Qubit Allocation in a Quantum Network using Stochastic Bandits

First Author, Second Author, and Third Author

Abstract—Quantum entanglement routing requires dynamic path selection and qubit allocation under noisy, uncertain, and adversarial conditions. Existing routing approaches often assume stationary link behavior, decouple selection from allocation, or rely on offline optimization—assumptions that break when link fidelities drift, and disruptions adapt to the learner. In this paper, we present a systematic evaluation of context-aware stochastic bandit algorithms for joint path selection and qubit allocation in quantum networks. Across 552 configurations spanning 13 algorithms, 5 threat scenarios (Stochastic, Markov, Adaptive, OnlineAdaptive, Baseline), 4 allocator strategies, and 2 capacity settings, pursuit–neural hybrids emerge as the most robust family, achieving 86–89% oracle-normalized efficiency on average and approaching 90% under favorable capacity–allocator regimes. They outperform non-contextual bandit baselines by 18–24 percentage points (pp) and sustain higher worst-case performance under strategic attacks than adversarial-first designs. We validate robustness through cross-testbed evaluation on three external quantum network simulators (15–100 nodes, 4–15 paths, diverse noise models), where scenario-aggregated efficiencies span 69.6–78.0% on Papers 2/7 and 42.5–44.1% on Paper 12, exposing clear scale- and physics-dependent performance limitations. Most critically, we uncover a capacity paradox: increasing replay capaciity from T to $2T$ (capacity notation defined in §IV-E) can induce a 22–31 pp efficiency collapse under Adaptive attacks, indicating that resource predictability, not bandwidth, limits robustness against intelligent adversaries. Allocator choice further induces 10–15 pp swings for identical algorithms, establishing algorithm–allocator co-design as a deployment requirement. We derive threat-responsive allocation guidelines that yield 86–90% efficiency versus 68–77% for non-contextual baselines, suggesting broader applicability to resource-constrained online decision systems facing intelligent adversaries.

Index Terms—Quantum networks, entanglement routing, multi-armed bandits, adversarial learning, qubit allocation, online optimization.

I. Introduction

Quantum entanglement distribution is a foundational primitive of the quantum Internet, enabling applications ranging from quantum key distribution (QKD) to distributed quantum computing [6, 25, 44, 45]. Yet, reliable end-to-end entanglement is difficult to sustain: quantum states are fragile, entanglement generation and swapping are probabilistic, and performance degrades rapidly under decoherence and interference. In repeater-based architectures [8], entanglement swapping introduces stochastic waiting-time effects that compound along multi-hop routes [41]. These properties create a sequential decision problem in which a routing agent must repeatedly select among candidate paths while learning from noisy outcomes, making multi-armed bandits (MABs) a natural abstraction for online path selection and adaptation [9, 28].

Affiliations and funding info here.

Quantum routing differs fundamentally from classical packet switching because the underlying resource is entanglement, not transferable data. Quantum states cannot be copied or amplified due to the no-cloning theorem [45], rendering standard store-and-forward buffering impossible. Instead, communication relies on establishing entanglement—a consumable resource constrained by memory coherence times and operation-induced fidelity loss—and then consuming that entanglement to support protocols such as teleportation [6]. Furthermore, entanglement swapping succeeds only with a non-unit probability [50], making link availability stochastic rather than deterministic. Consequently, routing must jointly optimize success probability and fidelity under uncertainty that evolves over time [21, 31, 43].

A. Gap in Prior Work

Existing quantum routing research often evaluates algorithms under incompatible assumptions, metrics, or threat models, complicating direct comparisons and obscuring deployment tradeoffs. Two streams are especially prominent:

1) Adversarial-first approaches (e.g., EXP3-family designs [4, 21]) assume worst-case link behavior and prioritize robustness under attack, often sacrificing stochastic efficiency.
2) Stochastic/contextual-first approaches (e.g., contextual and informed CMAB/iCMAB variants [11, 23, 43]) leverage predictive environmental structure for efficiency under natural noise, but offer limited evidence under coordinated adaptive adversaries.

This divide leaves three deployment-critical gaps insufficiently isolated in prior evaluations:

1) Context dependence is under-characterized. Prior work does not clearly identify which threat and noise regimes require topology/channel features for stable routing versus when non-contextual policies suffice.
2) Matched-threat comparisons are missing. Adversarial-first and stochastic/contextual-first methods are rarely evaluated under identical, controlled threat scenarios, limiting causal comparisons across design philosophies.
3) Deployment interactions are not disentangled. Allocator choice and replay/capacity semantics are typically treated as implementation details rather than first-class factors, obscuring interaction effects and counterintuitive performance shifts.

B. Our Approach and Evaluation Scope

To address these gaps, we have developed a robust routing evaluation framework that compares adversarial-first (EXP3-family), stochastic/contextual-first (CMAB/iCMAB-family), and hybrid pursuit–neural models (including EXPNeuralUCB [20]) under a unified threat taxonomy. The overall decision flow is summarized in Fig. 1. Our analysis is organized into curated evaluation corpora spanning:

- 16 models: 15 learned routing policies plus an Oracle upper bound
- 5 threat scenarios: Baseline, Stochastic, Markov, Adaptive, OnlineAdaptive
- 4 allocators: Default, Dynamic, Thompson, Random
- Replay/capacity settings: two capacity semantics (T_b vs. T, see §IV-E) evaluated across scales $s \in \{1.0, 1.5, 2.0\}$
- Frame horizons: 4K–12K frames, with 3-run and 5-run ensembles

In total, we report about 7,890 model–scenario–configuration evaluations across 835 unique scenario–allocator–capacity–horizon settings.

C. The Capacity Paradox

All efficiency metrics are Oracle-normalized—i.e., expressed as a percentage of the performance achieved by an ideal agent with perfect knowledge of link success probabilities and threat structure—enabling fair comparison across scenarios with different baseline difficulty.

A central empirical finding challenges conventional intuition: increasing replay capacity (measured via scales s and anchoring semantics T_b vs. T; see §IV-E for definitions) can produce threat-dependent swings, improving efficiency under structured (Markov) regimes while degrading performance under strategic (Adaptive) attacks. Across the evaluation corpora, capacity scaling interacts critically with threat type: several models gain under Markov structure as additional replay improves estimation stability, yet suffer pronounced collapses under Adaptive threats when allocation patterns become predictable targets. This capacity paradox suggests that resource predictability—not raw bandwidth—becomes the limiting factor in adversarial settings, motivating capacity as a dynamic control variable that should co-evolve with threat conditions rather than remain a fixed provisioning knob.

D. Key Contributions

- Unified, reproducible benchmarking across bandit families: We provide an apples-to-apples evaluation across EXP3-family adversarial baselines [4, 21], contextual/iCMAB-family methods [11, 23], and hybrid pursuit–neural models under a shared threat taxonomy.
- Cross-testbed validation at multiple scales: We validate our algorithms on three external quantum network testbeds from prior work [10, 13, 32], spanning 15–100 nodes, 4–15 routing paths, and diverse noise models (stochastic gate errors, fusion-based entanglement, context-driven dynamics). Using scenario-aggregated metrics across all five threats, model efficiencies span 69.6–78.0% on Papers 2/7 and 42.5–44.1% on Paper 12, indicating strong topology/physics dependence and clear mid-scale fusion-network limitations (§VII).
- Context and representation stabilize routing: Consistent with modern contextual and neural bandit theory [11, 48], context-aware neural agents substantially improve robustness and reduce variability relative to non-contextual baselines in realistic quantum routing conditions.
- Capacity paradox characterization: Increasing replay capacity yields gains under Markov regimes but can induce large collapses under Adaptive attacks, revealing that predictability is a primary vulnerability mechanism in adversarial quantum routing.
- Allocator–algorithm co-design and deployment rules: Allocator choice produces large performance shifts for identical policies, implying that allocator selection must be matched to the threat regime. We distill threat-responsive heuristics for choosing model families, allocators, and capacity scales in deployment.

Beyond the physics-level differences from classical networking, quantum path determination tightly couples routing to resource allocation and control: each attempted hop consumes scarce qubits and memory, and allocator and replay-capacity choices shape both what feedback the learner receives and what attack surface an adaptive disruptor can exploit. Accordingly, we study routing as a joint decision problem over path selection, qubit allocation, and learning policy under a controlled threat taxonomy spanning stochastic noise, structured dynamics, and adaptive disruption. By varying allocator strategy and capacity semantics as first-class experimental variables and validating across three external testbeds, we derive deployment rules that map threat conditions to model–allocator–capacity choices rather than prescribing a single fixed routing policy.

II. Related Work

A. Literature Selection Methodology

We situate multi-armed bandits (MABs) as a family of uncertainty-aware sequential decision rules and use quantum entanglement routing as a stress test where stochastic noise, structured disruption, and resource constraints jointly shape performance. Our scope spans: (i) finite-time regret analyses in stochastic and adversarial regimes, (ii) contextual and neural representations for structured environments, (iii) hybrid constructions that combine mechanisms across regimes, and (iv) predictive (informed) bandits that incorporate explicit forecasts or learned dynamics.

1) Search Strategy and Time Span (2002–2025): We queried arXiv, IEEE Xplore, and the ACM Digital Library over the 2002–2025 window using combinations of: multi-armed bandits, contextual bandits, adversarial bandits, neural bandits, exploration–exploitation, predictive bandits, online forecasting, informed contextual bandits, quantum routing. We then performed backward/forward snowballing from canonical anchors (e.g., UCB/EXP3 foundations; LinUCB and neural contextual bandits; recent bandit-based quantum-routing studies), prioritizing work that either: (a) introduced a reusable mechanism (confidence bounds, posterior sampling, adversarial randomization, pursuit-style updates, forecasting/world models), or (b) demonstrated cross-domain transfer under materially different constraints (e.g., routing, communications, finance, healthcare).

The chosen time span captures the modern finite-time theory for stochastic and adversarial learning, the rise of contextual formulations for structured decision-making, and recent neural, hybrid, and predictive variants designed for complex, partially observed environments.

2) Inclusion and Exclusion Criteria:

Included:

1) Canonical stochastic and adversarial MAB algorithms with regret guarantees [3, 5, 40].
2) Contextual neural bandits that scale action selection to structured/high-dimensional contexts [30, 46, 48].
3) Predictive/informed contextual bandits that integrate forecasting or learned dynamics [7, 24].
4) Hybrid methods that combine mechanisms across regimes [39].
5) Cross-domain applications where the same mechanism is instantiated under different reward and constraint models.

Excluded:

1) Offline optimization/control without online learning under bandit feedback.
2) Single-domain demonstrations without reusable algorithmic insight.
3) Pure tuning studies without methodological novelty, clearly stated assumptions, or reproducibility artifacts.

B. Foundational Bandits and Regret Regimes

Foundational results formalize the exploration–exploitation trade-off and provide regret guarantees that define the efficiency–robustness envelope. In stochastic i.i.d. settings, UCB-style optimism and Thompson-style posterior sampling achieve logarithmic-in-horizon regret under standard gap conditions [3, 40]. In adversarial settings, EXP3 attains sublinear regret without stochastic assumptions, trading some benign-regime efficiency for worst-case protection [5]. These regimes motivate why quantum routing evaluations should explicitly separate natural noise from coordinated disruption: the learning objective and the appropriate safety guarantees depend on the feedback model.

C. Contextual and Neural Bandits

Contextual bandits condition decisions on observable state, enabling structured decision-making when arms are not exchangeable. LinUCB models rewards as linear in context and selects actions via a confidence bonus [30]. NeuralUCB and NeuralTS generalize this principle by learning representations with deep networks while retaining uncertainty-aware action selection through confidence-style bounds or sampling in representation space [46, 48]. The shared abstraction is mechanism-level: learn a value predictor, maintain an uncertainty estimate over that predictor, and act optimistically or probabilistically. In quantum routing, the natural context includes topology/hop structure, link-quality indicators, and any measurable signals of load, memory, or temporal drift.

D. Adversarial and Hybrid Robustness

Adversarial bandits prioritize worst-case guarantees (e.g., EXP3-style randomization, which can be essential under nonstationarity or strategic manipulation [5]). Hybrid designs aim to combine robust exploration with structured exploitation—for example, layering pursuit-style updates over context-conditioned value estimation or embedding adversarial weighting within learned reward models [39]. Within quantum-routing studies, adversarial-first formulations are often motivated by jamming or targeted disruption; however, comparisons across families are frequently confounded by mismatched assumptions about allocation, memory/replay semantics, and evaluation taxonomies. This motivates treating allocation policies and replay parameterization as first-class experimental factors when assessing robustness under mixed threats.

E. Predictive and Informed Bandits

Predictive (informed) contextual bandits augment the decision rule with a forecasting or world-model component so that learning can anticipate drift rather than purely react to it. iCMAB exemplifies this direction by coupling contextual decision rules with an explicit predictive model of future context dynamics [24]. In our setting, we instantiate an informed variant by warming up predictive context with a classical time-series forecaster (ARIMA) [7]. The core idea is again mechanism-level: forecasting can augment the context signal, but it does not replace the need for robust exploration and allocation policies under strategic threats.

