Future Generation Computer Systems 185 (2026) 108714

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Generative AI in the age of quantum computing: A taxonomy, architectural elements and future directions ☆

Siva Sai [a,b], *, Ishika Goyal [e], Vinay Chamola [c,d,1], Rajkumar Buyya [b,2]

[a] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[b] Quantum Cloud Computing and Distributed Systems (qCLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Australia
[c] Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani, Pilani Campus, Vidya Vihar, Pilani, Rajasthan 333031, India
[d] APPCAIR, India
[e] Department of Computer Science and Engineering, Birla Institute of Technology and Science, Pilani, Pilani Campus, Vidya Vihar, Pilani, Rajasthan 333031, India

A R T I C L E I N F O

Keywords:
Quantum computing
Generative AI
Quantum GANs
Quantum Boltzmann machines
Quantum autoencoders
Diffusion models
NISQ
Quantum machine learning
Quantum advantage
Quantum applications

A B S T R A C T

Generative AI has emerged as a transformative paradigm for diverse applications, yet the escalating scale of modern models exposes critical computational and memory bottlenecks in classical hardware. This paper investigates the intersection of quantum computing and generative artificial intelligence (QGAI) to address these limitations and scale modern generative models. As models grow to billions of parameters, classical systems face bottlenecks in memory, energy, and training efficiency, while quantum computing offers exponential representational benefits for high-dimensional data. The paper analyzes five core quantum generative architectures — Quantum Circuit Born Machines, Quantum Generative Adversarial Networks, Quantum Boltzmann Machines, Quantum Variational Autoencoders, and Quantum Diffusion models, highlighting their design principles, learning mechanisms, and applications. QGAI models have demonstrated significant promise in domains such as drug discovery, human–machine interaction, IoT security, and financial modeling. Despite these advances, QGAI remains constrained by qubit noise, barren plateaus, and integration challenges. We conclude by identifying ten open research challenges and propose directions for achieving scalable, interpretable, and energy-efficient quantum generative learning.

1. Introduction

Generative artificial intelligence (GAI) has become a cornerstone of modern machine learning and creative automation. From Generative Adversarial Networks (GANs) that synthesize photorealistic imagery through Variational Autoencoders(VAEs) that learn disentangled latent spaces [1] to the recent success of diffusion–probabilistic models that outperform GANs on image quality benchmarks [2], the field has advanced at a remarkable pace. State-of-the-art language and multimodal models such as ChatGPT and DeepSeek now scale to hundreds of billions of parameters, enabling a few-shot reasoning and open-domain dialogue [3,4]. Yet this rapid scaling carries steep computational costs: training a single large-language model can emit large amounts of CO_2 and consume a lot of electricity [5]. Even with bespoke accelerators, classical hardware struggles to accommodate the memory footprint and wall-clock time required to fit such massive networks [6], motivating the search for novel computing paradigms.

Quantum computing (QC) offers a complementary path forward. Landmark demonstrations of quantum supremacy suggest that, for carefully chosen circuits, near-term quantum devices already outperform the largest classical supercomputers on sampling problems [7]. Although today's hardware remains noisy, small, and error-prone, quantum algorithms tailored to this Noisy Intermediate-Scale Quantum (NISQ) regime show promise in optimization, chemistry, and crucially probabilistic modeling [8]. In particular, parameterized quantum circuits (PQCs) can express broad families of probability distributions with far fewer tunable parameters than their classical analogues [9], making them intriguing candidates for enhancing GAI.

☆ This article is part of a Special issue entitled: 'Adv. Quantum Vol. IV' published in Future Generation Computer Systems.

* Corresponding author.
E-mail addresses: siva.sai@nus.edu.sg (S. Sai), f20220549@pilani.bits-pilani.ac.in (I. Goyal), vinay.chamola@pilani.bits-pilani.ac.in (V. Chamola), rbuyya@unimelb.edu.au (R. Buyya).
[1] Senior Member, IEEE
[2] Fellow, ACM

https://doi.org/10.1016/j.future.2026.108714
Received 11 October 2025; Received in revised form 8 July 2026; Accepted 13 July 2026
Available online 17 July 2026

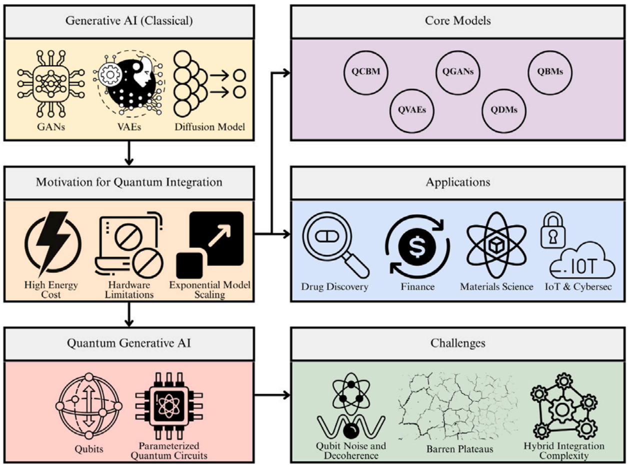

Fig. 1. Landscape of Quantum Generative Artificial Intelligence (QGAI).

Table 1
List of acronyms.

Acronym	Definition
GAI	Generative Artificial Intelligence
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
NLP	Natural Language Processing
NISQ	Noisy Intermediate-Scale Quantum
PQC	Parameterized Quantum Circuit
QC	Quantum Computing
QGAI	Quantum Generative Artificial Intelligence
QGAN	Quantum Generative Adversarial Network
QCBM	Quantum Circuit Born Machine
QBM	Quantum Boltzmann Machine
QVAE	Quantum Variational Autoencoder
QDM	Quantum Diffusion Model
QAE	Quantum Autoencoder
KL	Kullback–Leibler (divergence)
MMD	Maximum Mean Discrepancy
Q-ELBO	Quantum Evidence Lower Bound
HMI	Human–Machine Interaction
VQA	Variational Quantum Algorithm
VQE	Variational Quantum Eigensolver
QAOA	Quantum Approximate Optimization Algorithm

Over the past five years, a vibrant research sub-field, Quantum Generative AI (QGAI), has emerged at this intersection. Fig. 1 shows the current landscape of QGAI. Canonical examples include Quantum Circuit Born Machines (QCBMs), which sample from quantum Born probabilities to approximate classical data distributions with provably high expressive power [9]; Quantum GANs (QGANs), whose generators and discriminators may each be quantum, classical, or hybrid [10]; and Quantum Boltzmann Machines (QBMs), which leverage transverse-field Hamiltonians to model Gibbs distributions more richly than classical Restricted BMs [11]. More recent work extends diffusion–probabilistic modeling to quantum states [12,13].

Despite encouraging progress, formidable challenges remain. Barren-plateau phenomena, where gradients vanish exponentially with circuit width or depth, can stall optimization long before convergence [14]. Data-embedding overhead and shot noise limit throughput in hybrid workflows, while qubit decoherence imposes strict depth budgets on variational circuits. Moreover, integrating quantum accelerators with classical pre- and post-processing pipelines raises architectural, scheduling, and energy-efficiency questions that are still largely unresolved [8]. A systematic treatment of these issues is essential as the community transitions from proof-of-concept demonstrations to real-world deployments in drug discovery, finance, materials science, and human–machine interaction. Table 2 summarizes the major differences between classical and quantum GAI.

Quantum computers work in a totally different way from classical ones. Instead of bits that are either 0 or 1, quantum bits (qubits) can be in a superposition of both, and they can get entangled with each other in ways that classical bits cannot. This lets quantum systems represent and process high-dimensional probability distributions much more efficiently, with resources that grow only polynomially (or even less) as more qubits are added [15]. For generative modeling, this is of great importance because it means quantum computers can explore huge solution spaces and pick up on correlations that classical computers would struggle to even represent [14]. There are already quantum algorithms like amplitude estimation, QGANs, and variational circuits that show real promise in speeding up sampling, optimization, and inference tasks that are otherwise painfully slow on classical machines [16].

The benefits of quantum computing for GAI go beyond speed. Quantum models can use entanglement to create richer, more expressive latent spaces, which helps generate outputs that are not just diverse but also more coherent and contextually relevant [17]. Techniques like quantum neural networks, quantum Boltzmann machines, and quantum natural gradients are starting to tackle some of the classic problems in deep learning, like vanishing gradients and getting stuck in local minima. This means models can learn more stably and generalize better, which is something everyone in the field is after [14,17]. Another important advantage is energy efficiency. Training classical generative models consumes a lot of power, which is becomes a real concern for scaling up and keeping things sustainable . Quantum systems, with their parallelism and the ability to run certain computations with much less depth, could offer a much more energy-efficient way to do the same work. This matters a lot for fields like healthcare, finance, and scientific research, where big generative models could make a huge impact but are often held back by the sheer computational and energy costs [14,17].

The rest of the paper is organized as follows: Section 2 presents preliminaries on Generative Artificial Intelligence and Quantum Computing, providing the necessary foundations. Section 3 surveys quantum-enhanced generative architectures, including Quantum Circuit Born Machines, Quantum Generative Adversarial Networks, Quantum Boltzmann Machines, Quantum Variational Autoencoders, and Quantum Diffusion Models. Section 4 highlights applications of QGAI across diverse domains such as human–machine interaction, drug discovery, IoT security, resource allocation, and finance. Section 5 discusses key challenges and proposed solutions, covering scalability, error mitigation, training stability, data encoding, energy efficiency, interpretability, hybrid integration, and benchmarking. Section 6 identifies future research directions, including algorithmic innovations, NISQ feasibility, hybrid orchestration, and benchmarking for quantum advantage. Finally, Section 7 concludes the paper by summarizing the main insights and emphasizing the future potential of QGAI in advancing next-generation intelligent systems. Table 1 lists all the major acronyms used throughout this paper.

2. Background

This section provides a foundational overview of the two key pillars that underpin the Quantum Generative AI landscape: Generative Artificial Intelligence and Quantum Computing. Understanding their fundamental concepts, capabilities, and limitations is essential before diving into QGAI-specific models, architectures, and applications.

2.1. Generative artificial intelligence

Generative artificial intelligence (GAI) refers to a class of powerful models that learn to produce realistic and novel data, such as images, text, and audio. The rapid growth of GAI is driven by advancements in deep learning, massive datasets, and increased computational power, leading to widespread applications across various fields, including creative arts, scientific research, and commerce. This section explores the core principles of GAI, its primary model families, and the key limitations that continue to challenge its development and deployment.

Table 2
Classical vs. Quantum generative AI.

Aspect	Classical GAI	Quantum GAI
Training cost	Large models (e.g., GPT-3) require trillion-token datasets, high compute and energy	Parameterized quantum circuits can reduce sample and compute requirements
Sampling latency	Transformer and diffusion models incur large inference delays	Quantum circuits output samples directly via the Born rule
Expressivity	Limited in encoding entangled or highly non-local dependencies	Exponential encoding and entanglement capture complex correlations
Hardware constraints	Memory, wall-clock time, CO_2 emissions	NISQ devices: noisy, small qubit counts, decoherence constraints

Probabilistic foundations:
At the heart of GAI lies the goal of modeling a data-generating distribution $q(\mathbf{x})$ over a high-dimensional data space $\mathcal{X}$, such that a model can generate novel, high-quality samples $\tilde{\mathbf{x}} \sim p_\theta(\mathbf{x})$ that resemble real data. To accomplish this, GAI models typically define a parameterized distribution $p_\theta(\mathbf{x})$ and train it by minimizing some statistical divergence, such as the Kullback–Leibler (KL) divergence, Jensen–Shannon (JS) divergence, or Wasserstein distance, between the real and model distributions. This is generally done through stochastic gradient descent, often using surrogate loss functions suited to the model class [18].

Model families:
Several generative model paradigms have emerged as dominant approaches, as summarized in Fig. 2. Generative adversarial networks (GANs) train a generator G_θ and discriminator D_ϕ in a two-player minimax game, reaching equilibrium when real and synthetic samples are indistinguishable [18]. Variational autoencoders (VAEs) use probabilistic encoders and decoders to optimize an evidence lower bound, balancing reconstruction accuracy with latent-space regularization [19]. Diffusion models simulate a forward noising process and a learned reverse denoising process, often yielding state-of-the-art sample quality in vision and audio [2,20]. Transformer-based models (e.g., GPT) employ attention mechanisms to model sequences autoregressively or bidirectionally, enabling strong results in language, code, and captioning [21]. Finally, reinforcement learning for generative AI, including RLHF, aligns model outputs with preferences or downstream utility through policy optimization [22].

Classical limitations:
Despite the success of generative AI, three constraints persist. First, training cost remains high: large models (for example, GPT-3) are trained on trillion-token corpora and require substantial compute and energy. Second, sampling latency is significant because transformer and diffusion families are sequential or iterative at inference. Third, expressivity is limited for entangled or highly non-local dependencies that arise in quantum physics, genomics, and financial systems [23]. These constraints motivate quantum computing as a candidate for next-generation generative systems [24].

