.088	0.017

Table 4 confirms that the gain comes from phase-aware

measurement design. Under the small measurement budgets tested here, random Pauli measurements are close to random guessing, and the basic global-plus-random design gives only moderate improvement. In contrast, PG + R achieves high balanced accuracy and low overconfident error for both $N = 12$ and $N = 20$. This comparison shows that the large- N result is driven by phase-aware measurement information.

5.5 Hierarchical entanglement-structure diagnosis

Finally, we ask how the problem changes when the target label becomes more detailed. We extend the framework from binary GME diagnosis to hierarchical structure diagnosis, including GME, entanglement-depth, separability-number, and full block-size partition-label tasks. The experiments cover $N = 4, \dots, 8$, using the full integer-partition label set for each N ; the number of block-size partition labels therefore increases from 5 to 22.

For structure diagnosis, we compare random identity-inclusive measurements RI with the structure-aware strategy SA. The latter combines global phase-aware probes, local probes, pairwise probes, triple probes, and additional random probes. Class-conditional conformal prediction is applied separately to each task with $\alpha = 0.1$.

Fig. 2 makes clear that measurement design is crucial for

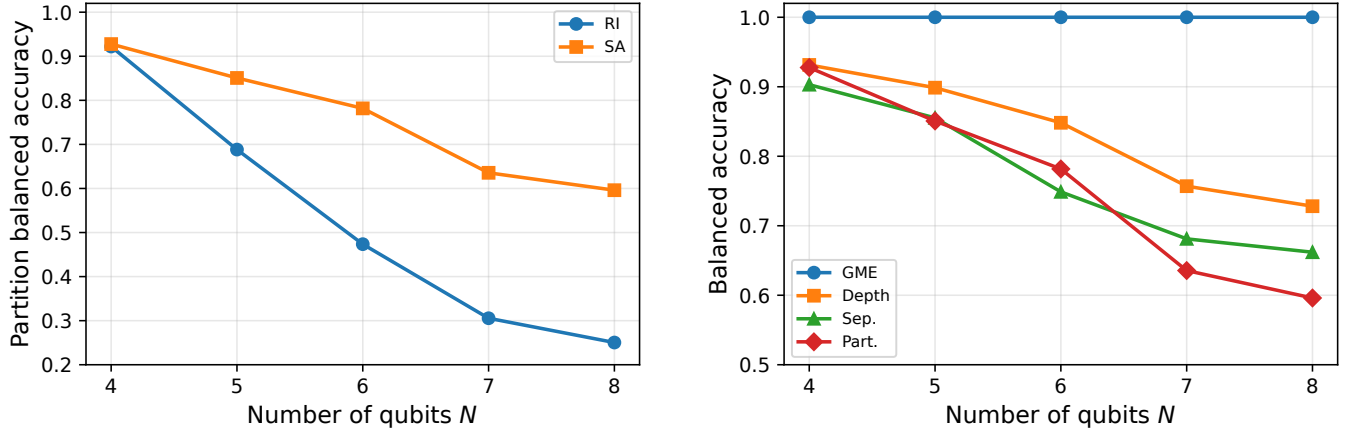


Fig. 2: Hierarchical structure diagnosis under finite-shot Pauli measurements. Left: full block-size partition-label diagnosis under random identity-inclusive (RI) and structure-aware (SA) measurements. Right: balanced accuracy for GME, entanglement-depth, separability-number, and full partition-label diagnosis under the structure-aware strategy.

fine-grained diagnosis. As N increases, RI deteriorates rapidly, while SA remains much more effective. At $N = 8$, the full block-size partition-label balanced accuracy improves from 0.250 under RI to 0.596 under SA, while coverage remains near or above the target level.

The same figure reveals a clear information hierarchy in measurement-limited diagnosis. Depth diagnosis is more reliable than separability-number and full block-size partition-label diagnosis, since the latter tasks require increasingly detailed structural information. Under SA, the GME balanced accuracy remains 1.000 and the GME overconfident error remains 0 for all $N = 4, \dots, 8$. Additional tabulated values for the partition-label and hierarchical tasks are provided in the Appendix.

For $N = 8$, $K = 160$, and SA, the full block-size partition-label task has

$$\begin{aligned} \text{Part. BAcc} &= 0.596, & \text{Part. Cov} &= 0.953, \\ \text{Part. OE} &= 0.008. \end{aligned}$$

The high reject rate,

$$\text{Part. Rej.} = 0.899,$$

shows that this measurement budget often leaves several fine-grained block-size partition labels statistically plausible. Additional measurement-budget studies are provided in the Appendix.