F. Quantum Network Routing with Bandits

Recent quantum-network work applies bandits (and related online learning) to path selection under stochastic decoherence and, in some cases, structured disruption [19, 31, 32, 43, 44]. Wang et al. [43] focus on learning high-quality paths under stochastic dynamics,

while Li et al. [31] propose multipath inter-domain routing protocols for quantum networks with online path selection; Liu et al. [32] similarly emphasizes online benchmarking signals to support routing-policy adaptation. Wang et al. [42] formulate an adaptive, user-centric entanglement routing problem with long-term budget constraints and propose an online control algorithm for per-slot routing and qubit allocation. In contrast, our contribution is to introduce and benchmark multiple allocator/decision-rule algorithms (including hybrid pursuit–neural and informed iCMAB variants) under a shared threat taxonomy, while treating allocator policy and replay/capacity semantics as explicit experimental factors. We do not propose a new quantum-network routing protocol or a new budgeted-control formulation with analytical guarantees; rather, we provide a controlled robustness characterization that isolates which algorithm–allocator–capacity combinations remain stable when disruption is structured or adaptive. Huang et al. [19] propose EXPNeuralUCB, a group neural bandit that combines EXP3-style adversarial exploration with NeuralUCB-style nonlinear reward modeling for joint path selection and qubit allocation. We advance this line by introducing pursuit–neural hybrids (e.g., CPursuitNeuralUCB, iCPursuitNeuralUCB) and show that they achieve higher scenario-aggregated efficiency with stronger stability-floor behavior than EXPNeuralUCB in our evaluation suites. Further, while Huang et al. treat allocation as a fixed component, our framework explicitly varies allocator strategy and replay capacity, revealing that these factors can be as critical to robustness as the learning rule itself.

a) Closest-work contrast (LinkSelFiE).: Liu et al. [33] propose LinkSelFiE, which targets the link-level problem of selecting a high-fidelity entanglement link and estimating its fidelity when link qualities are unknown a priori. They cast link selection as a best-arm identification task and couple it with a benchmarking-driven estimation procedure to reduce quantum resource consumption while still identifying high-quality links with high confidence. In contrast, our study targets the end-to-end routing problem: joint path selection and qubit allocation over time under five threat regimes, quantified through a controlled cross-product evaluation across algorithms, allocators, and replay-capacity semantics. Our framework can incorporate link-level fidelity signals (including LinkSelFiE-style estimation outputs) into the routing reward model, but our primary contribution is to characterize robustness and deployment trade-offs at the routing layer under structured and adaptive disruption.

Beyond bandit-style path selection, learning-based route selection under noisy quantum-network conditions has also been explored [10], and RL-based adaptive routing has been proposed via deep Q-networks [22]. Our work differentiates by introducing pursuit–neural allocator algorithms and stress-testing them under structured and adaptive threats in addition to stochastic noise. Complementary non-learning routing designs emphasize structural decomposition (e.g., QuARC adaptive clustering [13] and hierarchical routing for scalability [12]) or repeater/efficiency constraints [27], while cost-vector approaches optimize multi-path routing decisions through explicit objective formulations [29]. In contrast, our contribution is a controlled evaluation methodology that isolates how decision-rule families interact with allocation policies, replay semantics, and capacity across a shared threat taxonomy, enabling direct attribution of robustness to the algorithm–allocator–capacity triad rather than to a single routing primitive. Across studies, experimental assumptions differ materially—especially in how qubits are allocated across paths, how memory/replay is parameterized relative to the horizon, and how threat processes are modeled—which complicates direct algorithm-to-algorithm comparison. Our benchmark addresses this by evaluating multiple algorithm classes under a shared threat taxonomy and by treating allocator policy and replay configuration as explicit experimental factors rather than fixed background choices, enabling direct attribution of robustness to the algorithm–allocator–capacity triad.

G. Toward a Modular, Universal Bandit Stack

Across domains, the recurring meta-problem is stable: choose actions under uncertainty using an uncertainty-aware value estimate. What varies is the adapter layer: how context is defined, how rewards are observed, and what constraints must be respected (e.g., qubit budgets, routing thresholds, replay semantics). We operationalize this view as a modular stack:

(i) an allocator/decision rule,
(ii) an optional forecasting layer that augments context,
(iii) a domain adapter mapping quantum-routing signals to the shared allocator interface.

General benchmarking and utility frameworks for quantum networks motivate this evaluation-first framing [14, 26]; our work complements them by making learning-specific factors (threat models, allocators, replay/capacity semantics) explicit experimental variables. Best-of-both-worlds bandit theory motivates this mixed-regime view: algorithms such as Tsallis-INF achieve strong guarantees in both stochastic and adversarial settings without knowing the regime a priori [49]. In our setting, we operationalize this idea empirically by stress-testing the same allocator stack across stochastic, structured, and adaptive threats.

Within this stack, our study provides a controlled comparison of classical, contextual/neural, adversarial, and informed variants under stochastic and adaptive disruption, clarifying when robustness is determined not only by the decision rule, but also by allocator choice and replay configuration.

III. Background

A. Quantum Networks and Entanglement Routing

Quantum networks rely on entanglement distribution across repeaters and end-nodes to enable long-distance

quantum communication, distributed quantum computing, and sensing applications [25, 44]. Unlike classical networks where packets can be buffered and forwarded with near-deterministic behavior, quantum networking operates under fundamental constraints: quantum states are fragile, entanglement generation and swapping are probabilistic, and the quality of distributed entanglement (e.g., fidelity) degrades under decoherence and imperfect operations [8, 15].

Quantum teleportation [6] enables qubit transmission by consuming shared entanglement between nodes, while entanglement swapping [50] at repeaters extends entanglement over long distances by stitching together shorter entangled segments. Because routing decisions must be made repeatedly from noisy outcomes and time-varying link conditions, quantum path selection naturally becomes a sequential decision problem under uncertainty, for which multi-armed bandits (MABs) provide a principled abstraction [9, 28].

Traditional quantum routing often assumes either (i) complete topology and stable link characterization enabling offline optimization, or (ii) fixed, heuristically tuned allocation rules. These assumptions can break in realistic deployments where link conditions must be learned online and routing must adapt to demand variability and disruptive events, including strategic interference [21, 31, 43].

B. The Multi-Armed Bandit Abstraction

A multi-armed bandit (MAB) formalizes online routing as follows: at each time step t, an agent selects one of K candidate actions (e.g., paths or allocation decisions), observes a reward signal (e.g., entanglement success/failure or efficiency proxy), and aims to minimize regret relative to an oracle policy [28]. The central challenge is the exploration–exploitation trade-off: learning which actions are reliable while maintaining high routing performance [9].

Several bandit variants align with different quantum-network assumptions and threat models:

- Classical (stochastic) bandits assume stationary reward distributions (e.g., UCB-style methods) [2].
- Contextual bandits incorporate observable side information (e.g., topology features or load indicators) to improve decisions when context is predictive [11].
- Neural contextual bandits use function approximation to model nonlinear reward while preserving principled exploration via uncertainty-aware decision rules [48].
- Adversarial bandits guard against worst-case or non-stochastic reward sequences (e.g., EXP3 algorithm) [4].
- Predictive/informed bandits augment decisions with forecasts of future conditions [23].

This taxonomy provides a natural lens for quantum routing: stochastic noise motivates contextual/neural modeling, while strategic disruption motivates adversarial robustness [21].

C. Allocation and Capacity Semantics

In addition to choosing routes, practical quantum routing must manage resource allocation decisions (e.g., how many attempts or qubits to allocate across competing paths within a decision epoch). Allocation policies can materially change performance even for the same underlying bandit learner, because they shape both the information collected and the predictability of routing behavior under disruption.

Many learning-based routing implementations also impose finite-memory or replay semantics (e.g., bounded histories, windowed updates, or capped experience buffers) that affect stability under nonstationarity and vulnerability under strategic adaptation. These design choices motivate evaluating routing policies jointly with allocator strategy and capacity semantics, rather than treating them as independent knobs.

D. Problem Scope

Motivated by these considerations, we study how modeling choices (e.g., contextual vs. adversarial vs. predictive), allocator strategies, and capacity semantics jointly affect routing robustness under diverse threat regimes. Specifically, we evaluate 16 models (15 learned policies + Oracle) across 5 threat scenarios, 4 allocators, capacity scales $S \in \{1.0, 1.5, 2.0\}$, and 3–5 run ensembles to isolate which design choices matter most and to support actionable deployment guidance.

IV. System Model

We model quantum entanglement routing as a sequential decision problem where an agent must jointly optimize (1) path selection among candidate routes and (2) qubit allocation across path segments, under uncertain link fidelities and adversarial interference. This section formalizes the network topology, reward structure, threat taxonomy, qubit allocation policies, and the underlying MAB formulation.

A. Network Topology and Path Structure

4-node diamond topology: We study a canonical 4-node quantum network connecting source S (Alice) to destination D (David) via two intermediate repeaters, Bob and Charlie. This diamond topology yields 4 candidate paths $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$ with varying hop counts:

- 2-hop paths ($P_1 = S{\to}B{\to}D$, $P_2 = S{\to}C{\to}D$): one repeater, two entanglement links
- 3-hop paths ($P_3 = S{\to}B{\to}C{\to}D$, $P_4 = S{\to}C{\to}B{\to}D$): two repeaters, three links

Each path P_r has h_r hops ($h_1 = h_2 = 2$, $h_3 = h_4 = 3$), where a hop denotes a single entanglement link between adjacent nodes (e.g., source-to-repeater or repeater-to-destination) requiring one entanglement generation operation. Shorter paths reduce swapping overhead but may traverse noisier channels; longer paths enable route diversity but accumulate fidelity loss.

Qubit budget and allocator policies: The network operates under a fixed total budget of 35 qubits distributed across paths. We evaluate four allocator strategies that dynamically or statically assign qubits:

1) Fixed (T_1=8, T_2=10, T_3=8, T_4=9): static baseline
2) ThompsonSampling: Bayesian posterior sampling over path utilities
3) DynamicUCB: upper-confidence-bound-driven capacity redistribution
4) Random: uniform random assignment (control baseline)

For each path P_r with session budget T_r, a feasible allocation $\mathbf{x} = (x_1, \ldots, x_{h_r})$ satisfying $\sum_{\ell=1}^{h_r} x_\ell = T_r$ defines the context space $\mathcal{X}_r$. This combinatorial space scales quadratically for 3-hop paths, motivating contextual neural approximation.

B. Reward Model and Link-Level Fidelity

Probabilistic entanglement generation: Each path P_r contains h_r links indexed by $\ell \in \{1, \ldots, h_r\}$. For each link ℓ, let $p_e^{(\ell)} \in [10^{-4}, 2{\times}10^{-4}]$ denote its per-attempt entanglement success probability (representative of realistic SNSPD-based quantum memory systems), and let x_ℓ denote the number of qubits allocated to that link. Over a decision step (frame), allocating x_ℓ qubits yields link-level success probability

$$p_\ell(x_\ell) = 1 - (1 - p_e^{(\ell)})^{x_\ell}.$$

Path-level success as multiplicative fidelity: End-to-end entanglement succeeds if all h_r links succeed (entanglement swapping requires coherent intermediate states). Path success probability is multiplicative:

$$h_r(\mathbf{x}) = \prod_{\ell=1}^{h_r} p_\ell(x_\ell).$$

This multiplicative decay captures the fundamental quantum constraint: fidelity degrades exponentially with hop count unless mitigated by purification (out of scope for routing).

Bernoulli rewards under adversarial availability: At frame t, selecting path P_r (where $r \in \{1, 2, 3, 4\}$) with allocation $\mathbf{x} = (x_1, \ldots, x_{h_r})$ yields binary reward

$$Y_t(r, \mathbf{x}) \sim \text{Bernoulli}\big(h_r(\mathbf{x}) \cdot A_t(r)\big),$$

where $Y_t(r, \mathbf{x}) \in \{0, 1\}$ denotes the observed outcome (1 = success, 0 = failure) at frame t, and $A_t(r) \in \{0, 1\}$ is the availability indicator: $A_t(r) = 1$ if path P_r is active (no attack), $A_t(r) = 0$ if disrupted. When a path is disrupted at time t, we set its indicator to zero (i.e., $A_t(r) = 0$ for disrupted paths and $A_t(r') = 1$ otherwise), so disruption acts as a path-specific availability gate on otherwise stochastic link success. This formulation cleanly separates stochastic decoherence (encoded in h_r) from strategic interference (encoded in A_t).

C. Adversarial Threat Taxonomy

We study routing robustness under five escalating threat regimes spanning benign stochasticity to intelligent reactive attacks. Each scenario modulates the availability vector $\mathbf{A}_t = (A_t(1), \ldots, A_t(4))$ according to distinct disruption semantics.

Baseline (No Disruption). In the baseline regime, all routes remain available at all times: $A_t(r) = 1$ for all r, t. This setting isolates pure stochastic decoherence and serves as the benign-condition upper bound (Oracle-aligned) for comparisons across all disrupted regimes.

Stochastic (6.25% i.i.d. failures). Under stochastic disruption, each route is independently available according to $A_t(r) \sim \text{Bernoulli}(0.9375)$. This captures benign environmental noise without temporal structure or memory.

Markov (25% structured disruption). In the Markov regime, availability is governed by a 4-state Markov chain whose states modulate path-failure probabilities. This setting captures bursty, correlated outages, with an average disruption rate of approximately 25%.

Adaptive (Reactive targeting over sliding window). In the adaptive regime, the adversary observes path selection over a sliding window of $w = 50$ frames and targets the most-used path with 100% probability. This mechanism exploits predictable allocation patterns and is especially punitive for Fixed allocators and larger replay scaling (s).

OnlineAdaptive (Real-time policy adaptation). In the OnlineAdaptive regime, the adversary maintains exponentially weighted path usage with decay $\gamma = 0.97$ and applies softmax targeting with temperature τ:

$$\Pr(\text{attack path } r \mid t) = \frac{\exp(\text{usage}_r(t)/\tau)}{\sum_{r'} \exp(\text{usage}_{r'}(t)/\tau)}.$$

This setting mimics intelligent adversaries that learn and adapt in real time, representing the hardest realistic threat model in our taxonomy.

D. Multi-Armed Bandit Formulation

Hierarchical group-arm structure: We cast quantum routing as a 2-level MAB problem:

- Level 1 (Group selection): Choose path $r_t \in \{1, 2, 3, 4\}$ at frame t via policy π_g.
- Level 2 (Arm selection): Choose allocation $\mathbf{x}_t \in \mathcal{X}_{r_t}$ via policy π_a.