2.2. Quantum Computing (QC)

Quantum computing leverages quantum-mechanical principles to perform computation. Unlike classical bits, qubits can exist in superpositions of 0 and 1, and entanglement allows correlations that do not admit classical descriptions, enabling potential advantages on specific problem classes.

Fundamentals of quantum information:
Quantum information is carried by qubits, which are two-level systems represented as $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ with $|\alpha|^2 + |\beta|^2 = 1$. A register of n qubits inhabits a 2^n-dimensional Hilbert space, providing exponential representational capacity. Computation proceeds via unitary gates that evolve the state reversibly, and measurement projects the state to classical outcomes according to the Born rule.

Quantum algorithms:
Several algorithms illustrate potential advantages. Shor's algorithm factors large integers in polynomial time, challenging RSA security and delivering an exponential speedup over the best known classical methods [25]. Grover's algorithm searches unstructured databases in $O(\sqrt{N})$ queries, yielding a quadratic speedup [26]. Variational quantum algorithms (VQAs) target the NISQ regime by combining parameterized quantum circuits with classical optimisers; prominent examples include the Variational Quantum Eigensolver (VQE) and the Quantum Approximate Optimization Algorithm (QAOA) [8].

Hardware capabilities and constraints:
Leading architectures for QC include superconducting qubits (IBM, Google) [27], trapped ions (IonQ) [28], photonic circuits (Xanadu) [29], and neutral atoms (QuEra) [30]. Practical systems, however, face limitations in qubit count and quality, with error rates above fault-tolerant thresholds; coherence times and gate fidelities restrict feasible circuit depth; and the classical–quantum interface introduces latency that necessitates efficient co-processing strategies.

Why QC for GAI?:
Two principles are especially relevant for generative modeling. Exponential encoding allows an n-qubit state to represent a distribution over 2^n outcomes, supporting compact modeling of complex manifolds [31]. In addition, sampling arises natively: quantum circuits emit outcomes according to their internal probability amplitudes, avoiding explicit softmax layers or temperature tuning and enabling direct generation from the modeled distribution [32].

Table 3 presents a comparative overview of key Quantum Generative Models (QGMs). It highlights the differences in their architectures, training objective functions, and the nature of their generated outputs. This provides a structured understanding of how various QGMs approach generative learning in quantum settings. The performance and trainability of any QGAI model are governed by four interdependent pipeline components: the data encoding scheme, the circuit depth and entanglement topology, the hybrid training loop design, and the error mitigation strategy. The choice of encoding scheme determines both the qubit footprint and the expressive capacity of the quantum layer. Amplitude encoding compresses 2^n data dimensions into n qubits, offering exponential storage efficiency, but requires deep state-preparation circuits and exhibits sensitivity to small input perturbations [9]. Angle encoding maps each real-valued feature to a single-qubit rotation gate, producing shallow circuits compatible with NISQ hardware at the cost of a linear qubit-to-feature ratio, while basis encoding is the most hardware-efficient option but restricts the model to binary inputs. Circuit depth and entanglement topology mediate the fundamental tension between expressivity and trainability. Shallow circuits with nearest-neighbor entanglement are compatible with current error rates but capture only local correlations, whereas all-to-all entanglement allows richer feature maps yet simultaneously accelerates decoherence and induces barren plateaus whose severity grows exponentially with depth [14,33]. The matrix-product-state architecture [34] represents a practical middle ground, extending effective expressivity through bond-dimension virtual qubits

Table 3
Comparison of key quantum generative models.

Model	Architecture	Objective function	Target output
Quantum GAN (QGAN)	PQC + Classical discriminator	Minimax (Wasserstein/JS)	Classical or quantum data
Quantum VAE (QVAE)	Quantum encoder + Decoder	ELBO (Fidelity + Regularization)	Quantum state or distribution
Quantum Diffusion (QDM)	PQC with temporal unfolding	Fidelity loss (Reverse Process)	Noisy or pure states
Quantum Circuit Born Machine (QCBM)	PQC only	MMD or KL divergence	Discrete distribution

without proportionally increasing physical qubit requirements. The choice of classical optimiser in the hybrid training loop profoundly affects convergence behavior. Gradient-based methods with parameter-shift rule estimators [35] converge rapidly on smooth loss landscapes but are susceptible to barren plateaus in deep or randomly initialized circuits, while gradient-free methods including CMA-ES, SPSA, and particle-swarm optimization are more noise-tolerant but scale poorly with parameter count [36,37]. Error mitigation is the fourth lever available to practitioners. Zero-noise extrapolation and probabilistic error cancellation [38] reduce the effective noise rate at the cost of additional circuit executions, which for generative models already requiring many shots per training step can be prohibitive, while noise-resilient ansätze offer a complementary approach that avoids shot overhead at the cost of reduced expressivity [8]. The appropriate choice across all four components depends on the target hardware's specific error model, the available shot budget, and the distributional structure of the target task, and should be determined empirically rather than assumed from general principles.

3. Architectures

Quantum-enhanced generative AI (GAI) architectures aim to leverage quantum computing principles to improve model efficiency, scalability, and representational power (see Fig. 3).

3.1. Quantum circuit Born machines

Quantum Circuit Born Machines (QCBMs) [32] are unsupervised quantum generative models that leverage parameterized quantum circuits (PQCs) to approximate the probability distribution of a classical dataset. By producing tunable discrete distributions p_θ, a QCBM seeks to match a given target distribution q, and has proven effective on both synthetic and real-world data. The key advantage is the high expressivity of quantum circuits, which allows them to capture complex correlations that are difficult for classical models to represent [9]. Fig. 4 represents the training mechanism of QCBMs. A typical QCBM consists of alternating layers of single-qubit rotation gates (e.g., R_X, R_Y, R_Z) and multi-qubit entangling gates (e.g., CNOT, CZ). Shallow circuits with local connectivity can model simple distributions, whereas deeper, all-to-all entanglement can encode far richer correlations. Training proceeds by minimizing a divergence metric, often Kullback–Leibler (KL) divergence or maximum mean discrepancy (MMD), between samples from p_θ and the target distribution. Gradients (or gradient-free updates) are used by optimizers such as Adam, CMA-ES, SPSA, or particle-swarm optimization, with the choice dictated by noise tolerance and problem scale [32]. Benedetti et al. [9] first formalized the QCBM framework and demonstrated it on trapped-ion hardware, establishing early training protocols.

Optimization strategies have also diversified. Particle-swarm optimization and CMA-ES showed early promise on noisy devices [37], while Bayesian optimization excelled on high-parameter landscapes [39]. Riofrío et al. [40] provided a systematic comparison of QCBMs and QGANs, showing that shallow QCBMs struggle with complex correlations whereas deeper circuits improve fidelity at the cost of optimization difficulty, highlighting a fundamental depth–trainability trade-off. Finally, Zhou et al. [34] used a matrix-product-state (MPS) architecture to compress QCBMs for NISQ hardware. Their MPS-QCBM reproduced 2×2 Bars-and-Stripes data using only two or three physical qubits (with bond-dimension "virtual" qubits), matching conventional QCBM accuracy while reducing hardware requirements,at the expense of a larger parameter space that can slow optimization for bigger datasets.

3.2. Quantum generative adversarial networks

Quantum Generative Adversarial Networks (QGANs), first proposed by Lloyd and Weedbrook [10], transpose the classical GAN paradigm into the quantum domain. A parameterized quantum circuit (PQC) plays the role of the generator, emitting quantum states whose measurement outcomes define a synthetic data distribution, while a discriminator implemented either classically or quantum mechanically learns to tell these samples from those drawn from the target distribution. Training proceeds as a minimax optimization: the generator minimizes, and the discriminator maximizes, a distance between the two distributions. Because the generator lives in Hilbert space, QGAN training can exploit quantum-native routines such as the variational-quantum-eigensolver, enabling a more efficient search of high-dimensional probability landscapes than is possible with purely classical stochastic-gradient updates. Fig. 5 shows the schematic working of QGANs. Lloyd and Weedbrook outlined four architectural flavors, determined by whether the task data and the two agents are classical (C) or quantum (Q): QT-QGQD (quantum task, quantum generator, quantum discriminator), QT-CGQD, QT-CGCD, and CT-QGCD [10]. The fully quantum option offers the greatest expressive power but is the most hardware intensive; the hybrid variants trade quantum advantage for near-term feasibility.

Riofrío et al. [40] systematically benchmarked QGANs against Quantum Circuit Born Machines (QCBMs), showing that QGAN generators can approximate classically intractable distributions yet suffer more severely from mode collapse and training instabilities on noisy hardware. To mitigate those issues, several design improvements have been proposed:

(1) Tensor-network generators — Huggins et al. [41] replaced the PQC with matrix-product or tree tensor networks, enabling classical pre-training before transfer to quantum hardware and reducing circuit depth.

(2) Patch-based QGANs — Huang et al. [42] introduced a quantum patch-GAN that divides data into smaller patches, each handled by a shallow sub-generator, making image generation possible with limited qubits.

(3) Quantum Wasserstein losses — Chakrabarti et al. [43] adapted the Wasserstein distance to quantum settings to stabilize training. Kiani et al. [44] refined the metric to remove unitary-invariance artifacts and boost state fidelity.

(4) Efficient gradients — The parameter-shift rule, generalized to noisy hardware by Schuld et al. [35], provides unbiased gradient estimates without ancillary qubits and is now standard for PQC optimization.

3.3. Quantum Boltzmann machines

Quantum Boltzmann Machines (QBMs) generalize classical Boltzmann machines by replacing Ising spins with qubits governed by a transverse-field Hamiltonian. First proposed by Amin et al. [11], a QBM defines a quantum Boltzmann distribution whose Gibbs state can encode richer correlations than any classical counterpart. Visible qubits represent data, hidden qubits supply latent capacity, and trainable couplings appear as longitudinal and transverse fields. Learning proceeds by minimizing a distance, often quantum relative entropy between model and data distributions, but non-commuting Hamiltonian terms make gradients expensive. Practical training, therefore,

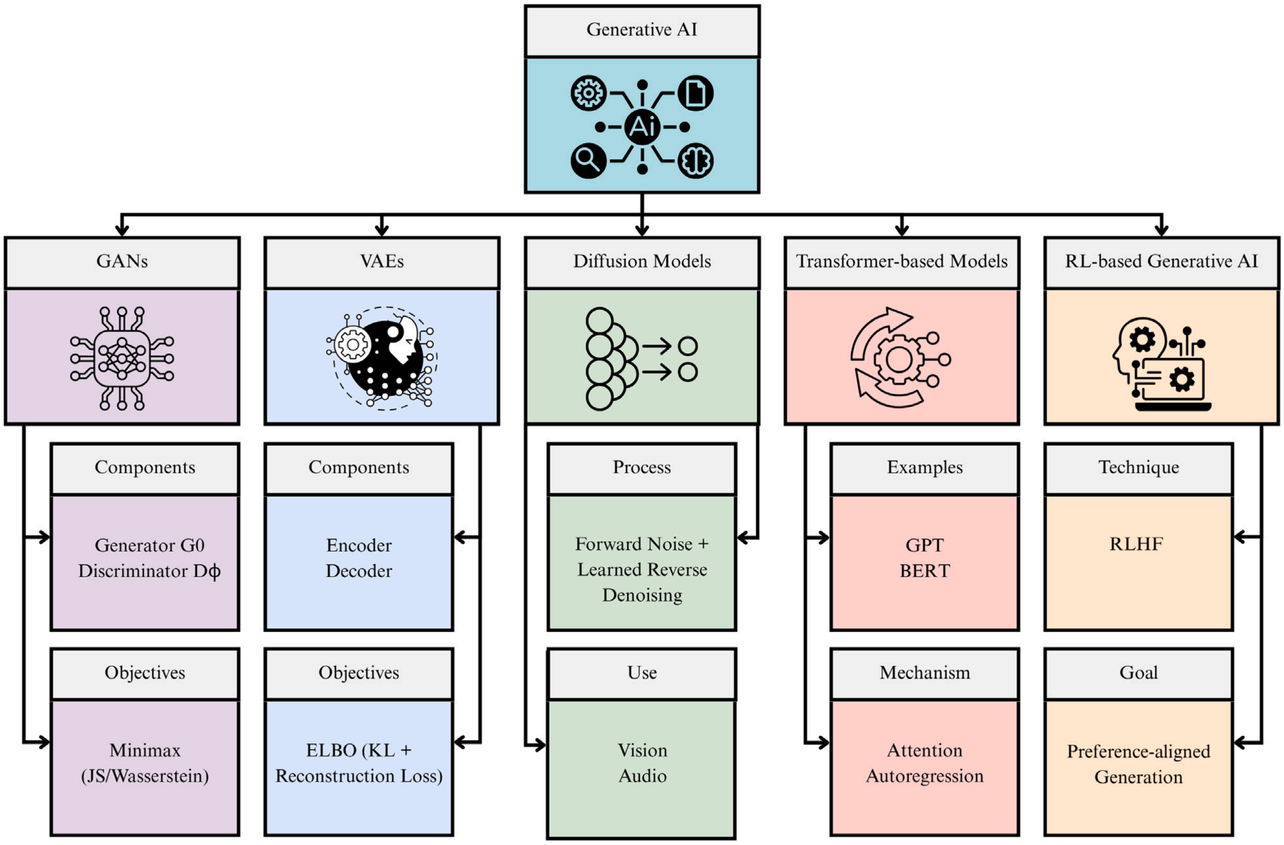

Fig. 2. Taxonomy of generative AI models.

uses approximations such as bound-based objectives or sampling on quantum annealers, where QBMs enjoy a natural hardware realization.