Taken together, the experiments show that coarse GME decisions are robust under physically designed measurements, while full block-size partition-label decisions become increasingly demanding as N grows. In this sense, the protocol does more than assign labels: it also shows which entanglement questions are supported by the available finite-shot Pauli evidence.

6 Discussion and outlook

The experiments support a measurement-to-decision view of multipartite-entanglement diagnosis. The finite-shot Pauli

record is the operational evidence, and the reported label set is a calibrated statement about what that evidence supports. The protocol separates three cases that ordinary classifiers often blur together: a single structural decision is justified, several structures remain plausible, or the sample lies outside the calibrated operating regime and should trigger abstention. This separation is important for using learning-based diagnostics responsibly in finite-data quantum experiments.

One main lesson is the diagnostic information hierarchy exposed by the benchmarks. Coarse GME diagnosis is comparatively stable when phase-sensitive global Pauli probes are available. Entanglement depth, separability number, and full block-size partition labels need progressively richer local and few-body information. The large- N hard benchmark shows the same point from the measurement-design side: with a small fixed measurement budget, random Pauli measurements and basic global probes do not reliably capture the GHZ phase information, whereas phase-aware X/Y -type global probes directly target the relevant coherence. The performance gains therefore do not come only from the score model; they also reflect the match between the measured physical correlations and the structural question being asked.

The same viewpoint also clarifies the scope of the method. Split conformal prediction controls in-distribution prediction sets under the usual exchangeability assumption. OOD-calibrated abstention extends the workflow to represented distribution shifts by adding an explicit rejection threshold. Within this calibration setting, the protocol provides calibrated diagnostic evidence under a specified Pauli-measurement model and calibration distribution. It can then serve as a diagnostic layer before stronger tools such as certification, adaptive measurement, or device-specific verification. The leave-one-family-out tests in the Appendix identify the most difficult transfers, especially random-circuit states and several $N = 6$ cases.

The benchmarks also point to several next steps. The state families used here are synthetic and mainly pure-state based; the white-noise admixture test in the Appendix is only a first mixed-state robustness check. Broader tests

on noisy many-body ensembles, variational-circuit states, matrix-product states, stabilizer and graph-state families, and hardware-generated states would help clarify the empirical operating range. The measurement designs used here are also fixed and manually constructed. Adaptive Pauli measurement selection, grouping of compatible observables, device-specific readout calibration, and integration with witness-based certification are natural extensions. These directions would bring efficient statistical diagnosis closer to formal quantum-state verification while keeping the finite-measurement, uncertainty-aware viewpoint developed here.

7 Conclusion

We have treated multipartite-entanglement diagnosis from finite-shot Pauli data as a calibrated quantum-information inference problem. The protocol uses binomial Pauli estimators as native experimental inputs and converts them into uncertainty-

aware diagnostic outputs through task-informed measurement design, score-based diagnostic modeling, class-conditional conformal prediction, and OOD-calibrated abstention. It reports whether the available finite measurement evidence supports a single structural conclusion, a broader prediction set, or no decision within the represented calibration regime.

In the benchmark families studied here, phase-aware global Pauli probes support GME diagnosis up to $N = 20$, while structure-aware local and few-body probes show the larger measurement demand of hierarchical entanglement-structure diagnosis. The results reveal an intrinsic information hierarchy: coarse global entanglement structure can be inferred from carefully chosen global coherence probes, whereas finer labels such as entanglement depth, separability number, and complete block-size partitions require substantially richer Pauli information. The protocol therefore gives a calibrated diagnostic layer for finite-measurement quantum experiments and is naturally complementary to tomography, entanglement witnesses, and device-independent certification.

Appendix

A Additional numerical implementation and calibration details

The main text defines the finite-shot Pauli estimators, diagnostic labels, conformal prediction sets, OOD-abstention rule, and measurement strategies. Here we give the implementation choices needed to read the Appendix results. All numerical experiments were carried out with Python programs written by the authors, and no external datasets were used. The main diagnostic model is trained on in-distribution benchmark samples, while OOD samples are kept for abstention calibration or final OOD testing. Table A1 summarizes the settings.

