QCircuitBench: A Large-Scale Dataset for Benchmarking Quantum Algorithm Design

Rui Yang Ziruo Wang Yuntian Gu Yitao Liang Tongyang Li





Center on Frontiers of Computing Studies, School of Computer Science, Peking University
School of Intelligence Science and Technology, Peking University
Institute for Artificial Intelligence, Peking University

Sat 6 Dec 8:30 a.m. CST — 11:30 a.m. CST

Motivation

Quantum Algorithm Design

- Deliver superpolynomial speedups.
- Challenging to design manually.

Large Language Models

- Abundant pre-training knowledge.
- Human-friendly interfaces.

Existing Benchmarks

- Target NISQ systems & tools—not AI training/evaluation.

Challenges

What challenges do we need to tackle?

Formulation: Natural Language? verbose, ambiguous (✗)
Math formulas? precise, but hard to verify automatically (✗)

Oracle Paradox: Theoretically: black-box.
Experimentally: explicit construction with quantum gates.

Classical Procedure: Quantum Algorithm = Quantum Circuit + Interpretation of Measurement Results.

Design Principles

Challenges

Formulation: Natural Language? (✗) Math formulas? (✗)

Oracle Paradox: Theoretically: black-box. Experimentally: explicit gates.

Classical Procedure: Quantum Algorithm = Quantum Circuit + Interpretation of Measurement Results.

Solutions

A code generation perspective

Represent quantum algorithms with quantum programming languages.

A Separate oracle.inc library

Preserve black-box abstraction while enabling compilation in OpenQASM.

Require post-processing functions

Include number of shots to characterize query complexity.

QCircuitBench

First large-scale benchmark for LLM-driven quantum algorithm design

- **Task Formulation**

  A general framework capturing the core aspects of quantum algorithm design.