Different types of QBMs are listed as follows:

Semi-restricted QBMs: Amin et al. [11] showed that forbidding hidden–hidden couplings simplifies training while retaining quantum advantage. With visible units clamped during optimization, the semi-restricted QBM converged faster than fully connected RBMs and captured higher-order correlations.

State-based training for tomography: Kieferová and Wiebe [45] introduced state-based training, where a QBM is fitted directly to measurement data from an unknown quantum state — circumventing full-state comparison and enabling scalable quantum-state tomography on near-term devices.

Reinforcement-learning: Crawford et al. [46] embedded QBMs in policy-evaluation loops, demonstrating higher success rates than classical RBMs on standard reinforcement-learning benchmarks by exploiting quantum-induced correlations.

Variational QBMs: Zoufal et al. [47] re-cast the QBM as a parameterized quantum circuit that prepares Gibbs states variationally. Exact gradients obtained via the parameter-shift rule reduce sampling overhead and improve scalability to larger lattices—useful, for instance, in quantum-chemistry Hamiltonians.

Evolved QBMs: Minervini et al. [48] proposed evolved QBMs (EQBMs), which apply a second Hamiltonian $H(\phi)$ to evolve the thermal state of a base Hamiltonian $G(\theta)$. The resulting state $\omega(\theta,\phi) = e^{-iH(\phi)}\rho(\theta)e^{iH(\phi)}$ combines imaginary and real-time dynamics, enlarging the model family while still admitting analytic gradients and natural-gradient updates driven by quantum Fisher information.

Sample-efficient training: Coopmans and Benedetti [49] reframed QBM learning as matching expectation values of Hamiltonian terms, proving that stochastic gradient descent converges in polynomial time and samples when those expectations are estimated with shadow tomography. Pre-training via mean-field or Gaussian-Fermionic models further mitigates barren plateaus.

3.4. Quantum variational autoencoders

Quantum Variational Autoencoders (QVAEs), proposed by Khoshaman et al. [50], extend the classical variational-autoencoder paradigm by inserting parameterized quantum circuits (PQCs) into the encoder and decoder. The quantum encoder maps classical or quantum input data into a low-dimensional latent space, and the decoder reconstructs the original input from this compressed representation. Training minimizes a quantum evidence lower bound (Q-ELBO) that balances reconstruction fidelity (via, e.g., state-fidelity metrics) against latent-space regularization to a chosen prior. By sidestepping a full comparison with exponentially large target states, QVAEs are well suited to near-term hardware: the PQCs provide compact dimensionality reduction while preserving essential data features [50].

Fig. 6 shows QVAE architecture along with classical VAE architecture. The classical VAE architecture includes a decoder and an encoder which outputs Gaussian parameters for approximating the latent space distribution, from which random noises are sampled for generating new data through the decoder network (the generation or sampling process is denoted in red box and arrows); (b) a patched variational quantum circuit (each patch has three qubits) serves as the quantum encoder/decoder network with corresponding quantum input, hidden, and output layers. A hybrid VAE/AE is realized by connecting the final fully connected (FC) layer with the preceding measurement outcome vector.

Different types of QVAEs are listed as follows:

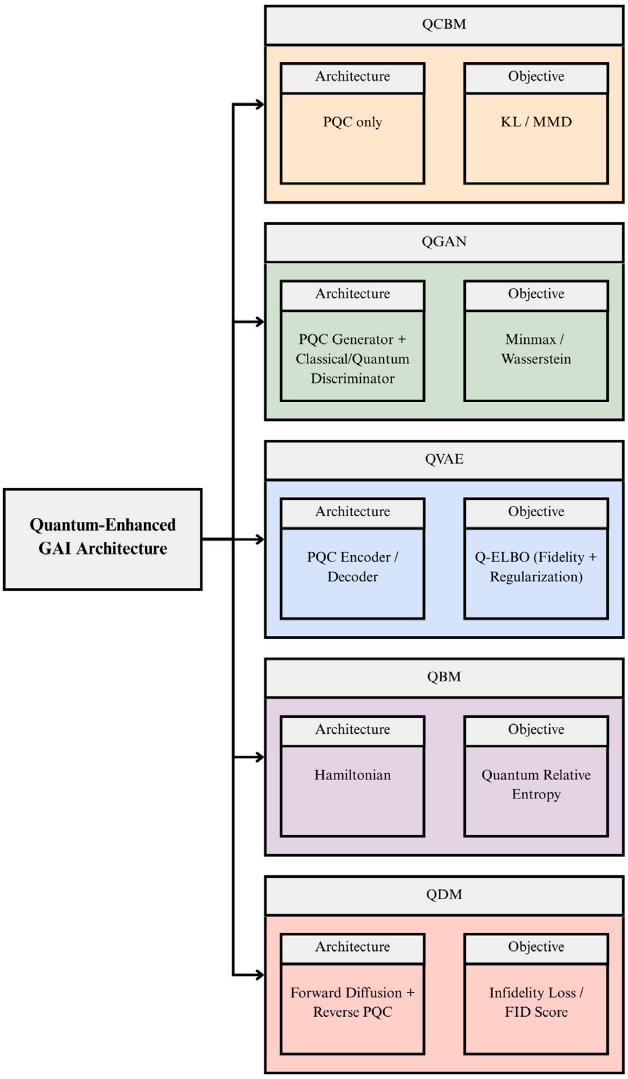

Fig. 3. Taxonomy of quantum generative architectures.

EF-QAE: Bravo-Prieto et al. [51] introduced the Enhanced-Feature Quantum Autoencoder (EF-QAE), which replaces global cost functions with local ones to avoid vanishing gradients during training . Encoding each data component into a single-qubit rotation and optimizing local fidelities yielded high-accuracy compression of both quantum many-body states and classical images.

Hybrid QVAE: Srikumar et al. [52] introduced a hybrid quantum autoencoder (HQA) that extracts information from quantum states by mapping them into a classical representational space. The approach enables clustering and semi-supervised classification of pure states, demonstrated on amplitude-encoded states, with potential extensions to more complex quantum data.

Noise-assisted and adiabatic QAEs: Cao and Wang [53] introduced a noise-assisted quantum autoencoder that improves recovery fidelity by aligning input and output mixedness using noise channels determined from measurement data. The approach utilizes measurement information more effectively, extends to both circuit and adiabatic models (implementable on quantum annealers), and is validated on thermal states of the transverse-field Ising model and Werner states. A projected quantum autoencoder is also proposed for pure-state ensemble compression.

ζ-QVAE: Quantum computing faces challenges in handling large real-world datasets due to limited hardware resources. To address this, the authors in [54] proposed ζ-QVAE, a fully quantum variational autoencoder that compresses both classical and quantum data into low-dimensional latent representations while enabling effective reconstruction. Unlike prior quantum compression models, ζ-QVAE employs regularized mixed-state latent spaces, supports multiple divergences, and leverages full density matrices for efficient optimization. This framework mirrors the capabilities of classical VAEs while remaining fully quantum, with applications demonstrated on genomics and synthetic data, showing improved latent space utilization and competitive or superior performance compared to classical models.

3.5. Quantum diffusion models

Quantum diffusion models (QDMs) translate classical diffusion–probabilistic frameworks into the quantum domain by replacing neural networks with parameterized quantum circuits (PQCs). Cacioppo et al. [55] introduced the first such model: a classical forward process adds Gaussian noise, while a quantum backward process iteratively denoises via a PQC, eliminating repeated state-encoding overhead and keeping the workflow NISQ-friendly. Amplitude encoding maps classical data (e.g., images) into quantum states, and the PQC then minimizes an infidelity loss between denoised and target states.

Following are the different variants of QDMs:

Latent and fully quantum QDMs: Cacioppo et al. [55] explored both a fully quantum generator and a latent variant that first compresses data with a classical autoencoder before quantum denoising. The latent model runs on today's NISQ hardware and supports conditional generation by encoding labels in auxiliary qubits. MNIST experiments produced recognizable digits, though the current Fréchet Inception Distance (FID) still lags behind classical diffusion models.

QuDDPM: Zhang et al. [12] proposed Quantum Denoising Diffusion Probabilistic Models, using quantum scrambling in the forward direction and a PQC-based reverse process trained with hybrid (classical and quantum) gradients. They achieved $>99\%$ state-generation fidelity for eight-qubit mixed states and reported polynomial sample complexity, outperforming QGAN baselines.

QGDM and RQGDM: Chen et al. [13] introduced the Quantum Generative Diffusion Model, whose non-unitary forward step uses depolarizing channels to reach a maximally mixed state. A trainable, timestep-embedded PQC then reconstructs the target, yielding 99.9% fidelity on pure-state tasks and 99.3% on eight-qubit mixed states, again surpassing QGANs. A resource-efficient variant (RQGDM) reduces auxiliary-qubit demands without sacrificing accuracy.

Q-Dense and QU-Net: Kölle et al. [56] introduced two quantum diffusion architectures, Q-Dense and QU-Net, designed to enhance image generation with quantum machine learning. Q-Dense employs amplitude embedding within dense quantum circuits, while QU-Net serves as a quantum analogue of the classical U-Net, incorporating quantum convolutional layers. To accelerate generation, the authors proposed a unitary single-sampling circuit that compresses the traditional multi-step reverse diffusion process into a single operation. Benchmarking on datasets such as MNIST, Fashion MNIST, and CIFAR-10 showed that these quantum models achieved superior performance compared to classical counterparts with a similar number of parameters, as measured by metrics like FID, SSIM, and PSNR. These results demonstrate the potential of QDMs to address efficiency and parameter-scaling challenges in generative modeling.

Fig. 7 illustrates the QU-Net architecture that integrates quantum convolutions. The model leverages down blocks, up blocks, and skip connections to process input slices, embedding them into a dense quantum circuit. This design enables hierarchical feature extraction and

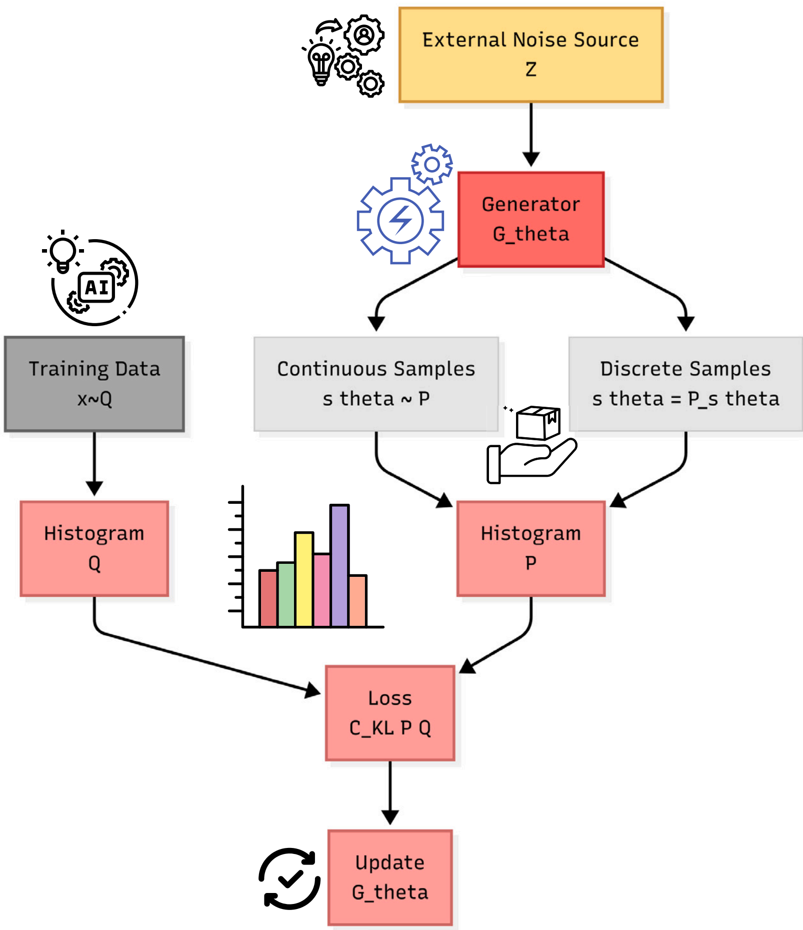

Fig. 4. Schematic of training of QCBM.

efficient reconstruction, akin to a quantum analogue of the classical U-Net. Table 4 categorizes the primary studies reviewed in this survey by the quantum computing framework used and by whether experiments were conducted via classical simulation only, on real quantum hardware, or both. This serves as a practical reference for researchers seeking to replicate or extend the cited works. For simulated studies, the simulator and system size are noted; for hardware implementations, the device and qubit count are provided. Table 5 summarizes all the above-mentioned quantum architectures.