The train/calibration/test split is kept disjoint in all Python-based experiments. Unless a table states otherwise, the reported entries are computed on held-out test samples after the conformal and abstention thresholds have already been fixed. The Pauli feature vector contains finite-shot parity estimators only; it does not use ideal density matrices or full tomography data. Formal runs use 2000 in-distribution states per N for the small-system in-distribution, noise, and OOD experiments, 600 samples per OOD family for OOD calibration and testing, 4000 states per N for the large- N GME benchmark, and 3200 states per N for hierarchical diagnosis. Results are summarized across five random repeats. For the large- N GME experiments, the $K = 40$ and $K = 80$ Pauli lists are generated as independent measurement designs rather than nested supersets. Small nonmonotonic differences across K should therefore be read as measurement-design and training variability.

The abstention score combines the maximum softmax score, predictive entropy, conformal-set size, distances in the raw Pauli-feature space, and distances in the learned embedding space. Its threshold is chosen from the ID calibration split and the designated OOD calibration families. Thus, the reported OOD results should be interpreted for the tested shift families.

Table A1: Compact implementation settings.

Category	Setting
Experiment ranges	The small-system in-distribution, noise, and OOD experiments use $N = 4, 5, 6$; the large- N GME benchmark uses $N = 8, 10, 12, 16, 20$; hierarchical diagnosis uses $N = 4, \dots, 8$. The main reported budgets use $K = 40, 80$ for GME-oriented tests and $K = 80, 160$ for hierarchical diagnosis. Formal runs use five random repeats; the train/calibration/test split is 60/20/20 for in-distribution samples, and the OOD calibration/test split is 50/50 for OOD families.
Model	The diagnostic model is a two-hidden-layer fully connected network with task-specific output heads. Hidden widths are 128 for the small-system and large- N experiments and 160 for the hierarchical diagnosis model.
Training	Models are trained with Adam, learning rate 10^{-3} , batch size 64, and class-weighted cross-entropy losses. Calibration is applied after training through split conformal prediction and OOD-calibrated abstention.
OOD abstention	The abstention module combines finite-shot Pauli features with confidence, entropy, conformal-set-size, and distance-based indicators to produce a thresholded OOD score.

B Theoretical derivations

This section gives two short derivations behind the calibrated decision layer and the phase-aware Pauli-measurement design used in the main text. The purpose is to separate the statistical calibration statement from the measurement features that are tied to GHZ-type quantum coherences.

B.1 Rank proof of class-conditional conformal coverage

Fix a diagnostic task and a class c . Let

$$S_i = 1 - q_\theta(c \mid X_i), \quad i = 1, \dots, n_c,$$

be the calibration nonconformity scores among samples whose true label is c , and let

$$S_{\text{test}} = 1 - q_\theta(c \mid X_{\text{test}})$$

be the corresponding score for a test sample conditional on $Y_{\text{test}} = c$. Under the stated class-conditional exchangeability assumption,

$$S_1, \dots, S_{n_c}, S_{\text{test}}$$

are exchangeable random variables. Let $S_{(1)} \leq \dots \leq S_{(n_c)}$ be the ordered calibration scores and set

$$k_c = \lceil (n_c + 1)(1 - \alpha) \rceil.$$

If $k_c > n_c$, the threshold is taken as $+\infty$, so the true class is included with probability one. Otherwise, the class-wise threshold is $\hat{q}_{\alpha,c} = S_{(k_c)}$.

The conformal set contains the true label c exactly when

$$S_{\text{test}} \leq \widehat{q}_{\alpha,c}.$$

By exchangeability, the rank of S_{test} among the $n_c + 1$ scores is uniformly distributed in the no-tie case; with ties, the same argument gives conservative coverage when ties are handled by the non-strict inequality above. Therefore

$$\mathbb{P}(S_{\text{test}} \leq S_{(k_c)} \mid Y_{\text{test}} = c) \geq \frac{k_c}{n_c + 1} \geq 1 - \alpha.$$

Thus

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}}) \mid Y_{\text{test}} = c) \geq 1 - \alpha.$$

This is a finite-sample calibration statement for the diagnostic prediction set under class-conditional exchangeability. Entanglement-witness or device-independent certification can be added when the required experimental assumptions and validation criteria are available.

