

# GroverGPT: A Large Language Model with 8 Billion Parameters for Quantum Searching

Haoran Wang\*,<sup>1</sup> Pingzhi Li\*,<sup>1</sup> Min Chen,<sup>2</sup> Jinglei Cheng,<sup>2</sup> Junyu Liu<sup>‡,2</sup> and Tianlong Chen<sup>‡1</sup>

<sup>1</sup>Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

<sup>2</sup>Department of Computer Science, The University of Pittsburgh, Pittsburgh, PA 15260, USA

\*These authors contributed equally to this work.

<sup>‡</sup>Co-corresponding authors.

tianlong@cs.unc.edu, junyuliu@pitt.edu

Quantum computing is an exciting non-Von Neumann paradigm, offering provable speedups over classical computing for specific problems. However, the practical limits of classical simulatability for quantum circuits remain unclear, especially with current noisy quantum devices. In this work, we explore the potential of leveraging Large Language Models (LLMs) to simulate the output of a quantum Turing machine using Grover’s quantum circuits, known to provide quadratic speedups over classical counterparts. To this end, we developed GroverGPT, a specialized model based on LLaMA’s 8-billion-parameter architecture, trained on over 15 trillion tokens. Unlike brute-force state-vector simulations, which demand substantial computational resources, GroverGPT employs pattern recognition to approximate quantum search algorithms without explicitly representing quantum states. Analyzing 97K quantum search instances, GroverGPT consistently outperformed OpenAI’s GPT-4o (45% accuracy), achieving nearly 100% accuracy on 6- and 10-qubit datasets when trained on 4-qubit or larger datasets. It also demonstrated strong generalization, surpassing 95% accuracy for systems with over 20 qubits when trained on 3- to 6-qubit data. Analysis indicates GroverGPT captures quantum features of Grover’s search rather than classical patterns, supported by novel prompting strategies to enhance performance. Although accuracy declines with increasing system size, these findings offer insights into the practical boundaries of classical simulatability. This work suggests task-specific LLMs can surpass general-purpose models like GPT-4o in quantum algorithm learning and serve as powerful tools for advancing quantum research.

## I. INTRODUCTION

Quantum computing harnesses fundamental quantum mechanical phenomena, such as superposition and entanglement, to solve certain computational problems exponentially faster than classical computers [1, 2]. For instance, Grover’s algorithm [2], a quantum algorithm for database searching, is proven to achieve a quadratic speedup over all known classical counterparts. However, the boundaries of *quantum advantage*—where quantum algorithms outperform all classical counterparts for a specific task—remain unclear. In practice, a closely related question concerns classical simulatability: if a quantum circuit can be efficiently simulated on classical computers, it is unlikely to demonstrate a significant advantage [3].

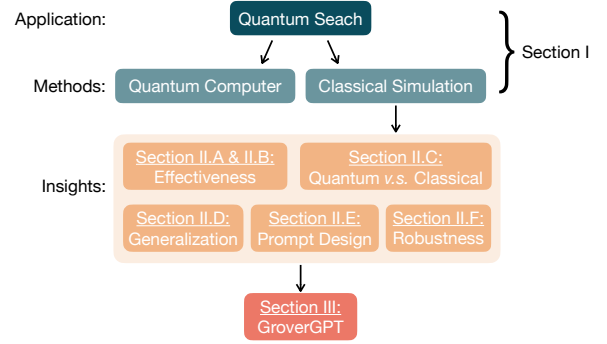


Figure 1. **Overview.** We investigate the classical simulation of quantum search through GroverGPT, a large language model approach. Starting from quantum search’s implementations via quantum machines and classical simulation, we evaluate GroverGPT along four dimensions: effectiveness of quantum search simulation, generalization from small to large qubit systems, comparative analysis between quantum and classical approaches, and the role of prompt engineering. Through these investigations, GroverGPT demonstrates promising capabilities in bridging quantum-classical computational boundaries.

Based on the construction of a quantum Turing machine, brute-force classical simulation—commonly referred to as state-vector simulation—incurs exponentially high memory costs for general quantum computing tasks. Smarter, approximate approaches, such as tensor networks, may perform better in practice [4], although they can still face exponentially high costs or exponentially large errors as the number of qubits scales for general tasks. The situation becomes even more complex with NISQ (Noisy Intermediate-Scale Quantum) devices [5]; it remains unclear whether existing noisy, near-term commercial quantum computers can be classically simulated for commercially valuable problems [6].

