AI REPORT

Project Name: 2B or not 2B
Team Name: FLMA8
GitHub Repository: https://github.com/sareensabra/2026-NVIDIA

THE WORKFLOW

We structured our AI-assisted development across three specialized agents:

Coda Agent
Coding / Implementation

- Responsible for generating quantum circuit implementations (VQE), CUDA-Q GPU kernels, and classical MTS code.
- We used the agent iteratively: first to draft functions, then to optimize kernels, and finally to refactor parallel FFT-based MTS routines for GPU execution.

ChatGPT Instance
Testing & Verification

- Assisted in unit test design, error analysis, and logic validation.
- Generated property tests for Hamiltonian symmetry, autocorrelation correctness, and CPU-GPU parity.
- Suggested test cases for verifying that AI-generated kernels were numerically consistent with theoretical expectations.

Gemini Agent
Complex Computations / Optimization Strategy

- Used for evaluating candidate ansatz configurations and estimating circuit depth vs. convergence trade-offs.
- Helped simulate multiple VQE seeds and rank them for downstream MTS seeding.
- Assisted with speedup calculations and analysis of FFT-based autocorrelation vs direct computation.
- Workflow Summary:
- VQE circuit generation → GPU kernel optimization (Coda) → verification with unit tests (ChatGPT) → hybrid analysis and benchmarking (Gemini) → iterative feedback and refactoring.

VERIFICATION STRATEGY

To ensure AI-generated code was correct and reliable, we AI Guard Rails and Unit Tests to manage logic errors

AI Guardrails

- Created a **skills.md file** documenting CUDA-Q API conventions, Hamiltonian definitions, and FFT rules.
- All AI prompts included references to these files to prevent hallucinations in function calls or syntax.

Unit Tests

Symmetry Test

- Verify that LABS energy is invariant under sequence negation and reversal

$$E(S) = E(-S) = E(reverse(S))$$

- Detects AI-generated circuits or kernels that miscompute autocorrelation or misimplement the Hamiltonian.

CPU-GPU Parity Test

- Compare energy outputs from CPU and GPU implementations for the same input sequences.
- Ensures AI-generated GPU kernels match the reference CPU logic; detects misaligned parallelization or implementation errors.

FFT vs Direct Autocorrelation Test

- Validate that the FFT-based MTS kernel produces results within ε tolerance of the O(N²) baseline.
- Confirms AI correctly rewrote autocorrelation into FFT while maintaining accuracy.

Seed Effectiveness Test

- Compare MTS convergence with VQE-generated seeds versus random initialization.
- Verifies that AI-produced VQE seeds improve classical optimization performance.

Small N Ground Truth Test

- For small problem sizes, assert that GPU/CPU energies match known optimal solutions.
- Catches logical or numerical errors in AI-generated code before scaling to larger N.

THE VIBE LOG

Win

★ The Coda agent automatically refactored the classical MTS autocorrelation computation into batched FFT kernels.

★ This saved hours of manual CUDA optimization and reduced wall-clock evaluation time for N=40 sequences by ~40%.

Learn

➔ Initially, ChatGPT produced incomplete unit tests that didn't check GPU vs CPU parity.

➔ After adding a structured skills.md and including Hamiltonian symmetry rules in prompts, the AI generated full property-based tests, which caught subtle numerical mismatches early.

Fail

✖ Gemini suggested a VQE ansatz with too many entangling layers, which caused circuits to exceed simulated GPU memory limits.

✖ We fixed this by manually transpiling circuits into smaller mini-circuits and parallelizing them, then updating AI prompts to include "max circuit depth" and "parallelize mini-circuits" guidance.