B.2 Expectation of phase-aware global X/Y Pauli probes

Consider the GHZ-like state

$$|\text{GHZ}_{N,a,\phi}\rangle = a|0\rangle^{\otimes N} + be^{i\phi}|1\rangle^{\otimes N}, \quad b = \sqrt{1 - a^2}.$$

Let $P = \bigotimes_{j=1}^N P_j$ be a full-weight Pauli string with $P_j \in \{X, Y\}$, and let n_Y be the number of local Y factors. Since

$$X|0\rangle = |1\rangle, \quad X|1\rangle = |0\rangle, \quad Y|0\rangle = i|1\rangle, \quad Y|1\rangle = -i|0\rangle,$$

we have

$$P|0\rangle^{\otimes N} = i^{n_Y}|1\rangle^{\otimes N}, \quad P|1\rangle^{\otimes N} = (-i)^{n_Y}|0\rangle^{\otimes N}.$$

The diagonal overlaps vanish because P flips every computational-basis bit. Hence only the two coherence terms contribute:

$$\begin{aligned} \langle P \rangle &= \langle \text{GHZ}_{N,a,\phi} | P | \text{GHZ}_{N,a,\phi} \rangle \\ &= ab [e^{i\phi} (-i)^{n_Y} + e^{-i\phi} i^{n_Y}] \\ &= 2ab \cos\left(\phi - \frac{n_Y \pi}{2}\right). \end{aligned}$$

Thus changing n_Y selects different phase quadratures of the GHZ coherence. This explains why the phase-aware global-plus-random design includes several full-weight X/Y -type probes instead of relying only on $X^{\otimes N}$, $Y^{\otimes N}$, or random full-weight Pauli strings.

C Robustness and out-of-distribution evaluation

C.1 Noise robustness

Table A2 gives the numerical values behind the noise-robustness trends in the main text. The results use the hybrid global-depth strategy with $K = 40$ and $S = 512$. Depolarizing noise has a relatively mild effect, while dephasing and readout noise affect depth-related diagnosis more strongly. GME coverage remains near the target level within finite-sample fluctuations, and the depth-task reject rate increases under harder noise conditions.

C.2 Mixed-state white-noise stress test

As a simple mixed-state robustness check, each pure benchmark state $|\psi\rangle\langle\psi|$ is replaced by

$$\rho_\eta = (1 - \eta)|\psi\rangle\langle\psi| + \eta \frac{I}{2^N},$$

before finite-shot Pauli sampling. This test checks how sensitive the calibrated GME module is to a basic white-noise admixture. Table A3 shows stable GME balanced accuracy, coverage, and overconfident error over the tested range $\eta \leq 0.20$.

Table A2: Noise robustness under the hybrid global-depth strategy with $K = 40$ and $S = 512$. Depol. and Deph. denote depolarizing and dephasing noise, respectively.

N	Noise	p	GME task		Depth task
			BAcc	Cov.	BAcc / Rej.
4	Clean	–	1.000	0.919	0.954 / 0.054
4	Depol.	0.10	1.000	0.922	0.954 / 0.052
4	Deph.	0.10	1.000	0.916	0.702 / 0.402
4	Readout	0.10	0.700	0.908	0.665 / 0.600
5	Clean	–	1.000	0.902	0.870 / 0.262
5	Depol.	0.10	1.000	0.888	0.865 / 0.294
5	Deph.	0.10	0.501	0.906	0.421 / 0.876
5	Readout	0.10	0.500	0.905	0.279 / 0.939
6	Clean	–	1.000	0.908	0.764 / 0.522
6	Depol.	0.10	1.000	0.906	0.754 / 0.538
6	Deph.	0.10	0.694	0.916	0.363 / 0.940
6	Readout	0.10	0.585	0.901	0.395 / 0.972

Table A3: White-noise mixed-state stress test for the GME module. Here η is the white-noise admixture strength in $\rho_\eta = (1 - \eta)|\psi\rangle\langle\psi| + \eta I/2^N$.

N	η	BAcc	Cov.	Rej.	OE
4	0.00	1.000	0.883	0.117	0.000
4	0.05	1.000	0.892	0.108	0.000
4	0.10	0.999	0.887	0.113	0.000
4	0.20	0.999	0.923	0.077	0.000
5	0.00	0.998	0.891	0.109	0.000
5	0.05	0.996	0.896	0.104	0.000
5	0.10	0.999	0.907	0.093	0.000
5	0.20	0.997	0.925	0.075	0.000
6	0.00	0.998	0.913	0.087	0.000
6	0.05	0.999	0.919	0.081	0.000
6	0.10	0.999	0.896	0.104	0.000
6	0.20	0.999	0.917	0.083	0.000

Table A4: Multi-score OOD rejection baseline. NR, FR, and OE denote novelty reject rate, final GME reject rate, and final GME overconfident error, respectively.