Selecting among the five quantum generative architectures requires weighing expressivity, trainability, hardware compatibility, and the structure of the target application. Quantum Circuit Born Machines are best suited when the objective is to approximate a discrete or structured classical distribution [9,40]. Quantum Generative Adversarial Networks are used for tasks that demand high sample diversity, such as synthetic attack-trace generation for IoT security [57] or de novo molecular design [58]. Their chief drawback is training instability: mode collapse and vanishing discriminator gradients [40]. Quantum Boltzmann Machines are uniquely advantageous when the data-generating process is itself quantum-mechanical, as in correlated-electron systems in quantum chemistry, or when the model must run natively on an annealer [11]. Quantum Variational Autoencoders are recommended when the primary goal is dimensionality reduction or latent-space learning on mixed classical–quantum data under severe circuit-depth constraints [50]. Quantum Diffusion Models currently achieve the highest fidelity scores in quantum state generation, exceeding 99% on eight-qubit tasks [12,13], and are the natural choice when output quality rather than training speed is paramount. Their iterative reverse process is more stable than GAN adversarial training, but the multi-step denoising loop incurs substantial shot overhead on NISQ hardware [56].

4. Applications

The integration of quantum computing and generative AI represents a significant advancement, establishing a new frontier for addressing problems that were previously intractable for classical systems. This emergent QGAI field leverages the fundamental principles of quantum mechanics, such as superposition and entanglement, to model and generate complex data with unprecedented efficiency. QGAI demonstrates its potential across a range of disciplines, from enhancing human–machine interaction to accelerating the discovery of new therapeutic compounds. This section reviews some of the most significant applications of QGAI, illustrating how these innovative models begin to address some of the most formidable challenges in science and technology. Before surveying specific application domains, it is important to qualify the nature of the performance claims that follow. The figures reported in this section are drawn from studies conducted under controlled simulation environments or small-scale NISQ hardware experiments. Unless explicitly stated otherwise, assertions of quantum

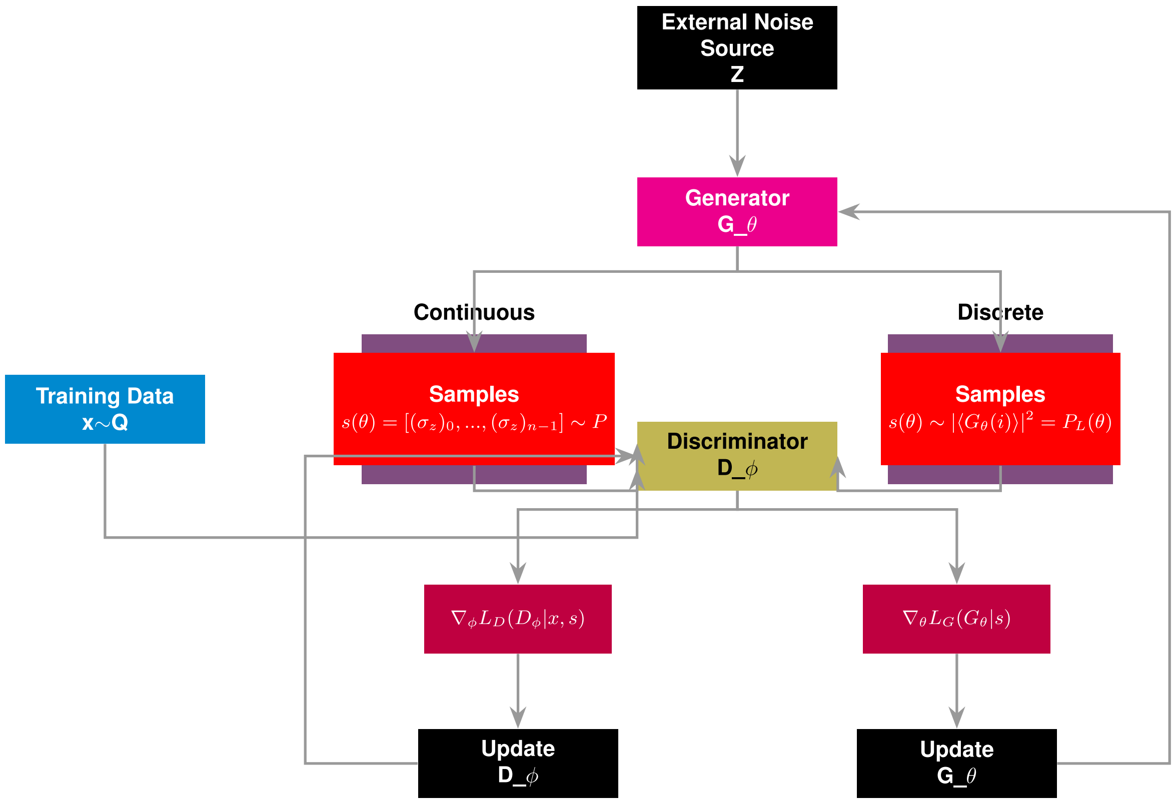

Fig. 5. Schematic of training of QGAN.

Table 4
Quantum frameworks, execution environments, and hardware platforms used in selected QGAI studies reviewed in this survey. Sim. = classical simulation only; HW = real quantum hardware; Both = simulation and hardware both reported. '–' indicates not explicitly stated in the source study. Only entries verified against primary sources are included.

Study	Architecture	Framework	Mode	Simulator/Device	Qubits
Zhu et al. [39]	QCBM	Custom PQC	Both	Trapped-ion device (UMD)	4
Lloyd & Weedbrook [10]	QGAN	Theoretical	Sim.	No hardware; theoretical	–
Huang et al. [42]	QGAN	Custom	HW	Superconducting (USTC)	12
Amin et al. [11]	QBM	Exact diagonalization	Sim.	Classical numerical sim.	10
Khoshaman et al. [50]	QVAE	Quantum Monte Carlo	Sim.	QMC simulation	–
Zhang et al. [12]	QDM	Qiskit	Sim.	Qiskit Aer statevector	8
Kölle et al. [56]	QDM	PennyLane	Sim.	Default qubit simulator	4–8
Li & Ghosh [60]	QVAE	PennyLane	Sim.	Classical statevector	3 per patch
Prajapat et al. [61]	QGAN+crypto	Custom	Sim.	Classical sim.	–
Romero et al. [62]	QAE	Custom	Sim.	Classical statevector	4

advantage over classical counterparts should be interpreted as potential or theoretical advantages rather than verified empirical superiority. Rigorous, large-scale benchmarking of QGAI models against state-of-the-art classical generative architectures, including denoising diffusion probabilistic models, large-scale GANs, and transformer-based systems, on production-grade datasets and under identical hardware conditions, remains an open research challenge for the field [14,40,59]. Where the surveyed literature provides concrete comparative metrics such as fidelity, FID score, parameter counts, or runtime, these are reported explicitly.

4.1. Human–machine interaction

Human–machine interaction (HMI) studies intuitive interfaces and collaborative workflows that let people and intelligent systems communicate, learn, and execute tasks seamlessly [63]. Classical AI, especially natural-language processing (NLP) and decision-making, underpins most current HMI frameworks, yet struggles with high-dimensional data, real-time adaptability, and cross-cultural language barriers [64]. For example, NLP engines often misinterpret dialectal variations or cultural nuances, reducing effectiveness in multilingual or rapidly changing settings [65]. Adaptive HMI must also balance clarity with complexity: poorly designed interfaces can overload users or cause miscommunication during shared tasks [66].

Quantum Generative AI (QGAI) can alleviate these issues by combining techniques like superposition and entanglement with powerful generative models. Quantum circuits process multimodal data (text, speech, images) in parallel, enabling QGAI systems to learn user preferences and close feedback loops in real time. QGAI also mitigates NLP shortcomings by generating culturally nuanced translations and disambiguating intent through quantum-optimized context analysis [67]. These advances make QGAI a promising foundation for next-generation HMI that demands robust cross-cultural interoperability and real-time adaptivity. Lee et al. [68] introduced a Quantum Computational Intelligence (QCI) agent with a content attention ontology model to improve human–machine interaction in Taiwanese/English language learning. It integrates a Generative AI image agent, SBERT (Sentence Bidirectional Encoder Representations from Transformers)-based similarity

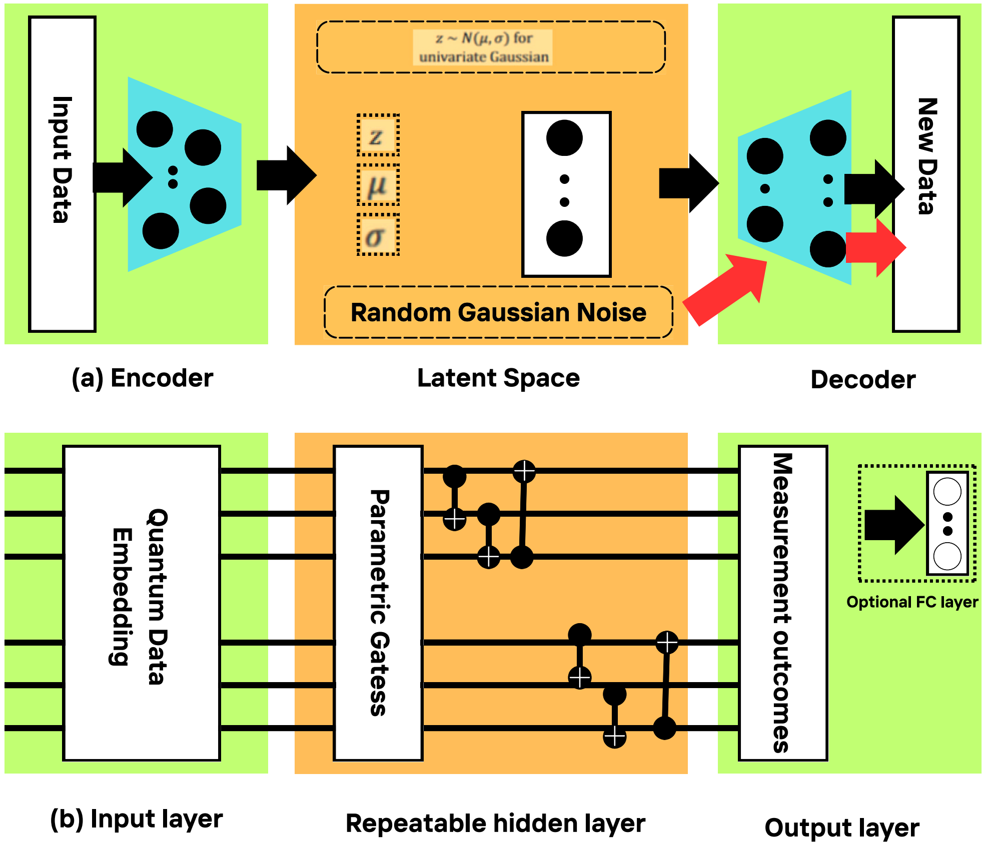

Fig. 6. QVAE architecture.

Table 5
Summary of quantum-enhanced GAI architectures.