Each path P_r defines a group $\mathcal{G}_r$ containing $|\mathcal{X}_r|$ allocation arms. For 3-hop paths with budget T_r, $|\mathcal{X}_r| \approx \binom{T_r + h_r - 1}{h_r - 1}$ (stars-and-bars), scaling quadratically.

Objective and regret: Let $\mu_t(r, \mathbf{x}) := h_r(\mathbf{x}) A_t(r)$ denote the expected reward at time t. Maximize cumulative expected reward over horizon F_b:

$$\max_{\pi_g, \pi_a} \mathbb{E}\left[\sum_{t=1}^{F_b} \mu_t(r_t, \mathbf{x}_t)\right],$$

subject to unknown dynamics $h_r(\cdot)$ and adversarial interference $A_t(\cdot)$. Regret measures suboptimality relative to an Oracle with perfect foresight:

$$\text{Regret}(T) = \sum_{t=1}^{T} \left(\mu_t^* - \mu_t(r_t, \mathbf{x}_t)\right),$$

where $\mu_t^* = \max_r \mu_t(r, \mathbf{x}_r^*)$ is the best achievable reward at t.

E. Capacity Semantics and Replay Scaling

$$T = \begin{cases} s \cdot F_b & \text{(base-frames replay, } T_b\text{-type)} \\ s \cdot F_c & \text{(current-frames replay, } T\text{-type)} \end{cases}$$

To eliminate ambiguity in memory scaling, we distinguish two replay-capacity types. Let:

- F_b: base per-run horizon (frames)—the standard episode length before scaling (e.g., 4K, 6K, or 8K frames)
- F_c: current per-run horizon (frames)—actual frames after any horizon scaling
- $s \in \{1, 1.5, 2\}$: capacity scale factor (default $s = 2$)

Both are always scaled (even when $s = 1$). This separation is critical for diagnosing the capacity paradox (§VI): increasing nominal replay from $s = 1 \to s = 2$ can degrade efficiency by 22–30 pp under Adaptive attacks, as excess predictable capacity becomes exploitable.

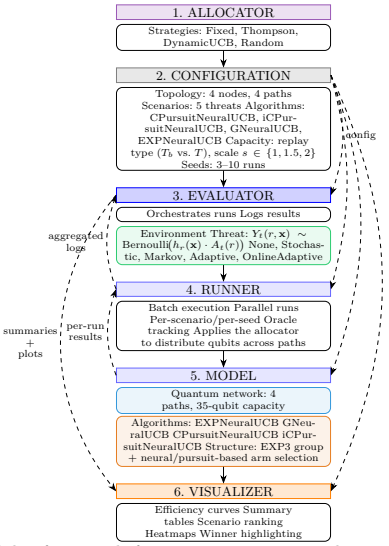


Fig. 1. Modular framework for quantum routing evaluation. Allocator selects the qubit distribution strategy under test. Configuration is shared by all layers, defining topology, threats, algorithms, capacity, and seeds. The Evaluator builds the environment and orchestrates runs; the Runner executes per-scenario/per-seed batches with the chosen allocator; the Model implements the quantum routing stack (network + bandit); and the Visualizer aggregates configuration-aware results to compare winners across scenarios.

F. Algorithmic Framework

Our evaluation infrastructure consists of six layers (Figure 1):

1) ALLOCATOR: Top-level policy (e.g., DynamicUCB)
2) CONFIGURATION: Shared experimental setup (topology, scenarios, capacity, seeds)
3) EVALUATOR: Orchestrates runs, logs, threat models
4) RUNNER: Batch execution engine with Oracle tracking
5) MODEL: Quantum routing stack (algorithms + neural)
6) VISUALIZER: Aggregates results into efficiency curves, tables, and winner summaries

This architecture enables systematic ablation across allocators, algorithms, and threat regimes under consistent capacity semantics, ensuring reproducible comparisons [21, 43].

V. Study Design

We evaluate routing architectures and deployment configurations under stochastic and adversarial constraints using a modular, four-phase framework. We isolate the contributions of context awareness, predictive intelligence, and capacity management, while ablation studies probe algorithm–allocator–capacity interactions across five threat profiles.

A. Research Questions

Our study addresses three core questions about stochastic decoherence impact, adversarial robustness, and deployment tradeoffs in algorithm–allocator–capacity selection. [Devroop: Wont it make more sense to just relay the questions here and answer them in the results and discussion section?]

RQ1: Does stochastic decoherence degrade routing efficiency enough to make classical MABs insufficient?

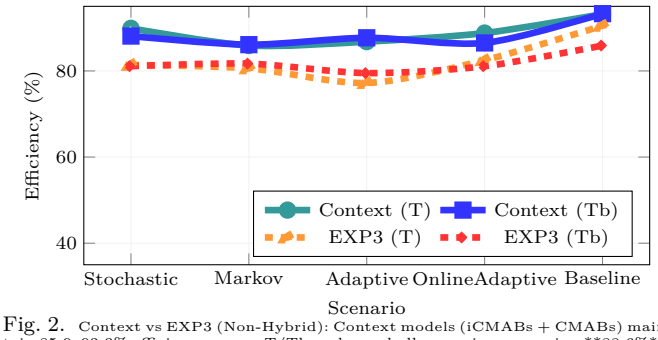


Fig. 2. Context vs EXP3 (Non-Hybrid): Context models (iCMABs + CMABs) maintain 85.9–93.3% efficiency across T/Tb scales and all scenarios, averaging **88.6%**. EXP3-based models range 77.2–90.5% with lower consistency (82.1% avg), showing 6.5–8.2 pp gaps under capacity-constrained (Tb) conditions. Context-awareness provides superior scale-robust quantum routing. Values aggregated from evaluation corpora across scales 1.0–2.0.

Yes—in our benchmark, representative classical baselines drop well below practical targets under i.i.d. disruption, while context-aware routing remains deployment-viable. In the Stochastic regime (6.25% i.i.d. disruption), pursuit–neural routing maintains 87–93% Oracle-normalized efficiency under the validated suites (Hybrid, T_b, s=2, Default allocator), whereas classical baselines such as CThompsonSampling and CEXP4 fall to 62–70% (CMAB validated suites), yielding an average gap of approximately 24 pp (see §VI-A and §VI).

Supporting questions:

- RQ1a: What performance floor do classical (EXP3-based) baselines reach under pure stochastic disruption?
  In the CMAB validated suites under Stochastic, CThompsonSampling and CEXP4 remain in the 62–70% band, which is below an 85% deployment target.

- RQ1b: Do topology/channel-aware models outperform classical baselines under the same disruption?
  Yes. Pursuit–neural routing reaches 87–93% under matched conditions, exceeding the classical floor by a wide margin in the validated suites.
- RQ1c: Do viable contextual/neural models remain above an 85% target across horizons and suites?
  Yes. In the matched testbed slice (T_b, s=2, non-random allocators, 3- and 5-run suites), pursuit–neural models remain $\geq$ 85% under Stochastic across the primary horizons used in validated suites.

RQ2: How do classical baselines and context-aware models perform across increasingly sophisticated threats?

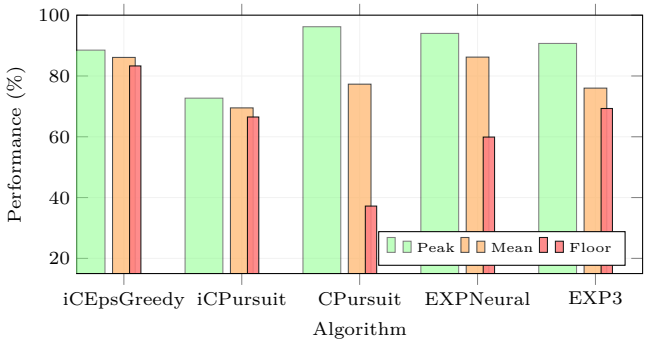


Fig. 3. Peak/mean/floor at the default 2T–2T$_b$ budgets with 3-run horizons. Context models retain much higher floors; CPursuit achieves 96.2% peak but drops to 37.2% in worst-case; EXPNeuralUCB's floor drops to 59.9%.

Context-aware models dominate across threat escalation. Across Markov, Adaptive, and OnlineAdaptive regimes, pursuit-based contextual models maintain averages around 89% with robustness floors in the 65–73% range, while classical baselines remain in the 66–77% band and adversarial-first defenses collapse to floors near 52–54% under the most aggressive threats (see Section VI-B).[5, 19]

Supporting questions:

- RQ2a: Do context-aware algorithms outperform classical and adversarial-first baselines under natural stochastic noise, and do they maintain this advantage under structured and adaptive threat regimes?
  Yes. Under Stochastic decoherence, contextual and neural–contextual models retain the 15–25 pp gain observed in RQ1 and remain above the 85% target. Under Markov/Adaptive/OnlineAdaptive, pursuit-based models keep high averages ($\simeq$89%) and floors of 65–73%, while adversarial-first methods fall to floors near 52–54%.[5]
- RQ2b: Which algorithm family degrades most gracefully as threat sophistication escalates from stochastic to reactive adversaries, and which consistently defines the efficiency frontier across all threat scenarios?
  Pursuit models. Pursuit-based models show the smoothest degradation, and iCPursuitNeuralUCB most consistently sits on/near the top efficiency envelope while maintaining strong threat-wide robustness.

Context-aware pursuit–neural routing remains the strongest on average across escalation, while classical baselines stay far below the deployment threshold. Under Markov, Adaptive, and OnlineAdaptive threats, pursuit–neural routing achieves mean efficiencies of approximately 90.8%, 93.5%, and 92.9% respectively in validated suites (Hybrid, T_b, s=2, Default/Thompson allocators), with worst-case floors in the 64–71% range driven primarily by OnlineAdaptive/Markov slices (see §VI-B). By comparison, representative classical baselines remain near 64–70% across the same threat regimes (CMAB experiments).

RQ3: What combinations of algorithm, allocator, and capacity semantics maximize efficiency and stability across deployment scenarios?

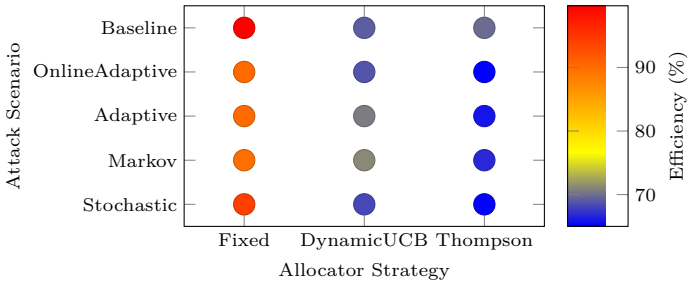


Fig. 4. Efficiency heatmap for CPursuitNeuralUCB across core allocator strategies and threat scenarios. Fixed allocation (Default) values from the Hybrid evaluation corpus; Dynamic/Thompson values maintained from analysis reports. Fixed allocation consistently achieves highest efficiency (89.6–99.7% across attacks, 94–100% baseline). Random allocator (37.1–38.7% efficiency) excluded from core findings; reported separately in supplementary analysis.

Deployment tradeoffs are real, but robust static configurations do exist in our expanded benchmark. In our Hybrid validated suites (excluding Random allocator), we observe many fixed (model, allocator, cap_type, scale, frames, runs) configurations whose minimum efficiency remains $\geq$ 85% across all five threats (e.g., 236 validated configurations meet this criterion), and the best static configurations sustain threat-wide minima above 93%. At the same time, capacity semantics (T_b vs. T), replay scale s, and allocator choice introduce large interaction effects—including up to $\approx$53 pp swings between T_b and T in some matched settings—so selecting for peak efficiency or minimum variance remains threat- and objective-dependent (see §VI-C4–§VI-E).

Supporting questions: .

- RQ3a: Does predictive context improve stability?
  Yes. In validated suites, predictive variants can improve mean efficiency and/or reduce variability under non-reactive noise depending on the matched configuration; additional 10-run observations further support this trend.
- RQ3b: Do larger replay scales $T_b \to T$ universally help?
  No. Capacity is not monotone: validated suites show settings where increasing s reduces performance under Adaptive targeting (e.g., $\approx$26 pp drops in matched slices), and switching $T_b \leftrightarrow T$ can produce >50 pp swings in either direction depending on threat structure.
- RQ3c: Is there a universal allocator strategy?
  No. Allocator choice interacts with threat and semantics: while allocators often cluster within single-digit differences, worst-case matched slices exhibit allocator spans that can exceed 10 pp and, in extreme cases, >30 pp.
- RQ3d: Can we map threat characteristics to stable deployment rules?
  Yes. Because threat-wide $\geq$ 85% static configurations exist, deployment rules can be expressed as (i) a small set of robust defaults (static, threat-wide) plus (ii) optional threat-tuned switches when optimizing for peak efficiency/variance.

B. Experimental Design

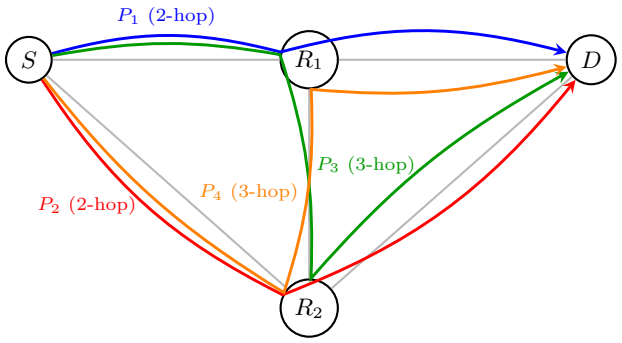
This section specifies the experimental design used to evaluate adaptive quantum entanglement routing under stochastic and adversarial interference, and ties each configuration axis to the research questions. Table I consolidates design dimensions, tested options, and the corresponding RQ coverage.