Family	NR	FR	OE
ID	0.052	0.121	0.000
Dicke	0.668	0.668	0.332
LC	0.412	0.415	0.585
RC	0.285	0.324	0.676
RG	0.324	0.324	0.676
W	0.210	0.210	0.790

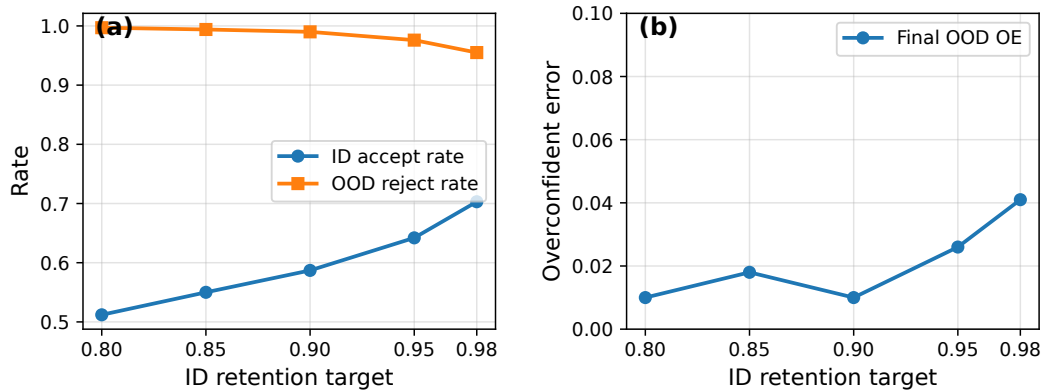


Fig. A1: Sensitivity of the OOD-calibrated abstention mechanism with respect to the ID-retention target γ . (a) ID accept rate and OOD reject rate as functions of γ . (b) Final OOD overconfident error under different choices of γ .

C.3 OOD baseline and sensitivity analysis

Table A4 reports a multi-score OOD rejection baseline based on Pauli-space distances, embedding-space distances, kNN variants, softmax confidence, and conformal set size. The baseline rejects part of the OOD samples, but it still leaves large overconfident errors for W and random-circuit states. This motivates the OOD-calibrated abstention mechanism used in the main text.

Fig. A1 shows stable OOD rejection and low final OOD overconfident error over $\gamma = 0.80, \dots, 0.98$, which indicates that the operating point is not sensitive to these threshold choices.

As an additional cross-family check, we perform leave-one-family-out OOD validation. For each held-out family among W, Dicke, LC, RG, and RC, the abstention layer and threshold are calibrated using ID calibration samples together with OOD calibration samples from the other four families. The held-out family is reserved for final testing. Table A5 reports the resulting performance. Here, ID retain means the fraction of in-distribution calibration samples retained after abstention thresholding; FA, FR, and OE denote final accept rate, final reject rate, and final GME overconfident error, respectively.

Table A5: Leave-one-family-out OOD validation. For each held-out family, the abstention layer is calibrated using all other OOD families and then tested on the held-out family only.

N	Held-out	ID retain	OOD reject	FA	FR	OE
4	W	0.905	1.000	0.000	1.000	0.000
4	Dicke	0.905	1.000	0.000	1.000	0.000
4	LC	0.905	0.969	0.031	0.969	0.031
4	RG	0.905	1.000	0.000	1.000	0.000
4	RC	0.905	0.689	0.258	0.742	0.258
5	W	0.905	1.000	0.000	1.000	0.000
5	Dicke	0.905	1.000	0.000	1.000	0.000
5	LC	0.905	0.942	0.042	0.958	0.042
5	RG	0.905	0.983	0.017	0.983	0.017
5	RC	0.905	0.703	0.272	0.728	0.272
6	W	0.905	0.653	0.347	0.653	0.347
6	Dicke	0.905	0.853	0.147	0.853	0.147
6	LC	0.905	0.892	0.108	0.892	0.108
6	RG	0.905	0.750	0.250	0.750	0.250
6	RC	0.905	0.453	0.503	0.497	0.503

Table A5 shows effective cross-family abstention for several held-out families, especially W, Dicke, LC, and RG at $N = 4, 5$. Random-circuit states and several $N = 6$ cases are harder; for example, held-out RC at $N = 6$ gives $OE = 0.503$.