Architecture	Introduction	Variants/Key contributions
Quantum Circuit Born Machines (QCBMs)	Unsupervised models using parameterized quantum circuits (PQCs) to approximate target distributions, leveraging high expressivity of quantum circuits to capture complex correlations [9,32].	Early work: Benedetti et al. [9], Zhu et al. [39], Liu and Wang [36]. Optimizers: particle-swarm, CMA-ES, Bayesian [37,39]. Riofrío et al. compared QCBMs vs QGANs [40]. Zhou et al.: MPS-QCBM for NISQ devices [34].
Quantum Generative Adversarial Networks (QGANs)	Quantum adaptation of GANs with PQC generator and classical/quantum discriminator; training as minimax optimization in Hilbert space [10].	Lloyd & Weedbrook: four variants (QT-QGQD, QT-CGQD, QT-CGCD, CT-QGCD) [10]. Benchmarked vs QCBMs (Riofrío et al.) [40]. Improvements: tensor-network generators [41], patch-based QGANs [42], quantum Wasserstein losses [43,44], parameter-shift gradients [35].
Quantum Boltzmann Machines (QBMs)	Generalize classical BMs with qubits and transverse-field Hamiltonians; encode Gibbs states with richer correlations; training via relative entropy or annealers [11].	Semi-restricted QBMs [11]. State-based training for tomography [45]. RL applications [46]. Variational QBMs [47]. Evolved QBMs (EQBMs) [48]. Sample-efficient training guarantees [49].
Quantum Variational Autoencoders (QVAEs)	Extend VAEs by embedding PQCs in encoder/decoder; trained with quantum ELBO for dimensionality reduction while preserving key features [50].	EF-QAE [51]. Hybrid QVAE (HQA) [52]. Noise-assisted/adiabatic QAEs [53]. ζ-QVAE for fully quantum latent spaces [53].
Quantum Diffusion Models (QDMs)	Translate diffusion–probabilistic models into quantum domain with PQC-based denoising; enable quantum-native reverse processes [55].	Latent vs fully quantum QDMs [55]. QuDDPM [12]. QGDM & RQGDM [13]. Q-Dense & QU-Net for images [56].

scoring, and Meta AI's Universal Speech Translator to support bilingual text–image generation and dialogue experiences. Using Particle Swarm Optimization (PSO), the system predicts learner performance, while the enhanced NUTN.TW-LLM (National University of Tainan Taiwan-Large Language Model) enables interactive and adaptive learning. Experimental results showed the model's efficacy, with accuracy for social media data analysis improving to 71% and a specific language learning model achieving 90% accuracy. Human inspectors also gave the GAI image agent high grades for its generated images. Overall, the findings suggested that students were interested in co-learning languages and QCI with machines.

4.2. Drug discovery

Drug discovery seeks and optimizes compounds for therapeutic use, typically via virtual screening, molecular-dynamics simulations, and laboratory validation [69,70]. The search space, however, overwhelms classical pipelines, which are expensive and sometimes inaccurate at quantum-level phenomena. Although classical machine learning accelerates screening, it can falter when modeling high-dimensional molecular systems or predicting electronic structures and reaction pathways [71].

Quantum Generative AI tackles these hurdles by encoding molecules (graphs or SMILES) directly into quantum states. QGANs and Quantum Variational Autoencoders (QVAEs) explore chemical space via parameterized quantum circuits, efficiently sampling regions that would be classically intractable [50,72]. These models have already pushed drug-likeness metrics such as Quantitative Estimate of Drug-likeness (QED) and partition coefficient (logP) beyond classical baselines [73].

Li et al. [60] proposed a Scalable Quantum Variational Autoencoder (SQ-VAE) and its vanilla variant (SQ-AE) for drug molecule design, combining quantum and classical architectures. By applying strategies like adjustable quantum depth and patched circuits, the models improve reconstruction and sampling of drug molecules, showing advantages in low-dimensional cases and better drug properties in high-dimensional learning compared to classical counterparts. The results showed that for low-dimensional datasets, the quantum models achieved computational advantages in training speed and representation quality. For high-dimensional molecules, the quantum generative autoencoders generated new molecules with better drug properties. For example, a scalable quantum VAE with an LSD of 18 generated molecules with a logP score of 0.780, outperforming a classical VAE's 0.357. Another variant with a LSD of 56 achieved a QED score of 0.204, compared to the classical VAE's 0.139.

Jain and Ganguly [58] introduced hybrid Quantum GAN architectures (QWGAN-HG, QGAN-HG-GP, and QWGAN-HG-GP) for molecular simulation and drug discovery. These models used a hybrid generator that combined a parameterized quantum circuit with a classical neural network to explore the QM9 dataset. The models were evaluated using statistical distance metrics and various drug property scores to assess their efficacy. The results showed that the proposed QWGAN-GP-HG-P4-L2 architecture performed best, achieving a Drug Candidate Score of 0.204 and a Validity Score of 22.469, which were significantly higher than the classical MolGAN's scores of 0.118 and 0.125, respectively.

4.3. IoT security

IoT security aims to protect connected devices, networks, and data in domains such as healthcare, smart-city infrastructure, and industrial automation [74]. These environments face unauthorized access, data breaches, and device spoofing because they are decentralized, resource-constrained, and often rely on legacy public-key protocols (e.g. RSA, ECC) that are susceptible to quantum attacks [75]. Classical countermeasures struggle to scale, and quantum computers threaten to break many standard cryptosystems by accelerating integer factorization and related problems [75,76].

Prajapat et al. [61] proposed a novel security protocol for AIoT-based healthcare systems by integrating Generative AI (GAI), blockchain, and quantum encryption. The protocol employs encoding, challenge–response authentication with hash/digital signatures, and dynamic key exchange to ensure secure data transmission. Experimental results show high performance with 99.4% accuracy, 99.10% precision, 98.66% recall, and 98.50% F1-score, highlighting its effectiveness in safeguarding patient data and reducing carbon-intensive inefficiencies.

In an indirectly related work [77], the authors proposed an unsupervised anomaly detection method for high-energy physics using a quantum generative adversarial network (QGAN) trained on Standard Model (SM) data. The methodology involved training a QGAN on a limited dataset of Standard Model events and then using it to detect anomalous Higgs and Graviton events. For the quantum approach, the researchers randomly selected 100 events from each dataset, which was ten times fewer than the 1000 events used for the classical benchmark. The experimental results indicated that the quantum model achieved comparable classification accuracy while using significantly less training data. For instance, on the 7-feature dataset, the QGAN achieved a Graviton detection AUC of 0.96 and a Higgs detection AUC of 0.92, which were comparable to or better than the classical model's respective AUCs of 0.96 and 0.87. The authors also found that the quantum model exhibited an improved expressive power, measured by a higher effective dimension.

4.4. Resource allocation

Quantum-computing networks must juggle qubit decoherence, fluctuating entanglement fidelity and heterogeneous service demands. Static allocation wastes resources, whereas naive dynamic policies overlook quantum-specific constraints such as variable qubit lifetimes or purification overheads. QGAI tackles the problem by marrying quantum generative models with reinforcement learning (RL). Quantum GANs and QVAEs embed high-dimensional resource states like numbers of free qubits, entangled pairs, and buffer times into parameterized circuits that predict near-optimal allocation under noise and uncertainty. They can decide, for example, when to devote extra qubits to error correction or how to route tasks across edge cloud nodes to minimize latency, all while exploiting correlations inaccessible to classical optimisers.

Xu et al. [78] proposed a multi-agent reinforcement learning (MARL) framework for decentralized resource allocation in heterogeneous quantum computing networks. By modeling each node as an agent in a partially observable environment, the method optimizes task offloading for quantum generative learning applications, such as quantum autoencoders. Experiments show that the MARL approach converges to optimal policies, reducing system costs by at least 80% compared to random allocation under various quality-of-service settings.

4.5. Miscellaneous

Finance: Quantum Circuit Born Machines (QCBMs) can model high-dimensional financial distributions such as joint asset returns and copulas with exponentially fewer parameters than classical approaches. Borras et al. [79] compared Quantum GANs and QCBMs on cryptocurrency data[3] (BNBBTC, ETHBTC, LTCBTC, NEOBTC, QTUMETH), equilibrium was achieved when generator and discriminator losses converged to

$$L_G = L_D = -\log\left(\frac{1}{2}\right) \approx 0.6931,$$

and QGANs required fewer training iterations than classical baselines, indicating a quantum advantage in efficiency. For the QCBM, COBYLA outperformed SPSA and Nelder–Mead in minimizing distribution discrepancies.

[3] https://www.tradingview.com/symbols/BNBBTC/.

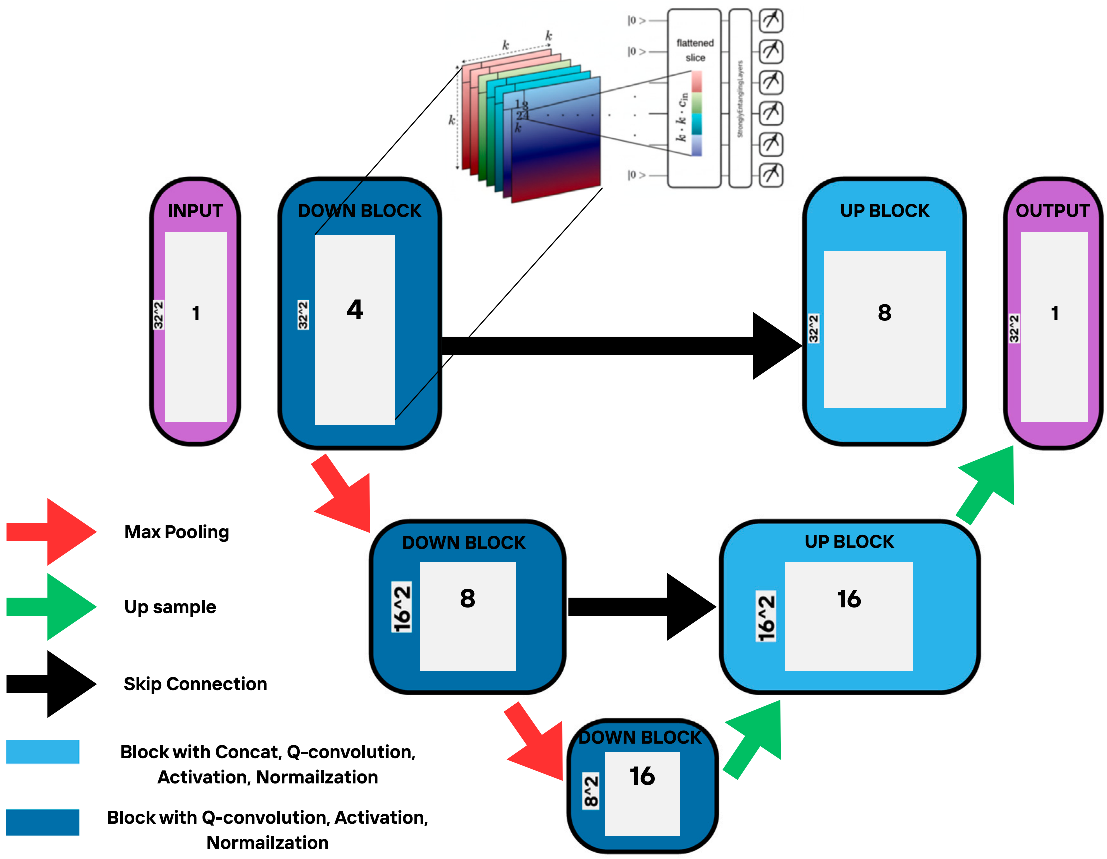

Fig. 7. QU-Net architecture and quantum convolution, embedding a flattened slice into a dense quantum circuit.

Quantum-state compression: Quantum autoencoders (QAEs) mitigate hardware noise by encoding states into lower-dimensional, error-resilient subspaces. Romero et al. [62] introduced the concept of a quantum autoencoder to compress quantum data, a task for which classical compression algorithms cannot be employed. The methodology involved training a programmable quantum circuit using classical optimization algorithms to find a unitary transformation that preserved quantum information through a smaller latent space. The cost function was defined as the average fidelity between the "trash state" and a reference state, which was measured via a SWAP test in a hybrid quantum–classical scheme. The model was applied to compress the ground states of the Hubbard model and molecular Hamiltonians, including the hydrogen molecule. The results showed that both circuits tested were able to achieve high fidelities. For example, when compressing the hydrogen molecule's wavefunction from 4 to 2 qubits, Circuit A achieved a fidelity error of 6.99 on the testing set, while Circuit B had an error of 6.07.

Steinbrecher et al. [80] introduced Quantum Optical Neural Networks (QONNs), mapping classical neural network concepts to quantum optical systems for tasks like quantum state compression, reinforcement learning, black-box quantum simulation, and one-way quantum repeaters. Using numerical simulations accelerated with Python, Cython, and GPU libraries, QONNs were trained via gradient-free optimization (BOBYQA) and evolutionary strategies, demonstrating strong generalization from limited training data. In a reinforcement learning task, a QRAM-encoded QONN achieved a mean fitness of 136.1 after 1000 generations, which was higher than both a direct-encoded QONN (61.9) and a classical network (37.1). The QONN also successfully compressed molecular hydrogen ground states, with iterative and global optimizations converging to fidelities of 90.0% and 92.2%, respectively. A QONN trained to implement a one-way quantum repeater reached numerical precision after 50 layers.

Quantum chemistry: Semi-restricted Quantum Boltzmann Machines (QBMs) have been trained on transverse-field Ising Hamiltonians to approximate correlated-electron ground states. Amin et al. [11] trained Quantum Boltzmann Machines (QBMs) with nondiagonal Hamiltonians using a lower bound on log-likelihood to enable sampling-based gradient estimation. The methodology involved training QBMs with a transverse field Ising Hamiltonian, which introduced quantum effects, and comparing the results to classical Boltzmann machines (BMs) on small systems. The experimental results demonstrated that QBMs consistently learned data distributions better than classical BMs, as measured by lower Kullback–Leibler (KL) divergence values. For a fully visible model with 10 qubits, the classical BM converged to a KL divergence of approximately 0.62, while the QBM achieved a significantly lower value of about 0.42. Table 6 summarizes all the applications of Quantum GAI models.