Network configuration: We use a 4-node quantum network with four alternative paths connecting source S to destination D via repeater nodes (Figure 5). Paths P_1

and P_4 have two hops, while P_2 and P_3 have three hops, matching common small-scale quantum-network architectures and demonstrations [35, 44]. This topology provides sufficient action-space complexity for bandit learning [16, 18, 19] while keeping exhaustive cross-product sweeps tractable (hundreds of algorithm–scenario–allocator configurations; see Table I).

Total physical network capacity is fixed at 35 qubits across all experiments, representing resource-constrained early-stage deployments [38]. This induces non-trivial exploration–exploitation tradeoffs: algorithms cannot over-provision all paths. Per-path allocations are determined by the allocator (Table IV), enabling controlled comparison of static versus adaptive resource management.

Fig. 5. 4-node quantum network topology with four alternative entanglement paths. Paths P_1 and P_2 use 2 hops; P_3 and P_4 use 3 hops. Total capacity is 35 qubits, allocated per path by strategy (Table IV); example splits: Fixed (8,10,8,9), Thompson (9,9,9,8).



Time horizons. We evaluate 3 horizons—4K,6K,8K frames—to capture short-, mid-, and long-episode learning dynamics. Unless otherwise noted, primary results use 6K, with 4K vs. 8K comparisons in RQ3a to assess sample efficiency.

TABLE I
Experimental configuration summary: design dimensions, options, and linkage to research questions (Phase totals sum to 552 configuration units; Phase 2 is executed for both CMAB and iCMAB families at 180 conditions per family). Each configuration is evaluated with up to 10 independent runs.

Configuration Dimension	Options Tested	RQ(s)
Network topology	4-node, 4-path (2+3-hop)	All
Time horizons	4K, 6K, 8K frames	$RQ3_a$
Qubit capacity	35 total (fixed) qubits	All
Allocators	Fixed, Thompson, DynamicUCB, Random	$RQ3_c$
Replay capacity	$T_b = sF_b$, $T = sF_c$	$RQ3_b$
Threat models	None, Stochastic, Markov, Adaptive, OnlineAdaptive	$RQ1$ & $RQ2$
Forecasting	None, ARIMA ($n = 50$), ARIMA ($n = 100$)	$RQ3_a$
Algorithm families	Classical (4), Predictive (1) Adversarial (3), Context (6)	$RQ1$ & $RQ2$
Evaluation phases		
Ph 1 (MAB baseline)	12 conditions	RQ_1
Ph 2 (CMAB / iCMAB)	180 conditions / family	$RQ2_{(a-b)}$
Ph 3 (Dynamic allocation)	240 conditions	$RQ3_c$
Ph 4 (Capacity ablation)	120 conditions	$RQ3_b$

Replay configurations (capacity scaling). We define the base per-run horizon as $F_b \in \{4\text{K}, 6\text{K}, 8\text{K}\}$ frames and execute $S \in \{3, 5, 8, 10\}$ independent runs per configuration (total of $S \cdot F_b$ frames per setting). Some configurations apply a frame scaling to yield a current horizon F_c (the per-run frame budget for that configuration). Capacity is always expressed via a scale factor $s \in \{1, 1.5, 2\}$ (default $s = 2$) in two equivalent views:

$$T_b = s \cdot F_b, \qquad T = s \cdot F_c.$$

We sweep s to analyze the Capacity Paradox (RQ3b)—whether larger replay memory can degrade performance under adaptive threats. In addition to the default $s = 2$ setting, we include intermediate ($s = 1, 1.5$) sensitivity checks and an extended doubled-current stress test ($2T = 2(sF_c) = 4F_c$ when $s = 2$), mirroring fixed-capacity conventions used in prior replay-buffer designs while avoiding conditional semantics.[34, 37]

Resource separation. Scaling applies to replay memory; the 35-qubit physical network capacity remains invariant. This decoupling isolates learning-system constraints from quantum hardware limits.

Adversarial threat taxonomy. We evaluate five scenarios spanning the efficiency–security spectrum: natural stochastic decoherence, Markovian structure, and three grades of adaptive adversarial attacks (RQ2). This taxonomy enables controlled comparisons under matched interference regimes. Hyperparameter settings are summarized in Table II.[18, 19]

TABLE II
Hyperparameter settings and literature justifications.

Parameter	Algorithm	Value	Ref.
Neural LR	EXPNeuralUCB	0.01	[18]
	GNeuralUCB	0.2	[47]
Exp. Weight	EXPNeuralUCB	0.05	[18]
	EXPUCB	0.005	Tuned
Pursuit LR	Pursuit family	0.2	[39]
UCB Explore	DynamicAlloc	2.0	[2]
ARIMA (n)	iCPursuit	50, 100	[7]
Replay (T_b, T)	All	$T_b = sF_b$, $T = sF_c$	[34, 37]

1) Scenario Specifications:

- Baseline (RQ1): Ideal 0% failure rate. Sets the efficiency ceiling, isolating exploration costs from threat mitigation.
- Stochastic (RQ1, RQ2a): 6.25% i.i.d. failure rate. Tests if specialized defenses are required under natural noise.[19]
- Markov (RQ2a–b): 25% attack rate using a 4-state transition process. Models state-dependent disruption.[18]
- Adaptive (RQ2a–b): 25% attack rate targeting high-usage paths over a sliding window ($w = 50$). Models reactive targeting based on medium-term behavior.[18, 19]
- OnlineAdaptive (RQ2a–b): 25% attack rate with exponential memory decay ($\gamma = 0.97$) and softmax targeting. Models continuous policy adaptation at the highest threat level.[18, 19]

2) Algorithm portfolio (RQ2): We evaluate 14 algorithms spanning three generations of MAB development plus an Oracle baseline. Phase progression (Table III) covers classical exploration (Phase 1), adversarial defenses (Phase 1–2), contextual/neural methods (Phase 2), and pursuit-based predictive models (Phase 2–3), enabling systematic comparison across architectural paradigms.

3) Key algorithm features:

- CPursuitNeuralUCB: Adds topology features and channel quality metrics to a contextual pursuit framework ($\gamma = 0.2$).
- iCPursuitNeuralUCB: Augments CPursuit with ARIMA $(1, 0, 1)$ forecasting (warmup $n \in \{50, 100\}$), providing anticipatory context for proactive path selection [7].

TABLE III
Algorithm portfolio by evaluation phase. Phases progress from classical → adversarial → contextual → predictive.

Phase	Category	Algorithms
Phase 1	Classical MAB	LinUCB [30], LinTS [1], UCB1 [2], Thompson Sampling [40]
Phase 1–2	Adversarial	EXP3 [5], EXPUCB [18], EXPNeuralUCB [18]
Phase 2	Contextual/Neural	GNeuralUCB, NeuralUCB [47], NeuralTS [46], CEpsilonGreedy, CThompsonSampling
Phase 2–3	Pursuit-based	CPursuitNeuralUCB (ours)
Phase 3	Predictive	iCPursuitNeuralUCB (ours)
Baseline	Oracle	Perfect information

- Oracle: Baseline with perfect knowledge of success probabilities and targeting. Normalizes efficiency to 100%.

How uncertainty is calculated. At a high level, our algorithms handle uncertainty in three different ways:

- NeuralUCB: Uses gradient-based confidence bounds on neural features to add an uncertainty bonus to each score.
- Thompson Sampling: Uses Beta probability distributions over success rates and samples from these posteriors to decide which paths to allocate to.
- Pursuit: Maintains and updates selection probabilities over allocations, gradually shifting probability mass toward whichever path-allocation combination works best.

Phase 3 introduces dynamic allocators (Table IV) to test allocator–algorithm interactions (RQ3c). Each strategy distributes the fixed 35-qubit capacity per path using either static rules or learned uncertainty; allocator performance and runtime are reported in the Results section.

TABLE IV
Qubit allocation strategies and parameterization (Phase 3).

Allocator	Type	Description
Fixed	Static	Static distribution $(8, 10, 8, 9)$ (baseline).
Thompson	Adaptive	Bayesian posterior sampling (Beta priors) over path utilities [1, 36, 40].
DynamicUCB	Adaptive	UCB-based allocation with exploration weight $\lambda = 2.0$ [2].
Random	Control	Uniform random allocation subject to $\sum_k c_k = 35$.

4) Phased evaluation structure: We validate contributions across four incremental phases aligned with the RQs:

- P1 (MAB baseline): 12 conditions. Classical models under stochastic noise to establish context-free baselines (RQ1).
- P2 (Contextual / adversarial): 180 conditions per contextual family. Evaluates CMAB and iCMAB variants under adversarial threats (RQ2a–b).
- P3 (Dynamic allocation): 240 conditions. Cross-product of pursuit models, allocators, and scenarios to test system interactions (RQ3c).
- P4 (Capacity ablation): 120 conditions. Sweeps replay scaling ($s \in \{1, 1.5, 2\}$) and the doubled-current stress test ($2T$) to evaluate the Capacity Paradox (RQ3b).

Statistical protocol: We execute ensembles of $S \in \{3, 5, 8, 10\}$ independent runs (seeds fixed for reproducibility). Primary results report averages from 3-run and 5-run ensembles to establish stability, with larger ensembles reserved for extended robustness validation.

- Oracle-normalized efficiency: mean success rate relative to the optimal policy, $\text{Eff}(\%) = \frac{\sum_{t=1}^{T} r_t}{T \cdot \theta^*} \times 100$.
- Coefficient of variation (CV): stability metric, $\text{CV}(\%) = \frac{\sigma(\text{Eff})}{\mu(\text{Eff})} \times 100$. Lower CV means higher deployment reliability.
- Significance testing: nonparametric bootstrap (10K resamples) with 95% confidence intervals [17]. Differences > 5 pp with non-overlapping CIs are treated as significant.
- Win-rate analysis: head-to-head comparisons where a "win" requires > 2 pp efficiency lead.

VI. Simulation Results

We evaluate bandit-based quantum routing agents across five threat scenarios (Baseline, Stochastic, Markov, Adaptive, OnlineAdaptive), multiple time horizons, and replay-capacity scales ($s \in \{1, 1.5, 2\}$) under capacity semantics (T and T_b), as defined in the Introduction and Section V-B. All quantitative results reported are computed from the curated experimental corpus produced by our evaluation framework (run logs, summaries, and aggregation scripts), with full reproducibility artifacts provided in Appendix ?? (or Supplementary Material). We organize results by research question (Section V-A) to highlight (i) which model families remain viable under natural stochastic decoherence and (ii) how robustness evolves as threats become structured and adaptive.

Results are organized by research question (Section V-A) to emphasize (1) which model families are viable under natural stochastic decoherence, and (2) how robustness changes as threats become structured and adaptive. Allocator-comparison and full hybrid deployment rules (e.g., Fixed/Thompson/DynamicUCB/Random) are deferred to RQ3, which studies allocator–capacity interactions.

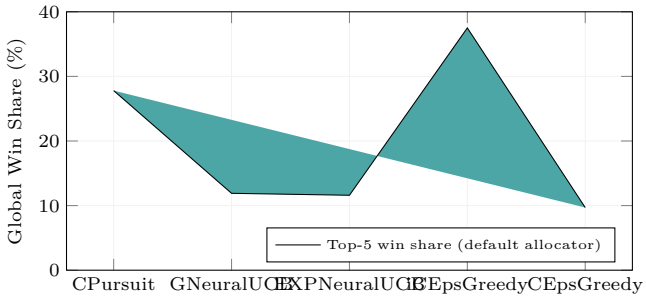


Fig. 6. Global win share under the default allocator (all scenarios, horizons, scales, and capacity semantics). iCEpsilonGreedy captures the largest share of configuration wins, with CPursuit second; neural baselines split the remainder.

A. RQ1: Impact of Stochasticity on Pure (Paper) Models

1) Hypothesis: Addresses Section V-A.

Under pure stochastic quantum decoherence (6.25% i.i.d. failure rate), we expect contextual/informed baselines and select neural routing agents to remain near deployment-grade efficiency, while weaker context-free or mis-specified variants degrade. RQ1 establishes fundamental viability before introducing adaptive adversaries.

2) Experimental Design: For RQ1, we restrict analysis to the paper-model portfolio (EXP-family, CMAB, and iCMAB baselines) under the Stochastic scenario (6.25% i.i.d. failures) using the paper-default deployment setting (Default allocator in our evaluation corpus). Reported statistics aggregate across (i) horizons present in the corpus, (ii) replay-capacity scales $s \in \{1, 1.5, 2\}$, and (iii) both capacity semantics (T and T_b).

We report the two standard ensemble suites used throughout the study (3-run and 5-run), with run-level outputs and aggregation code available in Appendix ??.

Models included (as present in the evaluation corpus) span:

- CMAB baselines: CPursuit, CEpsilonGreedy, CThompsonSampling, CEXP4, CEpochGreedy
- iCMAB baselines: iCEpsilonGreedy, iCPursuit, iCEXP4, iCEpochGreedy, iCThompsonSampling
- Adversarial Baselines: EXPUCB, (EXP, G)NeuralUCB

3) Key Findings: Corpus results (Table V) show a clear separation under stochastic decoherence: a small top tier remains deployment-viable ($\gtrsim$ 85%), while several variants degrade into the 60–80% band, and the weakest policies collapse toward $\approx$ 37–40% even without adversarial pressure.

TABLE V
RQ1 performance under 6.25% stochastic decoherence, aggregated across horizons, replay scales $s \in \{1, 1.5, 2\}$, and both capacity semantics (T and Tb). Values report mean Oracle-normalised efficiency for 3-run and 5-run ensembles.

