What is the comparative accuracy of Deepseek R1 and Codestral on multihop reasoning tasks in code generation, as evaluated on the HumanEval++ benchmark with varying levels of contextual complexity

Assignee Research

May 29, 2026

Abstract

Quantum programs are typically developed using quantum Software Development Kits (SDKs). The rapid advancement of quantum computing necessitates new tools to streamline this development process, and one such tool could be Generative Artificial intelligence (GenAI). In this study, we introduce and use the Qiskit HumanEval dataset, a hand-curated collection of tasks designed to benchmark the ability of Large Language Models (LLMs) to produce quantum code using Qiskit - a quantum SDK. This dataset consists of more than 100 quantum computing tasks, each accompanied by a prompt, a canonical

1 Introduction

This paper examines: Qiskit HumanEval: An Evaluation Benchmark For Quantum Code Generative Models. Research question: What is the comparative accuracy of Deepseek R1 and Codestral on multihop reasoning tasks in code generation, as evaluated on the HumanEval++ benchmark with varying levels of contextual complexity?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

15 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
There is no research work evaluating the performance of the quantum code generated by LLMs.	×	0.13
HumanEval [8] and MBPP [16], as well as their improved version EvalPlus [7], are the most known benchmarks for code gene	×	0.08
These benchmarks are manually curated to avoid any possible data leakage.	×	0.02
Another relevant benchmark for code generation is DS-1000 [17] which is aligned with real-world problems in data-science	×	0.06
The Qiskit HumanEval dataset is designed for an automated assessment of the LLM in producing executable and functionally	×	0.11
The Qiskit HumanEval dataset contains 28 tasks in the QUANTUM CIRCUIT GENERATION category.	×	0.10
The Qiskit HumanEval dataset contains 19 tasks in the SIMULATION AND EXECUTION category.	×	0.06
The Qiskit HumanEval dataset contains 7 tasks in the STATE PREPARATION AND ANALYSIS category.	×	0.07
The Qiskit HumanEval dataset contains 14 tasks in the ALGORITHM IMPLEMENTATION category.	×	0.06
The Qiskit HumanEval dataset contains 17 tasks in the GATE OPERATIONS AND MANIPULATION category.	×	0.06
The Qiskit HumanEval dataset contains 6 tasks in the VISUALIZATION AND POST-PROCESSING category.	×	0.06
The Qiskit HumanEval dataset contains 8 tasks in the ADVANCED CIRCUIT MANIPULATION category.	×	0.07
The Qiskit HumanEval dataset contains 2 tasks in the QUANTUM CIRCUIT SERIALIZATION category.	×	0.10
The Qiskit HumanEval dataset is organized into three difficulty levels: basic, intermediate, and advanced.	×	0.06
The Qiskit HumanEval dataset includes tasks that require an in-depth understanding of the BB84 algorithm.	×	0.05
The Qiskit HumanEval dataset is designed to provide a robust framework for assessing the performance of generated quantu	×	0.13

References

- http://arxiv.org/abs/2406.14712v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2306.08568v2