4.6. Case studies: End-to-end QGAI pipelines

To make the application survey more tangible, this subsection presents two concrete end-to-end pipeline descriptions showing how a QGAI model integrates into practical systems. The first case concerns QGAN-based anomaly detection in IoT security. Raw network-flow features such as packet inter-arrival times, byte counts, and protocol flags are first standardized and reduced via PCA to at most eight dimensions, matching the qubit budget of near-term devices. These reduced features are angle-encoded into single-qubit rotation gates, keeping state-preparation depth at $O(d)$ and within current noise budgets [9]. A PQC generator with two to three entangling layers is trained on normal IoT traffic to learn the benign distribution, while a classical discriminator learns to distinguish real normal-traffic samples from generator output following the hybrid CT-QGCD variant of Lloyd and Weedbrook [10]; quantum Wasserstein loss is used to improve robustness and scalability of adversarial training [43]. At inference, the discriminator score serves as an anomaly metric and traffic vectors below a validation-tuned threshold are flagged as anomalous. On current NISQ devices this pipeline is feasible only for $d \leq 8$ features and circuit depths of six to ten layers; shot budgets of 10^3–10^4 per inference step introduce latency of order seconds on cloud-accessed QPUs [8], making the approach suitable for batch analytics rather than real-time alerting. The second case concerns QVAE-based molecular design for drug discovery. Drug candidates are represented as SMILES strings, converted to Morgan molecular fingerprints, and compressed to eight to eighteen dimensions by a classical autoencoder front-end [60]. A QVAE with patched three-qubit circuits then encodes the compressed fingerprints into a quantum latent space, trained to minimize the Q-ELBO balancing reconstruction fidelity against latent-space regularization [50]. Novel candidates are generated by sampling from the quantum latent space and ranked using QED and logP scores [73]. Performance figures for Li and Ghosh [60] are reported in Section 4.2. The primary hardware constraint is that patched circuit designs are necessary to keep sub-circuit depths within coherence limits, and the classical decoder absorbs most reconstruction complexity on present-generation devices.

5. Challenges and solutions

While the previous sections explored the promise and early applications of quantum generative AI, the path toward its widespread adoption faces significant challenges. These obstacles stem from the fundamental limitations of near-term quantum hardware, including issues with scalability, noise, and training stability. The hybrid nature of QGAI also introduces major bottlenecks related to data encoding and the efficient integration of quantum and classical systems. This section examines these core challenges and looks at future research directions that aim to overcome them.

5.1. Scalability and hardware

Despite promising early results, quantum generative AI (QGAI) remains fundamentally constrained by the physical limitations of near-term quantum hardware. Current noisy intermediate-scale quantum (NISQ) devices typically provide only 50–100 usable qubits, many of which are prone to decoherence, cross-talk, and gate errors [7,8]. This restricts the feasible depth and width of parameterized quantum circuits (PQCs), making it difficult to scale quantum generative models for high-dimensional data. Furthermore, existing qubit topologies often impose connectivity constraints that hinder entanglement-efficient circuit designs. As a result, many QGAI frameworks rely on shallow circuits or hybrid quantum–classical workflows to remain viable. To enable large-scale generative modeling, significant advances in qubit fidelity, interconnect density, and error rates are essential. Emerging hardware platforms such as silicon spin qubits, photonic processors, and topological qubits offer promising avenues to transcend current limitations, though they are still in early stages of development [81].

5.2. Noise and error

Quantum noise remains one of the most critical barriers to effective training and deployment of QGAI models. Decoherence, gate infidelity, and measurement noise not only reduce output fidelity but also disrupt gradient estimation during optimization, which is central to variational training techniques like QGANs and QVAEs. While error-correcting codes are the gold standard in fault-tolerant quantum computing, their overhead remains prohibitive for most current devices. Consequently, variational error mitigation (VEM), zero-noise extrapolation, and probabilistic error cancellation are emerging as more practical approaches for near-term generative tasks [8,38]. Tailoring these techniques specifically for generative architectures, such as denoising diffusion models or adversarial networks, requires careful trade-offs between circuit expressivity and robustness. Moreover, advances in error-aware training algorithms, noise-resilient ansätze, and active feedback from classical co-processors may play a pivotal role in stabilizing training dynamics under real-world noise conditions.

5.3. Training stability and barren plateaus

A major impediment to training quantum generative models is the presence of *barren plateaus*, regions in the optimization landscape where gradients vanish exponentially with circuit depth or system size [14]. This issue severely limits the scalability of parameterized quantum circuits (PQCs), especially in high-dimensional generative tasks. Barren plateaus often arise in randomly initialized deep circuits or when cost functions lack locality. Overcoming them requires tailored ansätze with structured entanglement, layer-wise training strategies, or symmetries that preserve gradient signal. Recent research explores locality-aware cost functions, parameter initialization schemes, and adaptive learning-rate schedules to improve convergence [33]. In the context of QGAI, these innovations are particularly crucial since unstable training may degrade sample quality and reduce model fidelity. Exploring quantum-native optimization techniques such as quantum natural gradients or training via quantum feedback loops may further stabilize learning dynamics in generative settings.

5.4. Data encoding and readout bottlenecks

Bridging the classical–quantum interface is another pressing challenge. Efficiently encoding classical datasets, such as images, molecules, or financial time series, into quantum states is computationally expensive and often circuit-depth intensive. While techniques such as amplitude encoding, basis encoding, and angle encoding are commonly used, each has trade-offs in qubit requirements, expressivity, and hardware overhead [9]. Moreover, measurement readout from quantum circuits is inherently probabilistic and requires repeated sampling to estimate expectation values or reconstruct output distributions, which introduces latency and noise. These bottlenecks become particularly limiting in QGAI pipelines where large volumes of training samples are required. Emerging strategies such as quantum data compression, classical shadows, and low-depth feature maps aim to reduce input/output complexity. Innovations in hybrid encoders and classical pre-processing can also help shift some of the computational burden off quantum hardware [82].

5.5. Resource requirements and energy consumption

Although quantum algorithms offer potential speedups and energy efficiencies for specific generative tasks, their practical deployment is hampered by the significant resource overhead associated with quantum control systems. Operating qubits at cryogenic temperatures (e.g., below 15 mK) demands elaborate cooling infrastructures, and real-time classical control circuits introduce non-negligible power and synchronization costs. Furthermore, error correction, while essential for fault-tolerant operation, imposes large qubit overheads that dilute energy savings. For instance, implementing a single logical qubit may require hundreds or thousands of physical qubits in surface codes. As quantum workloads scale, the total energy footprint may rival or exceed that of high-performance classical GPU clusters unless more efficient architectures are developed. Designing sustainable, resource-aware quantum–classical systems with optimized thermodynamic and architectural efficiency is thus vital for the widespread adoption of QGAI solutions [82].

Table 6
Applications of QGAI models.

Domain	Description
Human–machine interaction	QGAI enhances NLP and multimodal interaction, enabling real-time adaptivity and culturally nuanced translations. Example: Quantum Computational Intelligence (QCI) agent integrating ontology, GAI, SBERT similarity, and speech translation for bilingual education [68].
Drug discovery	QGANs and QVAEs encode molecules into quantum states to explore chemical space. Hybrid quantum–classical GANs improve drug-likeness metrics (QED, logP, validity, uniqueness) compared to classical MolGAN [58,60].
IoT security	QGANs enable anomaly detection with fewer samples than classical GANs [66]. A GAI–blockchain–quantum encryption protocol achieves 99.4% accuracy, 99.1% precision, 98.66% recall, 98.5% F1-score, and 99.2% security in AIoT healthcare [61].
Resource allocation	QGANs and QVAEs with reinforcement learning optimize decentralized resource allocation in quantum networks. Multi-agent RL reduces system costs by at least 80% compared to random allocation [78].
Miscellaneous	Finance: QCBMs simulate stock-market copulas with 98% fidelity; QGANs model cryptocurrency data more efficiently [39,79]. Quantum autoencoders compress quantum states for noise resilience [62]. QBMs approximate correlated-electron ground states in quantum chemistry [11].

5.6. Model interpretability and explainability

As Quantum Generative AI expands from proof-of-concept demos to mission-critical deployments [83] in finance, medicine, and national security, black-box behavior becomes unacceptable. Classical explainability toolkits, saliency maps, SHAP values, counterfactuals, do not translate directly to the Hilbert space, where information is encoded as complex amplitudes rather than activations. Recent work proposes quantum feature-attribution techniques that combine classical shadows with integrated gradients to reveal which input dimensions dominate a parameterized quantum circuit's output probability mass. Other approaches leverage circuit cutting and local tomography to visualize entanglement flow or to extract low-rank latent representations that correlate with human-interpretable factors [17]. Developing lightweight, hardware-friendly diagnostics that scale beyond a few dozen qubits is therefore pivotal if QGAI is to earn stakeholder trust in high-stakes settings.

5.7. Integration with classical AI systems

Most near-term deployments will remain hybrid: a quantum co-processor handles the generative core, while classical hardware pre-processes data and post-processes measurement outcomes. Unfortunately, data marshaling between the host CPU/GPU and the QPU incurs latencies that can outweigh quantum acceleration gains. State-of-the-art orchestration frameworks such as Qiskit Runtime, Amazon Braket Hybrid Jobs, and PennyLane Lightning offer asynchronous job queues, batched parameter updates, and just-in-time compilation, yet scheduling, memory management, and co-optimization heuristics remain immature. Future research must focus on zero-copy data paths, streaming encoders that amortize I/O cost over many shots, and compiler passes that fuse classical preprocessing with quantum kernels to minimize turnaround time and energy footprint.

5.8. Benchmarking and standardization

Progress in QGAI is hampered by a patchwork of bespoke datasets, loss metrics, and reporting conventions. Directly porting ImageNet or CIFAR to quantum form is unrealistic, while small synthetic benchmarks mask scalability issues. The community therefore needs consensus suites that span discrete, continuous, and quantum-native data, along with fidelity, diversity, robustness, and energy metrics [40]. Initiatives such as QMLBench[4] for classification are nascent, but broader participation is required to ensure fair cross-model comparison, reproducibility, and transparent reporting of qubit counts, circuit depths, and hardware noise parameters.

[4] https://github.com/XanaduAI/qml-benchmarks.

5.9. Quantum advantage and real-world applications

Although small-scale demonstrations of quantum supremacy have appeared in sampling tasks [7,84], a *practical* quantum advantage in generative modeling remains unproven. Establishing such an edge requires carefully chosen application niches where (i) the target distribution is classically intractable, (ii) the quantum circuit depth remains within NISQ limits, and (iii) the end-to-end latency, including data transfer, outperforms highly optimized GPU pipelines. Near-term candidates include molecular conformer generation for lead optimization in drug discovery, Monte-Carlo copula simulation for tail-risk estimation in finance, and rapid lattice sampling for catalyst design in materials science [59]. Benchmarks on trapped-ion and photonic hardware already show order-of-magnitude speedups in small molecular subspaces [85]. Collaborative efforts between domain specialists, quantum hardware vendors, and cloud-service providers are therefore essential to deliver compelling real-world showcases of quantum-enabled generative AI.

Table 7 summarizes the key challenges faced by QGAI models along with the corresponding proposed solutions.

6. Future directions

To further advance the field, we identified some future discussions focused on algorithmic innovation, hybrid orchestration, and trustworthy QGAI as discussed below:

6.1. Algorithmic innovations and new paradigms

Beyond hardware progress, sustained advances in quantum algorithms will be pivotal for overcoming present-day bottlenecks. QDMs extend classical score-based and denoising diffusion frameworks by replacing neural score networks with parameterized quantum circuits that denoise in Hilbert space, promising exponentially richer latent dynamics [12,13]. Meanwhile, quantum-aware reinforcement learning (QRL) integrates quantum policy networks and quantum environment simulators, allowing agents to exploit entanglement-enhanced exploration and potentially quadratic sample-complexity reductions. Early studies report faster convergence on combinatorial-optimization benchmarks and improved robustness to stochasticity [87]. Other frontiers include quantum meta-learning, quantum natural-gradient descent, and adaptive ansätze that co-evolve with hardware noise profiles [88]. Taken together, these innovations expand the design space of QGAI well beyond direct porting of classical architectures, paving the way for hybrid models that can harness the best of both quantum and classical worlds.

Table 7
Challenges and proposed solutions in Quantum Generative AI (QGAI).