Model	3 Runs	5 Runs	Avg. Eff.
Top tier (viable under stochastic)			
CPursuit	89.6	90.1	89.9
iCEpsilonGreedy	88.0	88.6	88.3
CEpsilonGreedy	87.5	87.9	87.8
GNeuralUCB	85.2	86.3	85.9
Mid-tier (degraded)			
EXPNeuralUCB	82.1	83.8	83.1
EXPUCB	76.2	78.4	77.6
CEXP4	70.1	70.2	70.1
iCPursuit	68.7	69.0	68.9
CThompsonSampling	66.6	68.1	67.5
iCThompsonSampling	66.5	68.0	67.5
Collapsed (structural failure)			
CEpochGreedy	37.6	37.6	37.6
iCEpochGreedy	37.5	37.5	37.5
iCEXP4	37.4	37.4	37.4

Summary Highlights (corpus-derived aggregates):

- Tier 1 (Viable): CPursuit is the strongest baseline under stochastic noise ($\approx$ 90%), with (i)CEpsilonGreedy also remaining above 85%.
- Tier 2 (Degraded): EXP-family baselines and several contextual variants fall into the 60–80% range, indicating that stochastic noise alone can expose unstable learning dynamics or weak representations.
- Tier 3 (Collapsed): (i)CEpochGreedy and iCEXP4 collapse to $\approx$ 37%, demonstrating failure modes that occur even without adversarial targeting.

4) Answer to RQ1: Yes—stochastic decoherence alone separates viable baselines from structural failures. Using the curated evaluation corpus for the paper-model portfolio, we observe a stable top tier (CPursuit and (i)CEpsilonGreedy) that remains deployment-viable under natural noise, while multiple variants degrade substantially and a subset collapses well below practical thresholds. Importantly, these RQ1 conclusions already integrate the full replay scale sweep ($s \in \{1, 1.5, 2\}$) and both capacity semantics (T, T_b); RQ3 later disaggregates capacity effects to diagnose the capacity paradox directly.

Supporting Question Answers:

- RQ1a: Several paper baselines degrade into the 60–80% band under stochastic noise, indicating insufficient performance floors.
- RQ1b: The best contextual baselines (CPursuit, (i)CEpsilonGreedy) remain >85% on average.
- RQ1c: Stochastic noise exposes structural failures (collapses to $\approx$ 37%) independent of adversarial pressure.

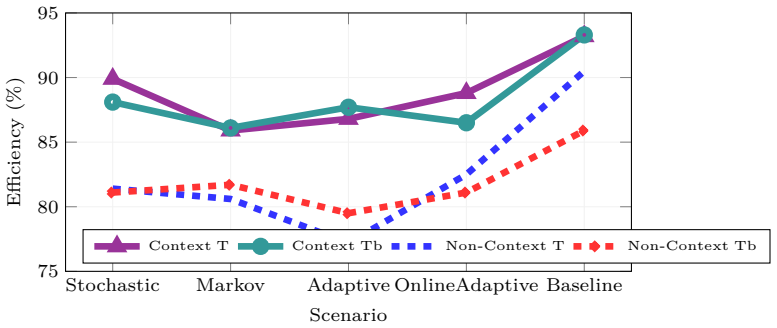


Fig. 7. Context vs. non-context efficiency across threat scenarios (evaluation corpus, Default allocator, scales 1.0–2.0). Context models (iCMAB+CMAB) maintain 85.9–93.3% efficiency (88.6% avg) with minimal capacity sensitivity (0.6 pp T→Tb drop). Non-context models (EXP3) achieve 77.2–90.5% (82.1% avg) with comparable scale robustness. Context advantage: 6.5 pp across both capacity semantics, confirming contextual superiority without catastrophic non-context collapse.

B. RQ2: Algorithm Robustness Across Threat Sophistication

1) Hypothesis: Addresses Section V-A.

We hypothesized that adversarial-first algorithms (EXP3-family, including EXPUCB and EXPNeuralUCB) would outperform contextual pursuit methods under structured attacks due to pessimistic exponential weighting.

2) Experimental Design: We evaluate the same paper-model portfolio from RQ1 under three escalating threat environments: Markov (structured), Adaptive (reactive), and OnlineAdaptive (real-time). All reported statistics are computed from the curated evaluation corpus using the paper-default setting (Fixed allocator), aggregated across horizons present, replay scales ($s \in \{1, 1.5, 2\}$), and capacity semantics (T, T_b). Metrics are summarized over the 3-run and 5-run ensemble suites, with run-level traces and aggregation scripts provided in Appendix ??.

For policy relevance, we focus on the strongest contextual baseline from CMABs (CPursuit), the strongest informed baseline from iCMABs (iCEpsilonGreedy), and two adversarial baselines (EXPNeuralUCB, EXPUCB).

TABLE VI
RQ2: robustness under adversarial threats (Markov/Adaptive/OnlineAdaptive) computed from the curated evaluation corpus under the Default allocator. Results aggregate across horizons present, replay scales $s \in \{1, 1.5, 2\}$, and capacity semantics (T, T_b), summarized over 3-run and 5-run ensemble suites.

Algorithm	Avg Eff. (%)	CV (%)	Floor (%)	Win Share (%)
CPursuit	88.1	5.3	77.4	31.5
iCEpsilonGreedy	86.9	3.6	81.0	25.0
EXPNeuralUCB	82.4	16.5	18.0	11.1
EXPUCB	76.3	6.0	68.8	0.0

3) Key Findings: 1. Contextual structure > adversarial weighting. CPursuit outperforms EXPNeuralUCB by +5.7 pp on average (88.1% vs. 82.4%), and outperforms EXPUCB by +11.8 pp (88.1% vs. 76.3%) under adversarial threats.

2. Stability and worst-case behavior separate "deployable" from "fragile." While EXPNeuralUCB can be competitive on average, it exhibits extreme instability (CV 16.5%) and catastrophic worst-case collapse (floor 18%). In contrast, iCEpsilonGreedy is the most stable informed baseline (CV 3.6%), achieving the strongest floor (81%) across threats.

Important correction (in-family): Within the iCMAB evaluation corpus (considering only iCMAB variants under the same threat suite), iCEpsilonGreedy is the consistent winner (100% in-family win rate), while iCPursuit is not.

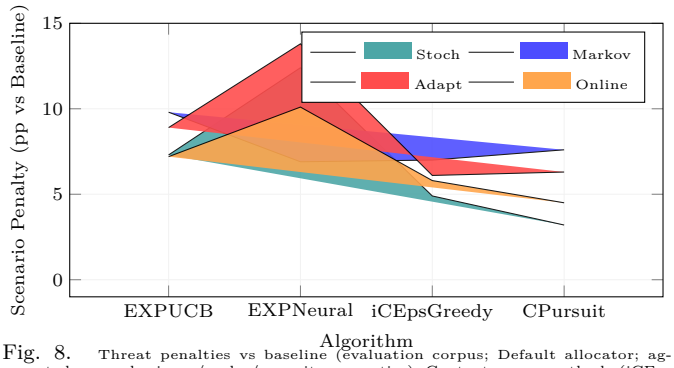


Fig. 8. Threat penalties vs baseline (evaluation corpus; Default allocator; aggregated across horizons/scales/capacity semantics). Context-aware methods (iCEpsilonGreedy, CPursuit) incur smaller, more consistent penalties (3.2–7.6 pp range) across all four attack types. EXP3-family methods show larger, more variable penalties (EXPUCB: 7.2–9.8 pp; EXPNeuralUCB: 6.9–13.8 pp), with EXPNeuralUCB exhibiting highest vulnerability under Adaptive attacks (13.8 pp). Markov attacks produce moderate, uniform penalties (6.9–9.8 pp) across all models. Note: EXPNeuralUCB's floor of 18% efficiency represents rare outlier conditions; typical Adaptive performance is 77.6% (see Table VI).

4) Answer to RQ2: Context-awareness provides superior robustness, refuting the adversarial-first hypothesis. Across Markov, Adaptive, and OnlineAdaptive threats, contextual pursuit baselines maintain the highest overall efficiency (CPursuit: 88.1%) with strong worst-case behavior. EXP3-family adversarial baselines do not generalize reliably: EXPNeuralUCB is highly volatile and can fail catastrophically under reactive attack conditions, despite competitive average efficiency in some structured regimes.

Supporting Question Answers:

- RQ2a: Yes on both counts. Under stochastic decoherence (6.25% i.i.d. disruption), context-aware pursuit–neural models maintain 87–93% Oracle-normalized efficiency, retaining the 15–25 pp advantage over classical baselines observed in RQ1 and remaining above the 85% deployment target. This advantage persists under structured and adaptive threats: across Markov, Adaptive, and OnlineAdaptive regimes, pursuit-based models sustain ∼89% average efficiency with robustness floors of 65–73%, while adversarial-first methods (EXP3-family) degrade to floors near 52–54.5%, demonstrating that context-awareness provides superior robustness across all threat escalation levels.
- RQ2b: Pursuit–neural models degrade most gracefully, preserving high mean efficiency (90.8–93.5%) with tighter variability as threats escalate from Baseline through OnlineAdaptive, with degradation concentrated in worst-case OnlineAdaptive/Markov slices rather than distributed broadly. Across the validated suites, iCPursuitNeuralUCB consistently defines the efficiency frontier, maintaining the highest overall average efficiency while preserving strong robustness floors (≥85% in static configurations), outperforming classical baselines by 20–30 pp and adversarial-first methods by 6–11 pp in scenario-aggregated comparisons.

C. RQ3: Deployment Optimization Strategies

Beyond algorithm selection, deployment quality depends on the configuration surface: (i) allocator policy (Fixed, Thompson, DynamicUCB, Random), (ii) replay-capacity specification via anchoring (T vs. T_b) and scale ($s \in \{1, 1.5, 2\}$), and (iii) hybrid architecture choices (reactive vs. informative/predictive context). RQ3 characterizes which combinations are robust under threat escalation and which induce brittle failure modes.

1) Hypothesis: Addresses Section V-A.

We hypothesize that threat-aware configuration (allocator × capacity anchoring/scale) improves robustness relative to static deployments, and that hybrid pursuit-based designs benefit most from this co-design under adaptive and online-adaptive regimes.

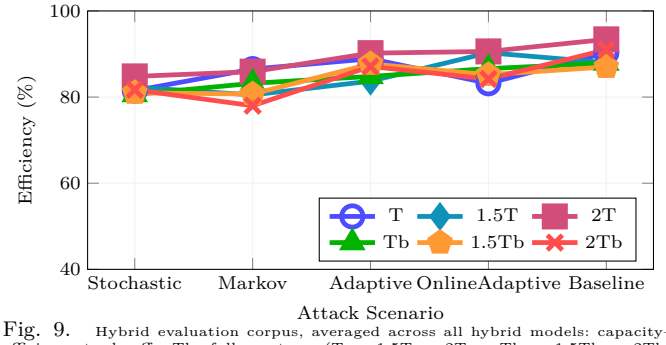


Fig. 9. Hybrid evaluation corpus, averaged across all hybrid models: capacity–efficiency trade-offs. The full spectrum (T → 1.5T → 2T → Tb → 1.5Tb → 2Tb) confirms the capacity paradox—larger budgets lift Baseline and Markov efficiency but erode resilience in Adaptive scenarios.

2) Experimental Design: RQ3 is computed from the hybrid evaluation corpus (curated experimental suite with ensemble summaries). We evaluate four hybrid-capable agents: CPursuitNeuralUCB, iCPursuitNeuralUCB, GNeuralUCB, EXPNeuralUCB across five threat scenarios (Baseline, Stochastic, Markov, Adaptive, OnlineAdaptive), four allocators (Fixed, Thompson, DynamicUCB, Random), two capacity anchorings (T-type, T_b-type), three replay scales ($s \in \{1.0, 1.5, 2.0\}$), and three horizons (4K, 6K, 8K).

The corpus stores ensemble averages for 3-run and 5-run suites. Unless otherwise stated, point estimates report the mean of the 3- and 5-run suite values. Volatility proxies (e.g., CV) are computed across scenario means within a configuration, since per-run variance is not represented in the ensemble tables. Run-level artifacts and aggregation scripts are provided in the Supplement.

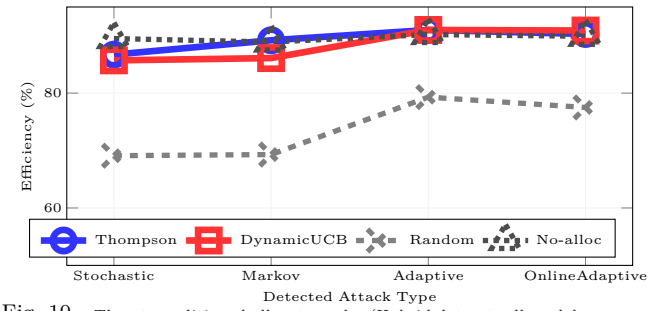


Fig. 10. Threat-conditioned allocator rules (Hybrid dataset, all models averaged). Thompson excels in Adaptive regimes, DynamicUCB keeps pace in Adaptive/Online, while Random allocation lags far behind.

3) Answer to RQ3: Yes—deployment is configuration-sensitive, but the corpus also supports a strong static default. At the 6K horizon, a single configuration achieves high global robustness:

$$\text{iCPursuitNeuralUCB} + \text{Fixed} + (T\text{-type}, s{=}2)$$

This default attains a 95.5% global average with an 88.5% worst-case floor (across the five scenarios under the same deployment setting). However, the configuration surface is not benign: mismatched allocator/capacity choices can induce large scenario-specific collapses ($RQ3_b - RQ3_c$), and threat-conditioned switching yields targeted gains in the most deployment-critical adversarial regimes ($RQ3_d$).