D Large-system benchmarks

We also provide additional values for the hard perturbed-GHZ benchmark. As a reference check, the GME module is first evaluated on ideal GHZ states.

For the ideal-GHZ benchmark using the global-plus-random strategy, all tested cases $N = 8, 10, 12, 16, 20$ with $K = 40, 80$ achieve GME balanced accuracy 1.000 and overconfident error 0.000, with empirical coverage fluctuating around the target level. For this reason, the main text focuses on the harder perturbed-GHZ benchmark.

Table A6 gives the complete numerical values for the hard perturbed-GHZ benchmark summarized graphically in the main text.

Table A6: Hard perturbed-GHZ benchmark using the phase-aware global-plus-random strategy. BAcc, Cov., Rej., and OE denote balanced accuracy, conformal coverage, reject rate, and overconfident error, respectively.

N	K	BAcc	Cov.	Rej.	OE
8	40	0.983	0.913	0.082	0.005
10	40	0.979	0.907	0.081	0.011
12	40	0.969	0.906	0.079	0.015
16	40	0.969	0.905	0.077	0.019
20	40	0.968	0.917	0.068	0.016
8	80	0.978	0.906	0.082	0.011
10	80	0.980	0.883	0.110	0.006
12	80	0.962	0.896	0.064	0.040
16	80	0.953	0.913	0.047	0.040
20	80	0.960	0.919	0.051	0.030

E Hierarchical and measurement-budget diagnostics

E.1 Detailed hierarchical-structure values

Table A7 gives the full block-size partition-label balanced accuracy and coverage for the structure-diagnosis trend shown in the main text. Table A8 gives additional hierarchical-structure values under structure-aware measurements, including the partition-label rejection and overconfident-error rates.

Table A7: Full block-size partition-label diagnosis with $K = 160$ and $S = 512$. RI and SA denote random identity-inclusive and structure-aware measurements, respectively.

N	Str.	BAcc	Cov.
4	RI	0.922	0.918
4	SA	0.928	0.918
5	RI	0.688	0.935
5	SA	0.851	0.932
6	RI	0.474	0.944
6	SA	0.782	0.935
7	RI	0.306	0.959
7	SA	0.636	0.939
8	RI	0.250	0.927
8	SA	0.596	0.953

Table A8: Hierarchical entanglement-structure diagnosis using structure-aware measurements with $K = 160$ and $S = 512$. Part. denotes full partition-label balanced accuracy. Part. Rej. and Part. OE are the corresponding rejection and overconfident-error rates.

N	GME	Depth	Sep.	Part.	Cov.	Rej.	OE
4	1.000	0.931	0.903	0.928	0.918	0.162	0.039
5	1.000	0.899	0.855	0.851	0.932	0.418	0.031
6	1.000	0.848	0.749	0.782	0.935	0.497	0.020
7	1.000	0.757	0.681	0.636	0.939	0.796	0.016
8	1.000	0.728	0.662	0.596	0.953	0.899	0.008

E.2 Measurement-budget dependence

Fig. A2 summarizes the measurement-budget dependence at $N = 8$ under the structure-aware strategy. The GME task reaches stable high performance with a moderate measurement budget, while finer-grained structure tasks improve more slowly. For full partition-label diagnosis, coverage remains near the target level within finite-sample fluctuations and overconfident error stays low, but the reject rate is high when the budget is limited.