Challenge	Description	Proposed Solutions/Approaches
Scalability and hardware limitations	NISQ devices are prone to decoherence, cross-talk, and gate errors. Limits circuit depth, width, and entanglement-efficient designs [8].	HQC(Hybrid Quantum Classical) systems address NISQ limitations by delegating the computationally difficult parts of a problem to the quantum computer while a classical computer handles the remaining tasks [86].
Noise and error mitigation	Decoherence, gate infidelity, and measurement noise reduce fidelity and disrupt gradient estimation during variational training [8].	Variational error mitigation, zero-noise extrapolation, probabilistic error cancellation, noise-resilient ansätze, and active feedback [38].
Training stability and barren plateaus	Gradients vanish exponentially with depth or system size, limiting PQC scalability in high-dimensional tasks [14].	Structured ansätze, layer-wise training, locality-aware cost functions, adaptive learning-rate schedules, and quantum natural gradients [33].
Data encoding and readout bottlenecks	Classical-to-quantum encoding (amplitude, basis, angle) is costly in qubits and depth. Readout requires repeated sampling, adding latency and noise [9].	Quantum data compression, classical shadows, low-depth feature maps, hybrid encoders, and classical pre-processing [82].
Resource requirements and energy consumption	Cryogenic cooling and control electronics add overhead; error correction requires hundreds of physical qubits per logical qubit. Energy demands may rival GPU clusters.	Sustainable, resource-aware quantum–classical systems; efficient hardware architectures to reduce thermodynamic cost [82].
Model interpretability and explainability	Black-box behavior in finance, medicine, and security is problematic; Hilbert-space amplitudes resist classical explainability tools.	Quantum feature attribution (classical shadows + gradients), circuit cutting, and local tomography for entanglement flow [17].
Integration with classical AI systems	Hybrid deployment requires data transfer between CPUs/GPUs and QPUs, incurring latency and overhead.	Orchestration frameworks (Qiskit Runtime, Braket Hybrid Jobs, PennyLane Lightning) with asynchronous queues and just-in-time compilation.
Benchmarking and standardization	Lack of standard datasets and metrics; synthetic benchmarks hide scalability issues.	Suites like QMLBench with fidelity, diversity, robustness, and energy metrics for fair comparison.
Quantum advantage and applications	Quantum advantage in generative modeling remains unproven; small-scale supremacy demos exist only in sampling tasks [7,84].	Domain-focused tasks (drug discovery, finance, materials science), benchmarking on trapped-ion and photonic hardware [59,85].
Algorithmic innovations and paradigms	Need for quantum-native algorithms beyond classical analogues.	QDMs [12,13], quantum RL with entanglement-enhanced exploration [87], meta-learning, and adaptive ansätze.

6.2. NISQ feasibility across application domains

A candid assessment of what current NISQ hardware can realistically support in each major application domain is essential for setting accurate expectations. In drug discovery, present devices with fifty to a hundred usable qubits can handle molecular fingerprints compressed to eight to eighteen dimensions, supporting proof-of-concept generation of small drug-like molecules in low-dimensional chemical subspaces [58,60]. In IoT security and cyberdefense, the feasibility window is more favorable because the relevant classification tasks operate on low-dimensional flow-level features rather than raw packet payloads [57,89]. QGAN-based anomaly detectors trained on eight to sixteen features are within current hardware reach, though shot-induced latency precludes real-time deployment on cloud QPUs; edge-deployed shallow circuits with classical pre-processing represent the more practical near-term architecture [90,91]. Across all domains, the consistent recommendation from the literature is that hybrid pre- and post-processing is not merely a workaround for hardware immaturity but a principled architectural choice that will remain relevant even as qubit counts and fidelities improve [8,86].

6.3. Hybrid orchestration, compiler co-design, and systems integration

A critical direction is the co-design of hybrid runtime stacks that minimize host–device latency while maximizing quantum utilization [92]. Promising avenues include batched parameter-shift execution, just-in-time transpilation with noise-aware gate selection, zero-copy pathways between CPU/GPU memory and QPU buffers, and streaming encoders that amortize state preparation over many shots. Compiler passes that fuse classical pre- and post-processing with quantum kernels can reduce I/O overhead and enable end-to-end auto-differentiation across heterogeneous boundaries. At the architectural level, topology-aware ansätze and placement strategies that respect device connectivity constraints can lower circuit depth, while schedule co-optimization across classical accelerators and QPUs targets wall-clock time, throughput, and energy per generated sample as joint objectives.

6.4. Error mitigation, verification, and trustworthy QGAI

Near-term deployment requires techniques that stabilize training and provide statistical guarantees about generated outputs under noise. Variational error-mitigation protocols, zero-noise extrapolation, and probabilistic error cancellation should be tailored to generative losses and sampling workflows. Complementary verification tools, such as shadow-based estimators for distributional properties, robust hypothesis tests for mode collapse, and noise-aware calibration loops, can quantify fidelity and diversity without prohibitive measurement budgets. Interpretability remains central: feature-attribution in Hilbert space, circuit-cutting visualizations of entanglement flow, and sensitivity analyses for data embeddings can help diagnose failure modes and build practitioner trust in safety-critical domains.

6.5. Data encoding, compression, and quantum-native corpora

The classical–quantum interface will benefit from learned and problem-specific embeddings that reduce state-preparation cost while retaining sufficient statistics for generation. Future work should explore variational data loaders, quantum-assisted tokenisation for sequences, and compression schemes that couple classical autoencoders with shallow quantum layers to trade qubits for classical latent structure. On the readout side, grouped measurements, shot-frugal estimators, and adaptive stopping rules can lower sampling requirements during training and evaluation. Community benchmarks comprising quantum-native datasets, along with standardized protocols for encoding images, molecules, and time-series, will enable fair comparisons across architectures and hardware backends.

6.6. Benchmarking, reproducibility, and evidence for advantage

To move beyond proof-of-concept demonstrations, the field needs principled benchmarks that report fidelity, diversity, robustness to distribution shift, energy per sample, and end-to-end wall-clock time on specified hardware. Reproducible evaluation suites with fixed encoders, circuit-depth budgets, and noise profiles can disentangle algorithmic progress from hardware drift. Application-driven targets, molecular conformer generation, tail-risk copula sampling, and materials lattice sampling, are well-suited for establishing compelling case studies of practical advantage when combined with transparency about qubit counts, connectivity, calibration data, and compiler settings. Public leaderboards and standard result cards should accompany releases to encourage rigorous, apples-to-apples assessment across QGAI systems.

7. Conclusions

Quantum Generative AI (QGAI) combines the expressive power of generative models with the computational advantages of quantum mechanics, offering a promising pathway beyond the limitations of classical AI. By harnessing principles like superposition and entanglement, QGAI architectures can represent and generate complex data distributions more efficiently than their classical counterparts. This paper has highlighted the core frameworks such as QCBMs, QVAEs, and QDMs and their applications in domains like drug discovery, IoT security, and human–machine interaction. While current progress is constrained by hardware noise, limited qubit counts, and algorithmic challenges like barren plateaus, the field is advancing rapidly. With continued improvements in quantum hardware and hybrid learning frameworks, QGAI is well-positioned to contribute significantly to future intelligent systems. Its potential to enable scalable, efficient, and privacy-preserving generative modeling makes it a compelling direction for ongoing research and development.

CRediT authorship contribution statement

Siva Sai: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization. **Ishika Goyal:** Writing – review & editing, Writing – original draft, Conceptualization. **Vinay Chamola:** Writing – review & editing, Supervision, Funding acquisition. **Rajkumar Buyya:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by CHANAKYA Fellowship Program of TIH Foundation for IoT & IoE (TIH-IoT) received by Dr. Vinay Chamola under Project Grant File CFP/2022/027 and the University of Melbourne stipend (supported by an ARC Discovery Project).

Data availability

No data was used for the research described in the article.

References

[1] Z. Ding, Y. Xu, W. Xu, G. Parmar, Y. Yang, M. Welling, Z. Tu, Guided variational autoencoder for disentanglement learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7920–7929.

[2] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[4] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, 2021, arXiv preprint arXiv:2108.07258.

[5] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training, 2021, arXiv:2104.10350.

[6] S. Sai, M. Prasad, G. Dashore, V. Chamola, B. Sikdar, On-Device generative AI: The need, architectures, and challenges, IEEE Consum. Electron. Mag. 14 (4) (2025) 21–32, http://dx.doi.org/10.1109/MCE.2024.3518761.

[7] F. Arute, K. Arya, R. Babbush, D. Bacon, J.C. Bardin, R. Barends, R. Biswas, S. Boixo, F.G. Brandao, D.A. Buell, et al., Quantum supremacy using a programmable superconducting processor, Nature 574 (7779) (2019) 505–510.

[8] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum 2 (2018) 79.

[9] M. Benedetti, E. Lloyd, S. Sack, M. Fiorentini, Parameterized quantum circuits as machine learning models, Quantum Sci. Technol. 4 (4) (2019) 043001.

[10] S. Lloyd, C. Weedbrook, Quantum generative adversarial learning, Phys. Rev. Lett. 121 (4) (2018) 040502.

[11] M.H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, R. Melko, Quantum boltzmann machine, Phys. Rev. X 8 (2) (2018) 021050.

[12] B. Zhang, P. Xu, X. Chen, Q. Zhuang, Generative quantum machine learning via denoising diffusion probabilistic models, Phys. Rev. Lett. 132 (10) (2024) 100602.

[13] C. Chen, Q. Zhao, M. Zhou, Z. He, Z. Sun, H. Situ, Quantum generative diffusion model: a fully quantum-mechanical model for generating quantum state ensemble, 2024, arXiv:2401.07039.

[14] M.S. Rudolph, S. Lerch, S. Thanasilp, O. Kiss, O. Shaya, S. Vallecorsa, M. Grossi, Z. Holmes, Trainability barriers and opportunities in quantum generative modeling, Npj Quantum Inf. 10 (1) (2024) 116.

[15] S. Chen, Are quantum computers really energy efficient?, 2023.

[16] Y. Zhao, H. Wang, K. Xu, Y. Wang, J. Zhu, F. Wang, Adaptive algorithm for quantum amplitude estimation, 2022, arXiv:2206.08449.

[17] X. Gao, E.R. Anschuetz, S.-T. Wang, J.I. Cirac, M.D. Lukin, Enhancing generative models via quantum correlations, Phys. Rev. X 12 (2) (2022) 021037.

[18] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[19] D.P. Kingma, M. Welling, Auto-encoding variational bayes, 2013, arXiv:1312.6114.

[20] S. Sai, U. Mittal, V. Chamola, DMDAT: Diffusion model-based data augmentation technique for Vision-Based accident detection in vehicular networks, IEEE Trans. Veh. Technol. 74 (2) (2025) 2241–2250, http://dx.doi.org/10.1109/TVT.2024.3491170.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.

[23] J. Schneider, Generative to Agentic AI: Survey, conceptualization, and challenges, 2025, arXiv:2504.18875.

[24] P. Singh, R. Dasgupta, A. Singh, H. Pandey, V. Hassija, V. Chamola, B. Sikdar, A survey on available tools and technologies enabling quantum computing, IEEE Access 12 (2024) 57974–57991, http://dx.doi.org/10.1109/ACCESS.2024.3390395.

[25] P.W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in: Proceedings 35th Annual Symposium on Foundations of Computer Science, Ieee, 1994, pp. 124–134.

[26] L.K. Grover, A fast quantum mechanical algorithm for database search, in: Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing, 1996, pp. 212–219.

[27] M. AbuGhanem, Superconducting quantum computers: who is leading the future? EPJ Quantum Technol. 12 (1) (2025) 102.

[28] C.D. Bruzewicz, J. Chiaverini, R. McConnell, J.M. Sage, Trapped-ion quantum computing: Progress and challenges, Appl. Phys. Rev. 6 (2) (2019).

[29] M. AbuGhanem, Photonic quantum computers, 2024, arXiv preprint arXiv:2409.08229.

[30] J. Wurtz, A.. Bylinskii, S.H. Cantu, F. Huber, A. Lukin, F. Liu, P. Weinberg, J. Long, et al., Aquila: Quera's 256-qubit neutral-atom quantum computer, 2023, arXiv:2306.11727.

[31] R.P. Feynman, Simulating physics with computers, in: Feynman and Computation, cRc Press, 2018, pp. 133–153.

[32] J.-G. Liu, L. Wang, Differentiable learning of quantum circuit born machines, Phys. Rev. A 98 (6) (2018) 062324.

[33] J.R. McClean, S. Boixo, V.N. Smelyanskiy, R. Babbush, H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. 9 (1) (2018) 4812.

[34] L.-H. Gong, L.-Z. Xiang, S.-H. Liu, N.-R. Zhou, Born machine model based on matrix product state quantum circuit, Phys. A 593 (2022) 126907.

[35] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A 99 (3) (2019) 032331.

[36] M.S. Daoud, M. Shehab, H.M. Al-Mimi, L. Abualigah, R.A. Zitar, M.K.Y. Shambour, Gradient-based optimizer (GBO): a review, theory, variants, and applications, Arch. Comput. Methods Eng. 30 (4) (2023) 2431–2449.