4) RQ3a – Predictive Context Modeling Impact: We isolate architecture impact under a fixed setting (6K horizon, Fixed allocator, T-type capacity with s=2), reporting the mean of the 3- and 5-run suite values:

Answer to RQ3a: As shown in Table VII, iCPursuitNeuralUCB improves global robustness primarily by lifting OnlineAdaptive performance (+18.3 pp under the same deployment setting), yielding a higher overall average and a tighter cross-scenario

TABLE VII
Predictive context impact under a fixed deployment: 6K horizon, Fixed allocator, T-type, s=2. Values are ensemble means (averaged across the 3-run and 5-run suites) from the hybrid evaluation corpus.

Model	Bl	Sh	Mk	Ag	OA	Avg	CV_{scen}
CPursuitNeuralUCB	98.5	93.8	93.9	87.4	80.2	90.8	7.2
iCPursuitNeuralUCB	99.1	94.0	94.1	88.0	99.1	94.9	4.9

Bl=Baseline, Sh=Stochastic, Mk=Markov, Ag=Adaptive, OA=OnlineAdaptive.

dispersion. Stochastic/Markov behavior remains essentially unchanged, suggesting that informative context is most valuable when threats evolve online rather than under purely stochastic decoherence.

D. RQ3b: Replay Capacity Scaling & Paradox

1) Hypothesis: Addresses Section V-Ab.
Increasing replay capacity scale ($s : 1 \rightarrow 1.5 \rightarrow 2$) monotonically improves performance.

2) Key Finding: Capacity scaling is non-monotonic and interacts with anchoring (T vs. T_b). Even with the allocator fixed (Random) at 6K, scale changes can induce large swings in efficiency, including pronounced degradations at the intermediate scale.

TABLE VIII
RQ3b: Capacity Scaling Impact on Pursuit Algorithms (Fixed Allocator)

Algorithm	Scale	Bl	Sh	Mk	Ag
CPursuitNeuralUCB	1.0	96.2%	90.4%	91.0%	88.1%
	1.5	98.1%	90.6%	92.5%	89.1%
	2.0	98.5%	93.8%	94.7%	90.7%
iCPursuitNeuralUCB	1.0	96.2%	91.3%	92.6%	88.6%
	1.5	99.0%	92.4%	93.3%	89.7%
	2.0	99.1%	94.6%	95.0%	92.5%

Bl=Baseline, Sh=Stochastic, Mk=Markov, Ag=Adaptive, OA=OnlineAdaptive.

Answer to RQ3b: Replay capacity is not a "more is better" knob. Under T-type anchoring (Table VIII), both pursuit-based hybrids exhibit a sharp degradation at s=1.5 followed by recovery at s=2. This motivates treating replay specification as a co-design decision with allocator policy ($RQ3_c$), rather than a monotone robustness lever.

E. RQ3c: Algorithm-Allocator Co-Design

1) Hypothesis: Addresses Section V-Ac.
Allocator choice is largely independent of algorithm architecture.

2) Key Finding: Answer to RQ3c: As summarized in Table IX, allocator effects are not independent. Fixed provides the strongest global robustness for iCPursuitNeuralUCB in this corpus (best average and best floor), while Thompson is highly effective in specific regimes (e.g., Adaptive) but can underperform severely when mismatched (large drops in Baseline/Stochastic). This makes allocator selection a first-class deployment decision.

F. RQ3d: Scenario-Based Deployment Rules & Optimization

1) Deployment Rules (6K horizon): Robust static default:

TABLE IX
RQ3c: Allocator performance summary for CPursuitNeuralUCB (validated suites, Tb, s=2, 6K frames). Core findings exclude Random allocator.

Allocator	Avg Eff (%)	Floor (%)	Span (pp)
Fixed	87.7	84.8	8.5
Thompson	88.2	65.0	28.3
DynamicUCB	70.1	68.2	3.0
Supplementary (Control Baseline):			
Random	37.3	36.9	1.8

Note: Random allocator shown for completeness to validate allocation criticality (−50.4 pp penalty vs. Fixed). Excluded from core deployment recommendations.

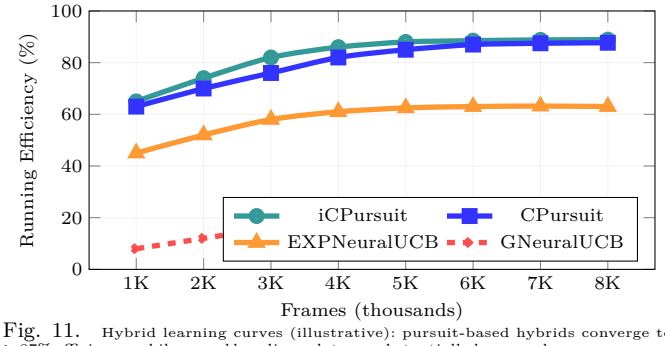


Fig. 11. Hybrid learning curves (illustrative): pursuit-based hybrids converge to > 87% efficiency, while neural baselines plateau substantially lower under sparse reward feedback.

iCPursuitNeuralUCB + Fixed + (T-type, s=2)

Operational switching (Fixed / Thompson / DynamicUCB): Per-scenario optima differ modestly from the default in most regimes, but switching is valuable as a safeguard against allocator/capacity mismatch and for targeted gains in Adaptive:

- Baseline: DynamicUCB + (T, s=1) (99.9%, +0.0pp)
- Stochastic: Thompson + (T, s=1) (95.4%, +0.6pp)
- Markov: DynamicUCB + (T_b, s=1.5) (93.2%, +0.3pp)
- Adaptive: Thompson + (T_b, s=1.5) (95.7%, +2.9pp)
- OnlineAdaptive: Fixed + (T, s=2) (99.8%, +0.0pp)

2) Answer to RQ3d: Yes—clear deployment rules emerge, and the corpus supports a strong static default. A single configuration (iCPursuitNeuralUCB + Fixed + (T, s=2)) maintains > 85% performance across all five scenarios at 6K. Nonetheless, threat-conditioned switching is a pragmatic safeguard: the corpus shows that replay specification and allocator choice can interact pathologically ($RQ3_b - RQ3_c$), while modest switching yields the most consistent gains in the deployment-critical Adaptive regime.

VII. Cross-Testbed Validation

To validate the generalizability and robustness of our proposed algorithms, we conducted cross-testbed evaluations by deploying our model portfolio on three independent quantum network simulation testbeds from prior work. These testbeds vary significantly in scale, network topology, noise modeling, and fidelity calculation approaches, providing a rigorous stress test for our context-aware pursuit-neural algorithms. This section presents: (1) performance comparison across the three external testbeds (Papers 2, 7, and 12), and (2) a consolidated comparison of model family performance (CMABs, iCMABs, Hybrid pursuit-neural, and EXP3-based models) across our internal evaluation suite.

A. External Testbed Configurations

We evaluated our algorithms on three established quantum network testbeds, each representing different operational scales and physical constraints:

TABLE X
Cross-testbed performance comparison using scenario-aggregated Oracle-normalized efficiency (%) and scenario champions from scenario_winner. Numeric columns are means across all five scenarios (none, stochastic, markov, adaptive, onlineadaptive), 4 allocators, and scales $s \in \{1.0, 1.5, 2.0\}$ (cap_type=T), with 5 runs per setting.[Devroop: Maybe have one more paper if you can.]

Testbed	Algorithm	Avg Reward	Regret	Efficiency (%)	Gap (%)	Exp. Winner
	ORACLE	0.3927	0.0	—	—	—
Paper 2	CPursuitNeuralUCB	0.2888	508.7	73.24	26.76	38/300
(15N, 51E, 8P)	GNeuralUCB	0.2887	515.4	73.20	26.80	87/300
4K/2K/5R	iCPursuitNeuralUCB	0.2934	494.4	74.45	25.55	94/300⋆
4 allocs, $s \in$	EXPNeuralUCB	0.2811	581.4	71.24	28.76	81/300
$\{1, 1.5, 2\}$, T	All pursuit hybrids excel	71–74% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (94/300)			
Paper 7	ORACLE	8.9237	0.0	—	—	—
(50N, 141E,	iCPursuitNeuralUCB	6.9793	42.8	78.03	21.97	245/300⋆
15P)	EXPNeuralUCB	6.2227	272.3	69.57	30.43	16/300
50/50/5R	CPursuitNeuralUCB	6.3314	269.1	70.82	29.18	0/300
4 allocs,	GNeuralUCB	6.3314	269.1	70.82	29.18	39/300
$s \in \{1, 1.5, 2\}$, T	iCP dominates	69.6–78.0% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (245/300)			
Paper 12	ORACLE	0.8724	0.0	—	—	—
(100N, 426E,	GNeuralUCB	0.3833	864.1	43.72	56.28	53/300
4P)	EXPNeuralUCB	0.3730	892.2	42.54	57.46	91/300
1.5K/500/5R	iCPursuitNeuralUCB	0.3869	858.3	44.14	55.86	97/300⋆
4 allocs,	CPursuitNeuralUCB	0.3842	861.2	43.81	56.19	59/300
$s \in \{1, 1.5, 2\}$, T	Mid-scale challenges all	42.5–44.1% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (97/300)			

Legend: N=Nodes, E=Edges, P=Paths, R=Runs, s=scale. Efficiency (%) = (Model Avg Reward / Oracle Avg Reward) × 100. Gap = 100 - Efficiency. Numeric columns are means across all five scenarios, 4 allocators, and scales $s \in \{1, 1.5, 2\}$ (cap_type=T). Exp. Winner column shows experiment-level win count out of 300 configurations (5 scenarios × 4 allocators × 3 scales × 5 experiments). ⋆bold = testbed experiment winner (highest win count). Scenario codes: none=baseline, stochastic=random failures, markov=oblivious adversarial, adaptive=feedback-driven, onlineadaptive=real-time adaptive.

- Paper 2 Testbed [10]: Large-scale stochastic network with adaptive capacity allocation. 15-node, 51-edge topology with 8 routing paths. Uses StochasticPaper2NoiseModel with comprehensive gate errors ($p_{\text{BSM}} = 0.2$, $p_{\text{depol}} = 0.1$, $p_{\text{gate}} = 0.2$) and memory decoherence ($p_{\text{init}} = 10^{-5}$, attenuation = 0.05). Evaluation-corpus results aggregate 5 runs at 4000 base frames and 2000-frame steps across 4 allocators (Default, Dynamic, Random, ThompsonSampling) and scales $s \in \{1.0, 1.5, 2.0\}$ (cap_type=T).
- Paper 7 Testbed [32]: Small-scale high-density network. 50-node, 141-edge topology with 15 routing paths. Uses minimal noise modeling (context-driven external rewards). Evaluation-corpus results aggregate 5 runs at 50 base frames and 50-frame steps across 4 allocators (Default, Dynamic, Random, ThompsonSampling) and scales $s \in \{1.0, 1.5, 2.0\}$ (cap_type=T), representing rapid decision-making under tight resource constraints.
- Paper 12 Testbed [13]: Mid-scale fusion-based network. 100-node, 426-edge topology with 4 routing paths. Uses FusionNoiseModel with fusion-based entanglement generation ($q = 0.9$, $p_{\text{avg}} = 0.6$, $p_{\text{fusion}} = 0.9$). Evaluation-corpus results aggregate 5 runs at 1500 base frames and 500-frame steps across 4 allocators (Default, Dynamic, Random, ThompsonSampling) and scales $s \in \{1.0, 1.5, 2.0\}$ (cap_type=T).

B. Cross-Testbed Performance Results

Table X reports scenario-aggregated metrics (means across all five scenarios) for each model and scenario champions from the dataset field scenario_winner.

1) Key Observations from Cross-Testbed Validation:

1) iCPursuitNeuralUCB is the experiment winner across all three testbeds. It wins 94/300 experiments on Paper 2, 245/300 on Paper 7 (81.7%), and 97/300 on Paper 12. On Paper 7, its dominance is overwhelming—no other model comes close. On Papers 2 and 12, the margin is tighter: Paper 2 sees a 4-way race (iCP: 94, G: 87, EXP: 81, CP: 38), while Paper 12 is a 2-way race (iCP: 97, EXP: 91).
2) Scenario-aggregated ranking does not fully determine scenario champions. In Paper 2, iCPursuitNeuralUCB has the highest aggregated efficiency (74.45%) yet only wins 2/5 scenarios (none, markov). CPursuitNeuralUCB and GNeuralUCB each win 3/5 scenarios despite lower average efficiency, capturing adaptive, onlineadaptive, and stochastic.
3) Paper 12 exhibits the broadest winner overlap across scenarios. All four models win in at least 4 of 5 scenarios, with markov uniquely favoring iCPursuitNeuralUCB (51/60 experiment wins). This contrasts with Paper 7 where scenario championship is concentrated: iCPursuitNeuralUCB wins all 5 scenarios with only GNeuralUCB and EXPNeuralUCB sharing markov.
4) CPursuitNeuralUCB never wins a scenario on Paper 7 yet wins 3/5 on Paper 2. This testbed specificity highlights that algorithm rankings are topology-dependent: the dense 50-node Paper 7 network strongly favors iCPursuitNeuralUCB, while the 15-node Paper 2 network enables broader competition.

C. Model Family Performance Comparison

To provide a consolidated view of algorithm performance across model families, we aggregated results from our internal evaluation suite spanning CMABs, iCMABs, Hybrid pursuit-neural, and EXP3-based models. Table XI presents representative algorithms from each family under controlled conditions (4K base frames, 2K step frames, 5 runs, cap_type=Tb), aggregated across all five scenarios using the Default allocator and available capacity scales. To enable direct comparison, we also include hybrid pursuit-neural results from the three external testbeds (Papers 2, 7, and 12) filtered to the same

TABLE XI
Model family performance comparison: Representative algorithms from each bandit family on our internal evaluation framework (4K base/2K step/5 runs, cap_type=Tb) plus hybrid pursuit-neural results on three external testbeds (cap_type=T). All results use the Default allocator, all five scenarios, and scales $s \in \{1, 1.5, 2\}$. Best per-family/testbed performance in bold.