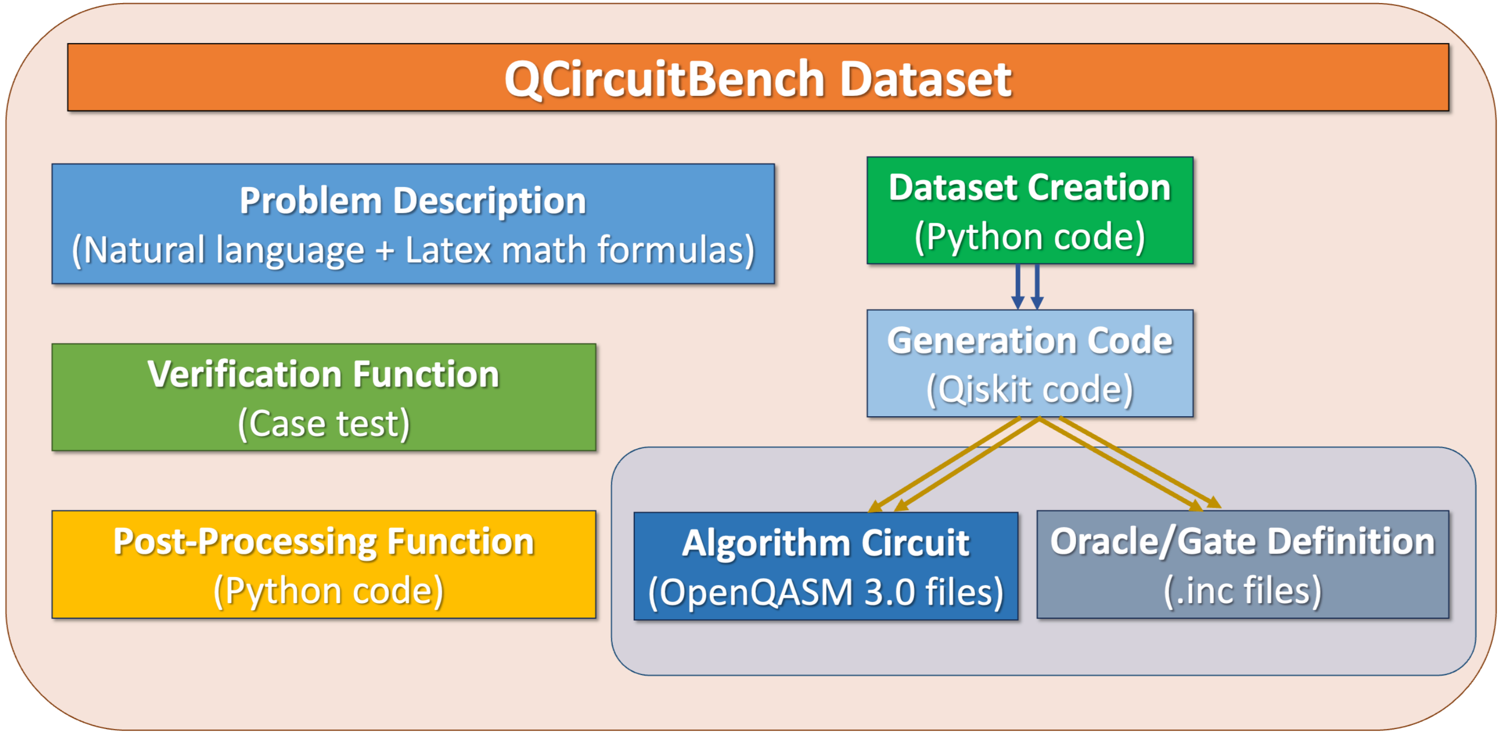
- **Rich Algorithm Coverage**

  Covering 3 task suites, 25 algorithms, 120,290 data points.

- **Automatic Verification**

  OpenQASM syntax, Python syntax, and task-specific semantic checks.

- **Training Potential**



Task Suite

❖ Oracle Construction

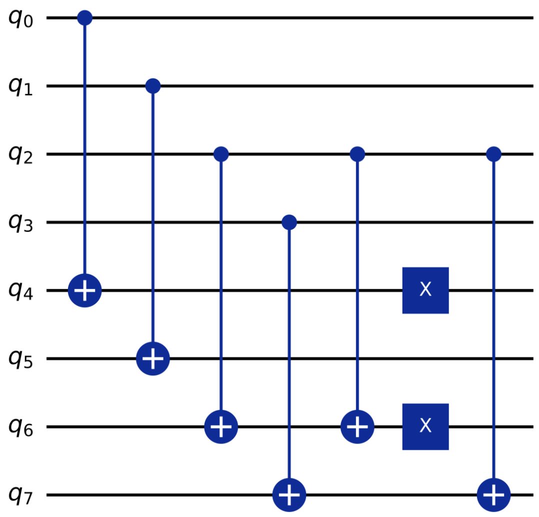
Encode Boolean function f as an oracle U_f such that $U_f|x\rangle|z\rangle = |x\rangle|z \oplus f(x)\rangle$.

❖ Quantum Algorithm Design

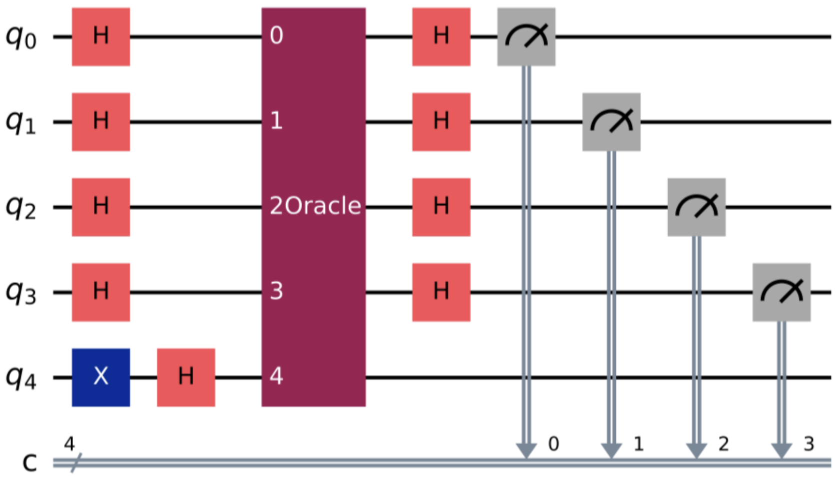
Covers textbook-level algorithms to advanced applications.