[37] X. Li, X. Yao, Cooperatively coevolving particle swarms for large scale optimization, IEEE Trans. Evol. Comput. 16 (2) (2011) 210–224.

[38] K. Temme, S. Bravyi, J.M. Gambetta, Error mitigation for short-depth quantum circuits, Phys. Rev. Lett. 119 (18) (2017) 180509.

[39] D. Zhu, N.M. Linke, M. Benedetti, K.A. Landsman, N.H. Nguyen, C.H. Alderete, A. Perdomo-Ortiz, N. Korda, A. Garfoot, C. Brecque, et al., Training of quantum circuits on a hybrid quantum computer, Sci. Adv. 5 (10) (2019) eaaw9918.

[40] C.A. Riofrio, O. Mitevski, C. Jones, F. Krellner, A. Vuckovic, J. Doetsch, J. Klepsch, T. Ehmer, A. Luckow, A characterization of quantum generative models, ACM Trans. Quantum Comput. 5 (2) (2024) 1–34.

[41] W. Huggins, P. Patil, B. Mitchell, K.B. Whaley, E.M. Stoudenmire, Towards quantum machine learning with tensor networks, Quantum Sci. Technol. 4 (2) (2019) 024001.

[42] H.-L. Huang, Y. Du, M. Gong, Y. Zhao, Y. Wu, C. Wang, S. Li, F. Liang, J. Lin, Y. Xu, et al., Experimental quantum generative adversarial networks for image generation, Phys. Rev. Appl. 16 (2) (2021) 024051.

[43] S. Chakrabarti, H. Yiming, T. Li, S. Feizi, X. Wu, Quantum Wasserstein generative adversarial networks, Adv. Neural Inf. Process. Syst. 32 (2019).

[44] B.T. Kiani, G. De Palma, M. Marvian, Z.-W. Liu, S. Lloyd, Learning quantum data with the quantum earth mover's distance, Quantum Sci. Technol. 7 (4) (2022) 045002.

[45] M. Kieferova, N. Wiebe, Tomography and generative data modeling via quantum Boltzmann training, 2016, arXiv:1612.05204.

[46] D. Crawford, A. Levit, N. Ghadermarzy, J.S. Oberoi, P. Ronagh, Reinforcement learning using quantum Boltzmann machines, 2016, arXiv:1612.05695.

[47] C. Zoufal, A. Lucchi, S. Woerner, Variational quantum Boltzmann machines, Quantum Mach. Intell. 3 (1) (2021) 7.

[48] M. Minervini, D. Patel, M.M. Wilde, Evolved quantum Boltzmann machines, Phys. Rev. A 113 (3) (2026) 032427.

[49] L. Coopmans, M. Benedetti, On the sample complexity of quantum Boltzmann machine learning, Commun. Phys. 7 (1) (2024) 274.

[50] A. Khoshaman, W. Vinci, B. Denis, E. Andriyash, H. Sadeghi, M.H. Amin, Quantum variational autoencoder, Quantum Sci. Technol. 4 (1) (2019) 014001.

[51] C. Bravo-Prieto, Quantum autoencoders with enhanced data encoding, Mach. Learn.: Sci. Technol. 2 (3) (2021) 035028.

[52] M. Srikumar, C.D. Hill, L.C. Hollenberg, Clustering and enhanced classification using a hybrid quantum autoencoder, Quantum Sci. Technol. 7 (1) (2022) 015020.

[53] C. Cao, X. Wang, Noise-assisted quantum autoencoder, Phys. Rev. Appl. 15 (5) (2021) 054012.

[54] G. Wang, J. Warrell, P.S. Emani, M. Gerstein, Quantum variational autoencoder utilizing regularized mixed-state latent representations, Phys. Rev. A 111 (4) (2025) 042416.

[55] A. Cacioppo, L. Colantonio, S. Bordoni, S. Giagu, Quantum diffusion models, 2023, arXiv:2311.15444.

[56] M. Kölle, G. Stenzel, J. Stein, S. Zielinski, B. Ommer, C. Linnhoff-Popien, Quantum denoising diffusion models, in: 2024 IEEE International Conference on Quantum Software, QSW, IEEE, 2024, pp. 88–98.

[57] J. Yang, E. Rezvani, M.A. Khan, B. Heidari, H.M. Albarakati, S. Prajapat, M. Abdel-Salam, L.Y. Por, QB-AutoIDS: A blockchain-based decentralized autonomous cyberattack detection system for consumer electronics using hybrid double Q-Learning and Bi-LSTM, IEEE Trans. Consum. Electron. 71 (4) (2025) 11487–11496, http://dx.doi.org/10.1109/TCE.2025.3600429.

[58] P. Jain, S. Ganguly, Hybrid quantum generative adversarial networks for molecular simulation and drug discovery, 2022, arXiv:2212.07826.

[59] M. Hibat-Allah, M. Mauri, J. Carrasquilla, A. Perdomo-Ortiz, A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models, Commun. Phys. 7 (1) (2024) 68.

[60] J. Li, S. Ghosh, Scalable variational quantum circuits for autoencoder-based drug discovery, in: 2022 Design, Automation & Test in Europe Conference & Exhibition, DATE, IEEE, 2022, pp. 340–345.

[61] S. Prajapat, P. Kumar, A.K. Das, G. Muhammad, Generative AI-Enabled quantum encryption algorithm for securing IoT-Based healthcare application using blockchain, IEEE Internet Things J. 12 (13) (2025) 24541–24551, http://dx.doi.org/10.1109/JIOT.2025.3555159.

[62] J. Romero, J.P. Olson, A. Aspuru-Guzik, Quantum autoencoders for efficient compression of quantum data, Quantum Sci. Technol. 2 (4) (2017) 045001.

[63] V. Dakulagi, K.H. Yeap, H. Nisar, R. Dakulagi, G. Basavaraj, M.V. Galindo, An overview of techniques and best practices to create intuitive and user-friendly human-machine interfaces, Artif. Intell. Multimodal Signal Process. Hum.-Mach. Interact. (2025) 63–77.

[64] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, H. Liu, Natural language processing (NLP) in management research: A literature review, J. Manag. Anal. 7 (2) (2020) 139–172.

[65] A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, D. Dippold, Natural language processing for dialects of a language: A survey, ACM Comput. Surv. 57 (6) (2025) 1–37.

[66] I. Grobelna, D. Mailland, M. Horwat, Design of automotive HMI: New challenges in enhancing user experience, safety, and security, Appl. Sci. 15 (10) (2025) 5572.

[67] Z. Albassami, A. Algarni, A. Qahmash, Z. Ahmad, A comprehensive review of AI-driven Q&A systems with taxonomy, prospects, and challenges, Knowl. Inf. Syst. (2025) 1–24.

[68] C.-S. Lee, M.-H. Wang, J.-K. Chiang, N. Kubota, E. Sato-Shimokawara, Y. Nojima, G. Acampora, P.-Y. Wu, S.-C. Chiu, S.-C. Yang, et al., Quantum computational intelligence with generative AI image for human-machine interaction, in: 2024 IEEE International Conference on Fuzzy Systems, (FUZZ-IEEE), IEEE, 2024, pp. 1–8.

[69] A.V. Sadybekov, V. Katritch, Computational approaches streamlining drug discovery, Nature 616 (7958) (2023) 673–685.

[70] N. Berdigaliyev, M. Aljofan, An overview of drug discovery and development, Future Med. Chem. 12 (10) (2020) 939–947.

[71] Y.A. Ivanenkov, D. Polykovskiy, D. Bezrukov, B. Zagribelnyy, V. Aladinskiy, P. Kamya, A. Aliper, F. Ren, A. Zhavoronkov, Chemistry42: an AI-driven platform for molecular design and optimization, J. Chem. Inf. Model. 63 (3) (2023) 695–701.

[72] S.A. Stein, B. Baheri, D. Chen, Y. Mao, Q. Guan, A. Li, B. Fang, S. Xu, Qugan: A quantum state fidelity based generative adversarial network, in: 2021 IEEE International Conference on Quantum Computing and Engineering, QCE, IEEE, 2021, pp. 71–81.

[73] G.R. Bickerton, G.V. Paolini, J. Besnard, S. Muresan, A.L. Hopkins, Quantifying the chemical beauty of drugs, Nat. Chem. 4 (2) (2012) 90–98.

[74] H. Rathore, S. Sai, A. Gundewar, Social psychology inspired distributed ledger technique for anomaly detection in connected vehicles, IEEE Trans. Intell. Transp. Syst. 24 (7) (2023) 7092–7107, http://dx.doi.org/10.1109/TITS.2023.3262398.

[75] J. Xiong, L. Shen, Y. Liu, X. Fang, Enhancing IoT security in smart grids with quantum-resistant hybrid encryption, Sci. Rep. 15 (1) (2025) 3.

[76] S. Sai, I. Goyal, S. Sharma, S.H. Manuri, V. Chamola, R. Buyya, Quantum machine learning for cybersecurity: A taxonomy and future directions, 2025, arXiv preprint arXiv:2512.15286.

[77] E. Bermot, C. Zoufal, M. Grossi, J. Schuhmacher, F. Tacchino, S. Vallecorsa, I. Tavernelli, Quantum generative adversarial networks for anomaly detection in high energy physics, in: 2023 IEEE International Conference on Quantum Computing and Engineering, QCE, Vol. 1, IEEE, 2023, pp. 331–341.

[78] M. Xu, D. Niyato, J. Kang, Z. Xiong, Y. Cao, Y. Gao, C. Ren, H. Yu, Generative AI-enabled quantum computing networks and intelligent resource allocation, 2024, arXiv preprint arXiv:2401.07120.

[79] S. Ganguly, Implementing quantum generative adversarial network (qgan) and qcbm in finance, 2023, arXiv preprint arXiv:2308.08448.

[80] G.R. Steinbrecher, J.P. Olson, D. Englund, J. Carolan, Quantum optical neural networks, Npj Quantum Inf. 5 (1) (2019) 60.

[81] B. Lekitsch, S. Weidt, A.G. Fowler, K. Mølmer, S.J. Devitt, C. Wunderlich, W.K. Hensinger, Blueprint for a microwave trapped ion quantum computer, Sci. Adv. 3 (2) (2017) e1601540.

[82] T. Zhang, J. Sun, X.-X. Fang, X.-M. Zhang, X. Yuan, H. Lu, Experimental quantum state measurement with classical shadows, Phys. Rev. Lett. 127 (20) (2021) 200501.

[83] S. Sai, R. Buyya, Quantum artificial intelligence for mission-critical systems: Foundations, architectural elements, and future directions, Future Gener. Comput. Syst. 184 (2026) 108602, http://dx.doi.org/10.1016/j.future.2026.108602.

[84] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, et al., Quantum computational advantage using photons, Science 370 (6523) (2020) 1460–1463.

[85] L.S. Madsen, F. Laudenbach, M.F. Askarani, F. Rortais, T. Vincent, J.F. Bulmer, F.M. Miatto, L. Neuhaus, L.G. Helt, M.J. Collins, et al., Quantum computational advantage with a programmable photonic processor, Nature 606 (7912) (2022) 75–81.

[86] G. De Luca, A survey of NISQ era hybrid quantum-classical machine learning research, J. Artif. Intell. Technol. 2 (1) (2022) 9–15.

[87] Y. Hu, F. Tang, J. Chen, W. Wang, Quantum-enhanced reinforcement learning for control: a preliminary study, Control. Theory Technol. 19 (4) (2021) 455–464.

[88] S. Sai, A. Sawaika, P. Singh, R. Buyya, Quantum federated learning: Architectural elements and future directions, in: Federated Learning: Foundations and Applications, Morgan Kaufmann, Cambridge, MA, USA, 2026, pp. 325–343.

[89] L. Yee Por, Z. Dai, S. Juan Leem, Y. Chen, J. Yang, F. Binbeshr, K. Yuen Phan, C. Soon Ku, A systematic literature review on AI-based methods and challenges in detecting zero-day attacks, IEEE Access 12 (2024) 144150–144163, http://dx.doi.org/10.1109/ACCESS.2024.3455410.

[90] M. Gabr, A. Diab, H.T. Elshoush, Y.-L. Chen, L.Y. Por, C.S. Ku, W. Alexan, Data security utilizing a memristive coupled neural network in 3D models, IEEE Access 12 (2024) 116457–116477, http://dx.doi.org/10.1109/ACCESS.2024.3447075.

[91] M.S. Peelam, A.A. Rout, V. Chamola, Quantum computing applications for Internet of Things, IET Quantum Commun. 5 (2) (2024) 103–112, http://dx.doi.org/10.1049/qtc2.12073.

[92] S. Raj, S. Sai, Y. Simmhan, K. Chard, R. Buyya, Quantum integrated high-performance computing: Foundations, architectural elements and future directions, 2026, arXiv preprint arXiv:2604.19814.