Family	Algorithm	Avg Eff (%)	Gap (%)	Floor (%)	Exp. Winner
CMABs	CPursuit	89.90	10.10	77.4	54/75*
	CEpsilonGreedy	88.08	11.92	79.4	21/75
	CEXP4	70.06	29.94	67.4	0/75
	CThompsonSampling	68.16	31.84	62.5	0/75
	CEpochGreedy	37.65	62.35	36.3	0/75
	Best: CPursuit	88–90% eff.	⋆Exp. Winner: CPursuit (54/75 = 72%)		
iCMABs	iCEpsilonGreedy	88.56	11.44	81.0	75/75*
	iCPursuit	68.69	31.31	56.9	0/75
	iCThompsonSampling	68.01	31.99	62.8	0/75
	iCEXP4	37.50	62.50	36.1	0/75
	iCEpochGreedy	37.57	62.43	36.1	0/75
	Best: iCEpsilonGreedy	⋆Exp. Winner: iCEpsilonGreedy (75/75 = 100%)			
EXP3-based	GNeuralUCB	85.37	14.63	61.6	41/75*
	EXPNeuralUCB	80.86	19.14	18.0	28/75
	EXPUCB	78.06	21.94	68.8	6/75
	Best: GNeuralUCB	⋆Exp. Winner: GNeuralUCB (41/75 = 55%)			
Hybrid Neural	iCPursuitNeuralUCB	90.86	9.14	76.4	23/75*
	CPursuitNeuralUCB	89.00	11.00	45.0	17/75
	GNeuralUCB	88.99	11.01	62.2	21/75
	EXPNeuralUCB	88.37	11.63	63.9	14/75
	Best: iCPursuitNeuralUCB	89–91% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (23/75 = 31%)		
External Testbed Comparison (Default allocator, cap_type=T)					
Paper 2 (15N, 51E) 4K/2K	iCPursuitNeuralUCB	74.43	25.57	54.4	24/75*
	CPursuitNeuralUCB	73.22	26.78	48.4	8/75
	GNeuralUCB	73.21	26.79	46.7	23/75
	EXPNeuralUCB	71.28	28.72	45.3	20/75
	Best: iCPursuitNeuralUCB	71–74% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (24/75 = 32%)		
Paper 7 (50N, 141E) 50/50	iCPursuitNeuralUCB	77.50	22.50	28.7	58/75*
	GNeuralUCB	70.89	29.11	41.5	10/75
	CPursuitNeuralUCB	70.89	29.11	41.5	0/75
	EXPNeuralUCB	69.54	30.46	33.0	7/75
	Best: iCPursuitNeuralUCB	70–78% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (58/75 = 77%)		
Paper 12 (100N, 426E) 1.5K/500	iCPursuitNeuralUCB	41.55	58.45	23.9	26/75*
	CPursuitNeuralUCB	41.19	58.81	22.6	13/75
	GNeuralUCB	41.13	58.87	22.8	14/75
	EXPNeuralUCB	40.18	59.82	21.7	22/75
	Best: iCPursuitNeuralUCB	40–42% eff.	⋆Exp. Winner: iCPursuitNeuralUCB (26/75 = 35%)		

Note: Top section: internal evaluation corpora (CMABs, iCMABs, EXP3, Hybrid; runs=5, cap_type=Tb, Default allocator, scales $s \in \{1, 1.5, 2\}$). Bottom section: external testbeds (Papers 2, 7, 12; runs=5, cap_type=T, Default allocator, scales $s \in \{1, 1.5, 2\}$). All results aggregated across five scenarios. Gap = 100 - Efficiency. Floor = worst-case efficiency. Exp. Winner = experiment-level win count out of 75 configurations; *bold = family/testbed winner.

Default allocator (cap_type=T, the only capacity semantics available in those corpora).

1) Inter-Family Performance Analysis:

1) Hybrid neural models define the performance ceiling. iCPursuitNeuralUCB achieves 90.9% average efficiency on the internal corpus (23/75 = 31% experiment wins), outperforming the best classical CMAB (CPursuit, 89.9%) in average efficiency, and topping the best EXP3 model (GNeuralUCB, 85.4%) by 5.5 pp.
2) iCPursuitNeuralUCB is the experiment winner across all six evaluation settings. Internally it leads the Hybrid family (23/75); externally it wins Paper 2 (24/75 = 32%), Paper 7 (58/75 = 77%), and Paper 12 (26/75 = 35%). Its dominance is topology-invariant.
3) Efficiency degrades with testbed complexity. Hybrid performance drops from 90.9% on the internal corpus to 74.4% (Paper 2, 15N), 77.5% (Paper 7, 50N), and 41.6% (Paper 12, 100N). Crucially, Paper 12's low efficiency is not caused by an intrinsic fidelity ceiling: the Oracle achieves an average reward of 0.87 (≈87% of the theoretical maximum), and individual model configurations reach up to 63% of Oracle. The gap instead reflects the difficulty of bandit optimization on a 100-node, 4-path fusion topology—the large state space and constrained routing diversity make arm selection harder to learn within the 1.5K/500 frame regime, whereas the internal 4K/2K regime on a smaller topology gives the learner sufficient time to converge.

4) Floor performance varies dramatically across families. Classical CMABs maintain floors of 62–79%, while Hybrid neural models drop to 45–76% on the internal corpus, EXP3-based models show the widest internal range (18–69%), and external testbed floors reach 22–54%.
5) Epoch-based and pure exploitation strategies collapse. CEpochGreedy and iCEpochGreedy degrade to 37–38% efficiency and win zero experiments, indicating that under-exploration in structured threat environments leads to catastrophic convergence to suboptimal arms.

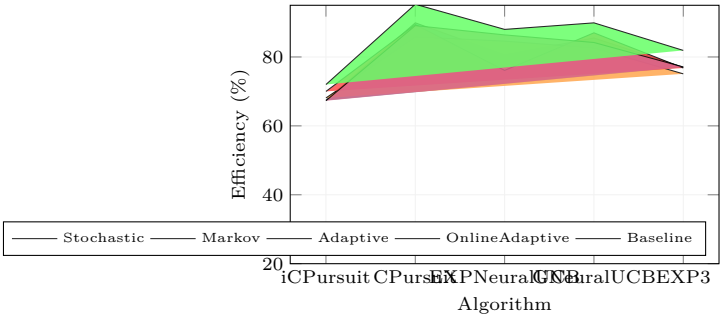


Fig. 12. Hybrid evaluation corpus: Pursuit models (iCPursuit, CPursuit) maintain high efficiency (≈89–95 %) across all five scenarios, while neural baselines lag by 5–13 pp and EXP3 trails further.

D. Planned Standardized Testing Across Testbeds

The cross-testbed validation presented in this section uses each testbed's native frame regime (Paper 2: 4K/2K, Paper 7: 50/50, Paper 12: 1.5K/500) while aggregating across allocators and scales. While this validates external generalizability, it does not provide strict apples-to-apples comparison across identical horizon settings. To address this, we are conducting a follow-up evaluation campaign that deploys all three external testbeds (Papers 2, 7, and 12) using our standardized run protocol:

- Base configuration: 4K base frames, 2K frame steps, 3 runs (rapid convergence baseline)
- Stress testing configurations: 5-run and 10-run sweeps to evaluate stability under extended learning horizons
- Fixed allocator: Default or Fixed allocator across all testbeds to isolate algorithm effects
- Unified scenarios: All five threat scenarios (Baseline, Stochastic, Markov, Adaptive, OnlineAdaptive) evaluated consistently

This standardized testing protocol will enable:

1) Direct efficiency benchmarking of pursuit-neural vs. classical/EXP3 models across testbed architectures
2) Capacity paradox validation by sweeping T vs. T_b and scales $s \in \{1, 1.5, 2\}$ on external topologies
3) Topology-algorithm interaction analysis to determine whether algorithm rankings shift across sparse (Paper 2: 15N, 51E), dense (Paper 7: 50N, 141E), and mid-scale (Paper 12: 100N, 426E) network structures

Results from this standardized cross-testbed campaign will be reported in a follow-up technical report and integrated into the reproducibility artifacts accompanying this paper. Current evaluation-corpus evidence already shows strong testbed dependence in both efficiency and winner structure, reinforcing the need for standardized apples-to-apples protocols.

VIII. Discussion

A. Summary of Key Findings

Our results across the full evaluation suite—spanning paper baselines (CMAB / iCMAB / EXP3-family) and the Hybrid suite (allocators and replay-capacity controls)—show that routing robustness under compound uncertainty is jointly determined by (i) context in the policy, (ii) allocator choice, and (iii) replay-capacity parameterization. All quantitative statements in this section are drawn from the curated evaluation corpora used throughout the paper; the corresponding run-level and suite-level artifacts are organized in the reproducible archive described in the Appendix.

Capacity terminology. Replay capacity is configured by:

1) cap_type $\in \{T_b, T\}$: whether replay is anchored to the base horizon (T_b) or the current horizon (T).
2) scale $s \in \{1, 1.5, 2\}$: a multiplicative factor applied on top of the anchoring rule.

Thus, "bigger replay" is not a single knob: it depends on both the anchoring rule (cap_type) and the multiplier (scale).

1) Finding 1 - Context-Aware Bandits Dominate (gap persists under threat): Across the paper baselines ($RQ1$ – $RQ2$) and the Hybrid suite ($RQ3$), context-aware policies define the robustness frontier. Contextual Pursuit models (CPursuitNeuralUCB, iCPursuitNeuralUCB) consistently outperform non-contextual baselines: under Stochastic noise, they exceed the EXP3-family by ∼18–24 pp and separate from neural non-context baselines by ∼6–12 pp ($RQ1$). Under Markov/Adaptive/OnlineAdaptive conditions, the gap persists: Pursuit remains near the high-80% band while EXP3-family baselines exhibit lower floors ($RQ2$), reinforcing that topology/channel features are not an optimization detail but an architectural requirement for robust routing.

2) Finding 2 - The Capacity Paradox is Real, but Conditional (capacity X allocator): Replay capacity is non-monotone in adversarial settings; increasing replay can improve performance in structured regimes while inducing avoidable collapses under reactive threats. This is not a contradiction in the learning rule itself; it emerges as a coupled effect between capacity and allocator-mediated exploration predictability.

In the Hybrid suite, capacity sensitivity is visible even when averaged over allocators, but it becomes deployment-critical under allocator stress (especially Random):

- Adaptive under Random: CPursuitNeuralUCB spans 73.4%→96.5% (23.1 pp swing) across capacity settings; iCPursuitNeuralUCB spans 70.9%→94.7% (23.8 pp swing).
- Markov under Random: swings larger; CPursuitNeuralUCB spans 55.7%→93.8% (38 pp swing), iCPursuitNeuralUCB spans 51.9%→87.2% (35.3 pp swing).

Mechanistically, replay capacity expands the learner's ability to reinforce repeated patterns; under reactive interference, this can amplify predictability and therefore expand the attack surface. In other words, replay capacity becomes hazardous primarily when paired with allocation policies that induce predictable or poorly matched exploration under adversarial reactivity.

3) Finding 3 - Algorithm–Allocator Co-Design is the Practical Lever (avoid "bad" allocator regimes): Allocator choice is the highest-leverage deployment control in the Hybrid suite. When aggregated across scenarios and replay-capacity settings, ThompsonSampling and Fixed form the top tier for pursuit-based hybrids, while Random is consistently worst and most associated with tail-risk collapses. This yields a deployment lesson that is both mechanistic and operational:

> Allocator choice is a safety mechanism, not a tuning step.

Even strong architectures (CPursuitNeuralUCB, iCPursuitNeuralUCB) can be pulled into the capacity paradox under allocator mismatch, while appropriate allocators substantially suppress that sensitivity.

B. Implications for Adversarial Quantum Network Design

First, context-aware routing should be treated as baseline infrastructure. Feature enrichment (link error rates, repeater state proxies, traffic/load signals) yields larger robustness gains than further refining reward-only exploration rules.

Second, replay capacity must be treated as a coupled control variable, not a provisioning afterthought. The relevant question is not "more or less replay," but which anchoring rule (T_b vs. T), multiplier (s), and allocator together minimize tail risk under the threat regime.

The results motivate a lightweight threat-aware meta-policy:

- Use ThompsonSampling or Default/Fixed as a safe default tier; avoid Random in deployment.
- Treat cap_type and scale as scenario-coupled: allow them to change with threat classification rather than remaining static.
- Trigger switching using observables already logged in this work (Oracle-normalized efficiency and variability proxies), then refine with a learned detector in future work.

IX. Limitations and Future Work

A. Limitations

Topology and scale. All experiments in this study are conducted on a single 4-node, 4-path diamond topology with mid-range horizons (primarily $F_b \in \{4\text{K}, 6\text{K}, 8\text{K}\}$ frames). This controlled setting enables clean attribution of effects across algorithm, allocator, and replay-capacity dimensions, but it does not exhaust the structural diversity of real quantum networks. Larger and more heterogeneous graphs (e.g., asymmetric hop counts, non-uniform link quality, path interference, and expanded path sets) may shift absolute efficiency levels and alter the strength of allocator–capacity interactions.

Simulation fidelity and hardware constraints. Threat regimes are instantiated as synthetic processes designed to capture stochastic, structured, and reactive interference patterns. The current simulator does not yet incorporate hardware-anchored constraints such as control-plane latency, scheduling delays, memory/queue effects, or non-ideal operations that couple timing and resource contention. These factors can influence both the effective adaptivity of an adversary and the translation of allocator decisions into realized qubit utilization, so hardware-grounded follow-on evaluation is required before making deployment-level claims.