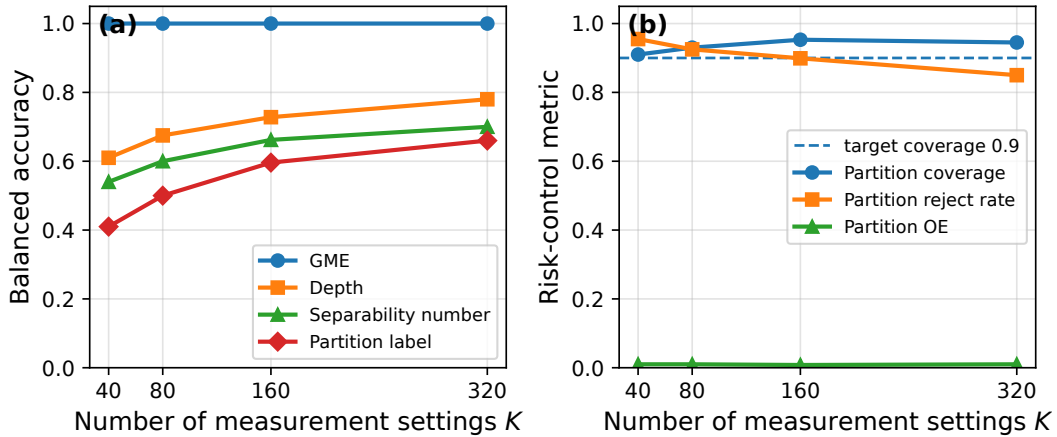


Fig. A2: Measurement-budget dependence at $N = 8$ under the structure-aware strategy. (a) Balanced accuracy for GME, entanglement-depth, separability-number, and full partition-label diagnosis. (b) Calibration metrics for the full partition-label task, including conformal coverage, reject rate, and overconfident error.

Statements and Declarations

Funding. This work was supported by the Natural Science Foundation of Jilin Province under Grant Nos. YDZJ202501ZYTS584, YDZJ202602CXJD091, and 20250102018JC. This work was also supported by the National Natural Science Foundation of China under Grant No. 12526104 and the Liaoning Provincial Joint Science and Technology Program under Grant Nos. 2024-MSLH-353 and 2025-MSLH-541.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics, Consent to Participate, and Consent to Publish declarations. Not applicable.

Data availability. No external datasets were used in this study. The numerical results reported in this paper were generated using Python programs written by the authors. Further details on the implementation are available from the corresponding author upon reasonable request.

Use of AI tools. During the preparation of this manuscript, AI-assisted tools were used for language editing and wording suggestions, and for assistance in drafting or debugging parts of the numerical and plotting code. The authors designed the scientific content, specified the algorithms and parameters, reviewed and modified the AI-assisted outputs, and verified the formulas, code, figures, numerical results, and conclusions. The authors take full responsibility for the final manuscript.

References

- [1] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, “Quantum entanglement,” *Rev. Mod. Phys.* **81**, 865–942 (2009). doi: 10.1103/RevModPhys.81.865.
- [2] O. Gühne and G. Tóth, “Entanglement detection,” *Phys. Rep.* **474**, 1–75 (2009). doi: 10.1016/j.physrep.2009.02.004.
- [3] N. Friis, G. Vitagliano, M. Malik, and M. Huber, “Entanglement certification from theory to experiment,” *Nat. Rev. Phys.* **1**, 72–87 (2019). doi: 10.1038/s42254-018-0003-5.
- [4] J. Preskill, “Quantum computing in the NISQ era and beyond,” *Quantum* **2**, 79 (2018). doi: 10.22331/q-2018-08-06-79.
- [5] G. Tóth, “Multipartite entanglement and high-precision metrology,” *Phys. Rev. A* **85**, 022322 (2012). doi: 10.1103/PhysRevA.85.022322.
- [6] P. Hyllus, W. Laskowski, R. Krischek, C. Schwemmer, W. Wieczorek, H. Weinfurter, L. Pezzè, and A. Smerzi, “Fisher information and multiparticle entanglement,” *Phys. Rev. A* **85**, 022321 (2012). doi: 10.1103/PhysRevA.85.022321.
- [7] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, “Efficient quantum state tomography,” *Nat. Commun.* **1**, 149 (2010). doi: 10.1038/ncomms1147.
- [8] S. Szalay, “Multipartite entanglement measures,” *Phys. Rev. A* **92**, 042329 (2015). doi: 10.1103/PhysRevA.92.042329.
- [9] G. Tóth and O. Gühne, “Detecting genuine multipartite entanglement with two local measurements,” *Phys. Rev. Lett.* **94**, 060501 (2005). doi: 10.1103/PhysRevLett.94.060501.
- [10] Y. Zhou, Q. Zhao, X. Yuan, and X. Ma, “Detecting multipartite entanglement structure with minimal resources,” *npj Quantum Inf.* **5**, 83 (2019). doi: 10.1038/s41534-019-0200-9.
- [11] Z. Ren, W. Li, A. Smerzi, and M. Gessner, “Metrological detection of multipartite entanglement from Young diagrams,” *Phys. Rev. Lett.* **126**, 080502 (2021). doi: 10.1103/PhysRevLett.126.080502.
- [12] T. Brydges, A. Elben, P. Jurcevic, B. Vermersch, C. Maier, B. P. Lanyon, P. Zoller, R. Blatt, and C. F. Roos, “Probing Rényi entanglement entropy via randomized measurements,” *Science* **364**, 260–263 (2019). doi: 10.1126/science.aau4963.
- [13] H.-Y. Huang, R. Kueng, and J. Preskill, “Predicting many properties of a quantum system from very few measurements,” *Nat. Phys.* **16**, 1050–1057 (2020). doi: 10.1038/s41567-020-0932-7.
- [14] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, “The randomized measurement toolbox,” *Nat. Rev. Phys.* **5**, 9–24 (2023). doi: 10.1038/s42254-022-00535-2.
- [15] H.-Y. Huang, R. Kueng, and J. Preskill, “Efficient estimation of Pauli observables by derandomization,” *Phys. Rev. Lett.* **127**, 030503 (2021). doi: 10.1103/PhysRevLett.127.030503.
- [16] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, “Machine learning and the physical sciences,” *Rev. Mod. Phys.* **91**, 045002 (2019). doi: 10.1103/RevModPhys.91.045002.
- [17] V. Gebhart, R. Santagati, A. A. Gentile, E. M. Gauger, D. Craig, N. Ares, L. Banchi, F. Marquardt, L. Pezzè, and C. Bonato, “Learning quantum systems,” *Nat. Rev. Phys.* **5**, 141–156 (2023). doi: 10.1038/s42254-022-00552-1.
- [18] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, “Neural-network quantum state tomography,” *Nat. Phys.* **14**, 447–450 (2018). doi: 10.1038/s41567-018-0048-5.