❖ Random Circuit Synthesis

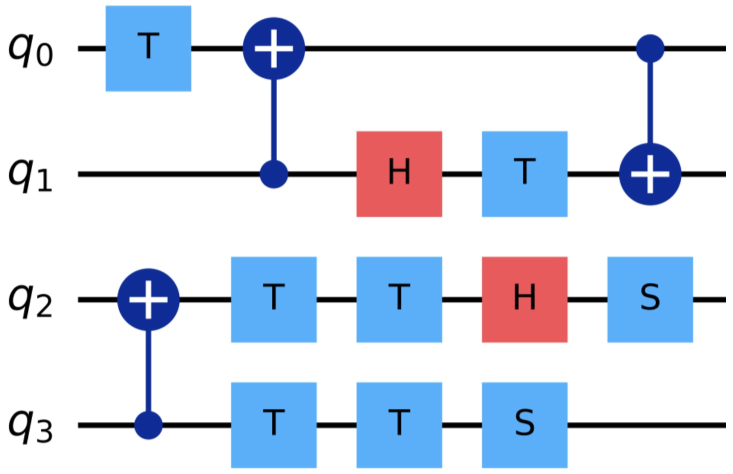
Reproduce quantum states from Clifford set {H, S, CNOT} / universal set {H, S, T, CNOT}.



(a) Simon's Problem (s=1100)



(b) Deutsch-Jozsa Algorithm



(c) Universal Circuits

Benchmark Results

Observations

- BLEU ≠ correctness: high BLEU score but low syntax/semantic accuracy.
- Few-shot > one-shot on structurally simple or template-like tasks; useless for hard ones.
- VQAs reveal challenges in modeling hybrid quantum-classical workflows.

Types of Errors

- Improvisation error: use unsupported namespaces/OpenQASM 3.0 features.
- Counting error: fail to identify ‘1’ bits in the secret string *(e.g., Bernstein-Vazirani)*.
- Data contamination: strong at generic Qiskit/Cirq, weak at gate-level QASM synthesis.

Fine-tuning Results

- LoRA-based fine-tuning on LLaMA3-8B.
- Improves scores, especially better at counting '1' bits *(Bernstein-Vazirani)*.
- Scores drop on random circuits, indicating challenge of encoding quantum state vectors within a language model and overfitting on tasks with high output diversity.

Table 2: Fine-tuning oracle construction scores.

Score	Model	Setting	Bernstein-Vazirani	Deutsch-Jozsa	Grover	Simon	Clifford	Universal	Avg
BLEU	gpt4o	few-shot(5)	95.6388 (±0.3062)	91.0564 (±0.6650)	92.0620 (±0.6288)	80.3390 (±2.0900)	39.5469 (±3.6983)	33.3673 (±3.1007)	72.0017
	Llama3	few-shot(5)	53.5574 (±5.2499)	69.8996 (±5.7812)	61.3102 (±5.4671)	26.3083 (±2.0048)	13.0729 (±0.9907)	13.4185 (±1.2299)	39.5945
	Llama3	finetune	76.0480 (±7.9255)	71.8378 (±2.4179)	67.7892 (±7.8900)	43.8469 (±3.2998)	10.8978 (±0.6169)	7.1854 (±0.5009)	46.2675
Verification	gpt4o	few-shot(5)	0.0000 (±0.0246)	0.4300 (±0.0590)	0.0000 (±0.1005)	-0.0200 (±0.0141)	-0.0333 (±0.0401)	-0.1023 (±0.0443)	0.0457
	Llama3	few-shot(5)	-0.2700 (±0.0468)	0.0900 (±0.0668)	-0.5200 (±0.0858)	-0.6600 (±0.0476)	-0.7303 (±0.0473)	-0.5056 (±0.0549)	-0.4327
	Llama3	finetune	-0.1300 (±0.0485)	-0.2000 (±0.0402)	-0.3300 (±0.0900)	-0.7400 (±0.0441)	-0.8741 (±0.0343)	-0.9342 (±0.0262)	-0.5347
PPL	Llama3	few-shot(5)	1.1967 (±0.0028)	1.1174 (±0.0015)	1.1527 (±0.0021)	1.1119 (±0.0017)	1.4486 (±0.0054)	1.4975 (±0.0051)	1.2541
	Llama3	finetune	1.0004 (±0.0002)	1.1090 (±0.0014)	1.0010 (±0.0006)	1.1072 (±0.0011)	1.2944 (±0.0053)	1.3299 (±0.0055)	1.1403

Takeaways

❖ **Novelty**

- First large-scale benchmark for LLM-driven quantum algorithm design.

❖ **Dataset Design**

- A perspective from code generation.
- Modular and extensible structure.
- Automatic verification functions.

❖ **Experiments**

- QCircuitBench poses significant challenges to SOTA LLMs.
- Fine-tuning experiments demonstrate early promise.

Thanks!