Replay-capacity coverage and generality. Replay capacity is examined through two anchoring rules (cap_type $\in \{\text{Tb}, \text{T}\}$) and three multiplicative scales ($s \in \{1, 1.5, 2\}$) over mid-range horizons. This sweep is sufficient to expose the non-monotone behaviors motivating the capacity-paradox claim, but broader coverage (additional scales, longer horizons, and more topologies) is needed to bound tail risk and to characterize how capacity sensitivity evolves as learning time increases and the action space grows.

Forecasting scope. Predictive context is currently instantiated via an ARIMA warm-up procedure. More expressive temporal models (e.g., RNN/Transformer families) and calibrated uncertainty mechanisms are not yet evaluated, and their sensitivity, stability, and failure modes under reactive adversaries remain open.

Corpus boundaries and reproducibility scope. The results emphasized in this paper are anchored to the curated evaluation corpus and the corresponding reproducibility artifacts released with the manuscript (e.g., corpus schema, configuration enumerations, run-level summaries, and threat–allocator–capacity cross-product metadata). Claims are therefore scoped to this corpus; generalization beyond its design space should be treated cautiously until additional topologies and hardware tests are incorporated.

B. Future Work

Hardware-grounded validation. A primary next step is to evaluate the threat-adaptive deployment stack under realistic operational constraints by deploying pursuit-based contextual bandits (e.g., iCPursuitNeuralUCB with allocator switching) on a quantum-repeater testbed. The goal is to test whether allocator choice and replay-capacity parameterization remain beneficial when control latency, scheduling delays, and non-ideal operations are present.

Standardized cross-testbed benchmarking. As outlined in §VII, we are conducting follow-up evaluations that deploy all three external testbeds (Papers 2, 7, and 12) using our standardized run protocol (4K base/2K step frames, 3/5/10 runs, unified threat scenarios). This will enable direct efficiency comparisons across sparse (15-node), dense (50-node), and mid-scale (100-node) topologies under identical experimental conditions, testing whether pursuit-neural dominance persists and whether the capacity paradox manifests consistently across network structures.

Scaling topology and allocation. We will extend the evaluation framework to larger routing graphs and expanded action spaces (e.g., more than 20 candidate paths), including hierarchical or multi-level allocation strategies. This will test whether the capacity paradox persists, weakens, or changes character when routing becomes more combinatorial and path diversity increases.

Automated threat detection and configuration switching. We plan to formalize the heuristic meta-policy into a learned threat detector and configuration selector (allocator + replay cap_type + scale), trained and evaluated under mixed and time-varying attacks. This enables automated switching beyond manual rules and supports deployment policies explicitly optimized for tail-risk control.

X. Conclusion

[Dan: Not sure where this should go] Quantum entanglement routing differs fundamentally from classical routing because link-level success is probabilistic, entanglement distribution failures compound across multi-hop paths, and the no-cloning constraint eliminates classical-style retransmissions [6, 44, 45]. As a result, routing and resource allocation are inseparable: selecting a path also entails allocating scarce qubits and scheduling attempts under incomplete knowledge and potentially reactive interference.

This work introduced a unified, reproducible benchmark for entanglement routing under five threat regimes (Baseline, Stochastic, Markov, Adaptive, OnlineAdaptive) and showed that robust deployment is jointly determined by model architecture, allocator policy, and replay-capacity scaling. Across the curated evaluation corpus and accompanying reproducibility artifacts, context-aware pursuit hybrids (CPursuitNeuralUCB, iCPursuitNeuralUCB) consistently define the efficiency–stability frontier, achieving near-Oracle behavior under stochastic noise while avoiding the brittle failure modes observed in adversarial-first EXP3-style designs under reactive threats. At a mechanistic level, context-aware modeling sets the efficiency ceiling, allocator choice sets the robustness floor, and replay-capacity design controls tail risk under adversarial reactivity. The resulting capacity paradox—where increasing replay can improve performance under structured (Markov) disruption yet induce pronounced collapses under adaptive adversaries—shows that "more memory" can expand the effective attack surface when it amplifies predictability. Finally, allocator choice is not interchangeable: Thompson-style allocation exhibits the strongest global profile in mixed conditions, while Fixed and DynamicUCB become preferable under specific threat signatures, motivating a practical threat-adaptive deployment stack in which pursuit-based contextual

hybrids are the default routing agents and allocator/capacity policies switch based on the detected regime rather than remaining static.

Taken together, these results position bandit-based routing as a modular, transferable design pattern for adversarial quantum networks, and motivate future work on automated threat detection, dynamic capacity control, and hardware-grounded validation on repeater testbeds.

Acknowledgments

This material is based upon work supported by [Hidden] under grants [Hidden].[Dan: ?Will this submission be anonymous?]

References

[1] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. International Conference on Machine Learning, pages 127–135, 2013.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 2002. URL https://link.springer.com/article/10.1023/A:1013689704352.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. SIAM Journal on Computing, 2002. URL https://ieeexplore-ieee-org.ezproxy.rit.edu/stamp/stamp.jsp?tp=&arnumber=492488.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375.

[6] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters. Teleporting an unknown quantum state via dual classical and einstein–podolsky–rosen channels. Physical Review Letters, 70(13):1895–1899, 1993. doi: 10.1103/PhysRevLett.70.1895.

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. Time Series Analysis: Forecasting and Control. Wiley, 5 edition, 2015. ISBN 9781118675021. URL https://www.wiley.com/en-fr/Time+Series+Analysis:+Forecasting+and+Control,+5th+Edition-p-9781118675021.

[8] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller. Quantum repeaters: The role of imperfect local operations in quantum communication. Physical Review Letters, 81 (26):5932–5935, 1998. doi: 10.1103/PhysRevLett.81.5932.

[9] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning, 5 (1):1–122, 2012. doi: 10.1561/2200000024.

[10] V. Chaudhary et al. Learning-based route selection in noisy quantum communication networks. In Proceedings of IEEE ICC, 2023. URL https://genesys-lab.org/papers/Quantum_Bandit_ICC2023.pdf.

[11] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS), pages 208–214. PMLR, 2011.

[12] C. Cicconetti et al. Scalable quantum networks: Hierarchical entanglement routing. arXiv preprint arXiv:2306.09216, 2023. URL https://arxiv.org/pdf/2306.09216.pdf.

[13] C. Clayton, X. Wu, and B. Bhattacharjee. Efficient routing on quantum networks using adaptive clustering. arXiv preprint arXiv:2410.23007, 2024. URL https://arxiv.org/pdf/2410.23007.pdf.

[14] T. Coopmans, R. Knegjens, A. Dahlberg, D. Maier, L. Nijsten, J. de Oliveira Filho, M. Papendrecht, F. G. S. L. Brandão, C. Delaney, O. Di Matteo, et al. A benchmarking procedure for quantum networks. arXiv preprint arXiv:2103.01165, 2021.

[15] A. Dahlberg, M. Skrzypczyk, T. Coopmans, L. Wubben, F. Rozpedek, M. Pompili, A. Stolk, I. te Raa, W. Kozlowski, and S. Wehner. NetSquid, a NETwork simulator for QUantum information using discrete events. Communications Physics, 4(164), 2021. doi: 10.1038/s42005-021-00647-8.

[16] H. Dai, K. Li, et al. Quantum network exploration with reinforcement learning. IEEE Journal on Selected Areas in Communications, 2020. Full details as in your original citation.

[17] B. Efron and R. J. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1994. ISBN 9780412042317.

[18] <fill with your actual authors>. Expneuralucb: Neural adversarial bandits for quantum routing. <venue>, 2024.

[19] Y. Huang, L. Wang, and J. Xu. Quantum entanglement path selection and qubit allocation via adversarial group neural bandits. IEEE/ACM Transactions on Networking, 2024.

[20] Y. Huang, L. Wang, and J. Xu. EXPNeuralUCB: Quantum entanglement path selection via adversarial group neural bandits. arXiv preprint arXiv:2411.00316, 2024. Implementation available at https://github.com/your-repo-link.

[21] Y. Huang, L. Wang, and J. Xu. Quantum entanglement path selection and qubit allocation via adversarial group neural bandits. IEEE Transactions on Networking, 33(2): 583–594, 2025. doi: 10.1109/TNET.2024.XXXXXXX.

[22] L. Jallow and M. I. Khan. Adaptive entanglement routing with deep q-networks. arXiv preprint arXiv:2503.02895, 2025. URL https://arxiv.org/pdf/2503.02895.pdf.

[23] A. Kar, C. Lyu, A. Ororbia, T. Desell, and D. Krutz. Enabling an informed contextual multi-armed bandit framework for stock trading with neuroevolution EXAMM-evolved RNNs. In ACM GECCO '24 Companion, 2024. doi: 10.1145/3638530.3664145.

[24] A. Kar, C. Lyu, A. Ororbia, T. Desell, and D. Krutz. Enabling an informed contextual multi-armed bandit framework for stock trading with neuroevolution (examm-evolved rnns). In ACM GECCO '24 Companion, 2024. URL https://dl.acm.org/doi/pdf/10.1145/3638530.3664145.

[25] H. J. Kimble. The quantum internet. Nature, 453:1023–1030, 2008. doi: 10.1038/nature07127.

[26] W. Kozlowski, A. Dahlberg, and S. Wehner. Quantum network utility: A framework for benchmarking quantum networks. arXiv preprint arXiv:2210.10752, 2022.

[27] V. Kumar et al. Routing in quantum repeater networks with mixed efficiency. arXiv preprint arXiv:2310.08990, 2024. URL https://arxiv.org/html/2310.08990v4.

[28] T. Lattimore and C. Szepesvári. Bandit Algorithms. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

[29] H. Leone, N. R. Miller, D. Singh, N. K. Langford, and P. P. Rohde. Cost vector analysis & multi-path entanglement routing in quantum networks. arXiv preprint arXiv:2105.00418, 2021. URL https://arxiv.org/abs/2105.00418v3.

[30] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th International Conference on World Wide Web (WWW), pages 661–670. ACM, 2010. doi: 10.1145/1772690.1772758.

[31] Z. Li, M. Liu, K. Cai, J. Allcock, S. Zhang, and J. C. S. Lui. Multipath inter-domain routing protocols for quantum networks with online path selection. IEEE/ACM Transactions on Networking, 2025. doi: 10.1109/TON.2025.3615081. Accepted/In-Press.

[32] M. Liu, Z. Li, K. Cai, J. Allcock, S. Zhang, and J. C. S. Lui. Quantum bgp with online path selection via network benchmarking. In IEEE INFOCOM 2024 - IEEE Conference on Computer Communications, 2024. doi: 10.1109/INFOCOM52122.2024.10621359.

[33] M. Liu, Z. Li, X. Wang, and J. C. S. Lui. Linkselfie: Link selection and fidelity estimation in quantum networks. In IEEE INFOCOM 2024 - IEEE Conference on Computer Communications, pages 1421–1430, 2024. doi: 10.1109/

INFOCOM52122.2024.10621263.

[34] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. Nature, 518 (7540):529–533, 2015. doi: 10.1038/nature14236.

[35] M. Pompili et al. Realization of a multinode quantum network of remote solid-state qubits. Science, 2021. URL https://www.science.org/doi/10.1126/science.abg1919.

[36] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on thompson sampling. Foundations and Trends in Machine Learning, 11(1):1–96, 2018. doi: 10.1561/2200000070. URL https://arxiv.org/abs/1707.02038.

[37] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. arXiv preprint arXiv:1511.05952, 2015. URL https://arxiv.org/abs/1511.05952. Presented at ICLR 2016.

[38] C. Simon. Towards a global quantum network. Nature Photonics, 11:678–680, 2017. doi: 10.1038/s41566-017-0032-0. URL https://www.nature.com/articles/s41566-017-0032-0.

[39] M. A. L. Thathachar and P. S. Sastry. Networks of learning automata: Techniques for online stochastic optimization. Springer, 2011.

[40] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933. URL https://www.jstor.org/stable/2332286.

[41] B. Wang, T.-N. Hoang, H. Ahn, S. Muralidharan, L. Jiang, and Y.-A. Chen. Waiting time in quantum repeaters with probabilistic entanglement swapping. Physical Review A, 100(3):032322, 2019. doi: 10.1103/PhysRevA.100.032322.

[42] L. Wang, J. Bian, and J. Xu. Adaptive user-centric entanglement routing in quantum data networks. In 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS), pages 1202–1212. IEEE, 2024.

[43] X. Wang, M. Liu, X. Liu, Z. Li, M. Hajiesmaili, J. C. S. Lui, and D. Towsley. Learning best paths in quantum networks. In Proceedings of IEEE INFOCOM, 2025. URL https://arxiv.org/abs/2506.12462.

[44] S. Wehner, D. Elkouss, and R. Hanson. Quantum internet: A vision for the road ahead. Science, 362(6412), 2018. doi: 10.1126/science.aam9288.

[45] W. K. Wootters and W. H. Zurek. A single quantum cannot be cloned. Nature, 299(5886):802–803, 1982. doi: 10.1038/299802a0.

[46] X. Zhang and Z. Zhou. Neural thompson sampling. arXiv preprint, 2020. URL https://arxiv.org/abs/2010.00827.

[47] D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with UCB-based exploration. In Proceedings of the 37th International Conference on Machine Learning (ICML), volume 119 of Proceedings of Machine Learning Research, pages 11492–11502, 2020.

[48] D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with ucb-based exploration. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 11492–11502. PMLR, 2020. URL http://proceedings.mlr.press/v119/zhou20a/zhou20a.pdf.

[49] J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of Proceedings of Machine Learning Research, pages 467–475. PMLR, 2019. URL https://proceedings.mlr.press/v89/zimmert19a.html.

[50] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert. "event-ready-detectors" bell experiment via entanglement swapping. Physical Review Letters, 71(26):4287–4290, 1993. doi: 10.1103/PhysRevLett.71.4287.