- [19] J. Gao, L.-F. Qiao, Z.-Q. Jiao, Y.-C. Ma, C.-Q. Hu, R.-J. Ren, A.-L. Yang, H. Tang, M.-H. Yung, and X.-M. Jin, “Experimental machine learning of quantum states,” *Phys. Rev. Lett.* **120**, 240501 (2018). doi: 10.1103/PhysRevLett.120.240501.
- [20] J. Gray, L. Banchi, A. Bayat, and S. Bose, “Machine-learning-assisted many-body entanglement measurement,” *Phys. Rev. Lett.* **121**, 150503 (2018). doi: 10.1103/PhysRevLett.121.150503.
- [21] S. Lu, S. Huang, K. Li, J. Li, J. Chen, D. Lu, Z. Ji, Y. Shen, D. Zhou, and B. Zeng, “Separability-entanglement classifier via machine learning,” *Phys. Rev. A* **98**, 012315 (2018). doi: 10.1103/PhysRevA.98.012315.
- [22] D. Koutný, L. Ginés, M. Moczala-Dusanowska, S. Höfling, C. Schneider, A. Predojević, and M. Ježek, “Deep learning of quantum entanglement from incomplete measurements,” *Sci. Adv.* **9**, eadd7131 (2023). doi: 10.1126/sciadv.add7131.
- [23] Y.-J. Luo, J.-M. Liu, and C. Zhang, “Detecting genuine multipartite entanglement via machine learning,” *Phys. Rev. A* **108**, 052424 (2023). doi: 10.1103/PhysRevA.108.052424.
- [24] Y.-J. Luo, X. Leng, and C. Zhang, “Genuine multipartite entanglement verification with convolutional neural networks,” *Phys. Rev. A* **110**, 042412 (2024). doi: 10.1103/PhysRevA.110.042412.
- [25] R. Li, J. Du, Z. Qin, S. Zhang, C. Du, Y. Zhou, and Z. Xiao, “Entanglement structure detection via computer vision,” *Phys. Rev. A* **110**, 012448 (2024). doi: 10.1103/PhysRevA.110.012448.
- [26] R. Li, S. Zhang, Z. Qin, C. Du, Y. Zhou, and Z. Xiao, “Low-cost detection of high-dimensional multipartite entanglement structures,” *Phys. Rev. Applied* **23**, 044033 (2025). doi: 10.1103/PhysRevApplied.23.044033.
- [27] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World* (Springer, New York, 2005). doi: 10.1007/b106715.
- [28] A. N. Angelopoulos and S. Bates, “Conformal Prediction: A Gentle Introduction,” *Found. Trends Mach. Learn.* **16**, 494–591 (2023). doi: 10.1561/22000000101.
- [29] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, “Distribution-free, risk-controlling prediction sets,” *J. ACM* **68**, Article 43 (2021). doi: 10.1145/3478535.