Thus, the practical frontier of classical simulatability can only be approached with *large-scale high-performance computing*. For example, several claims of demonstrating quantum advantages in sampling algorithms [7] have been challenged by tensor network methods implemented on classical devices [8, 9]. On the other hand, novel approaches utilizing Large Language Models (LLM) [10–12], one of the most powerful large-scale AI tools, present new possibilities for simulating quantum circuits [13–16]. Although notable progress has been achieved, these early efforts have primarily concentrated on prompt engineering using commercial models from OpenAI and similar providers, building open-source datasets, or developing toy models with only a few thousand parameters—significantly smaller than modern industrial-level models, which typically have at least a few billion parameters.

In this work, we present an initial investigation into the clas-

sical simulation limits of Grover’s algorithm using a quantum-focused LLM. *We introduce GroverGPT, the first LLM, to the best of our knowledge, capable of simulating quantum algorithms at an industrial medium scale.* GroverGPT combines quantum circuit simulation with natural language processing (*i.e.*, a branch of AI that enables computers to understand and generate human language) to explore how effectively classical systems can emulate quantum search algorithms. The model is built upon LLaMA’s 8-billion architecture, a state-of-the-art language model developed for processing and generating text trained on a dataset exceeding 15 trillion tokens [12]. To simulate a quantum Turing machine, the input of the model will be a classical description of the quantum circuit (in our case it is Grover), and the output will be a sequence of probability distribution for each bit string.

Our data construction methodology integrates three key components: 1) *quantum circuit representations* – diagrams showing sequences of quantum operations, 2) *quantum assembly language* – QASM [17], a standardized programming language for describing quantum operations similar to classical computer assembly code, and 3) natural language interaction. These components are unified through a carefully designed pre-training pipeline based on the Llama architecture [12]. It allows us to begin studying the capability of classical systems to learn and generalize quantum principles without maintaining explicit quantum states.

As a first step toward understanding these simulation capabilities, we analyze our approach using quantum-specific metrics – *search accuracy* ( $\alpha$ ), *infidelity* ( $\epsilon$ ), and *marked infidelity* ( $\epsilon^k$ ). Our preliminary results show promising performance in simulating quantum search on moderate-sized systems, with encouraging signs of generalization from training on small systems (3  $\sim$  6 qubits) to somewhat larger quantum registers (*e.g.*, 20 qubits). While these initial findings suggest interesting possibilities for classical simulation of quantum algorithms, they also reveal important limitations that appear to be fundamental rather than technical. This exploration provides early insights into the challenges and opportunities at the boundary between classical and quantum computation. This initial investigation contributes to quantum computing research through four key aspects:

- A comprehensive dataset comprising 97K quantum search examples across different qubit sizes (3  $\sim$  20 qubits), including quantum circuit simulations, QASM representations, and natural language descriptions, which we release to facilitate further research in quantum algorithm simulation.
- A novel experimental framework for studying classical simulation limits of quantum algorithms through language model-based approximation. We released a pre-trained 8-billion-parameter language model (GroverGPT) specialized in quantum search simulation.
- Initial empirical observations about how classical systems might learn and generalize quantum principles without explicit quantum state representation. We explore a potential approach to quantum algorithm simulation that aims to balance infidelity with computational efficiency.

The structure of this paper is organized in Figure 1. As a summary, we obtain the following insights:

- The model can outperform general purpose models like OpenAI’s GPT-4o (Section II A and II B).
- The model can generalize towards up to 20 qubits by only looking at few-qubits examples, but the capability drastically decreases with the number of qubits (Section II C). This result indicates that the model somehow knows the structure of the algorithms, but the capability is still limited by the exponential growth of the Hilbert space.
- The model can learn features of quantum searching instead of just learn a classical searching algorithm (Section II D).
- The model can learn some structures of quantum circuits with the help of QASM languages as inputs, and some prompt engineering frameworks have been developed to make it learn better (Section II E).
- The model’s robustness is comprehensively evaluated (Section II F), highlighting the impact of training strategies and dataset diversity on performance consistency.

Technical methods are summarized in Section III, and related knowledge and experimental details have been summarized in Appendix for knowledge completeness. Conclusion and outlook is provided in Section IV.

## II. GROVERGPT

### A. Pre-Training Strategy Selection for GroverGPT

As illustrated in Figure 2, we develop GroverGPT through a structured pre-training pipeline built upon the Llama-3.1-8B [12] foundation model. Our training dataset encompasses quantum systems ranging from 3 to 20 qubits, carefully separating training (3  $\sim$  10 qubits) and testing (6  $\sim$  20 qubits) sets. The pre-training process integrates three key components: quantum circuit simulations from Grover’s algorithm implementations, corresponding QASM code generated via Qiskit [18], and natural language conversations about quantum search problems. This multi-modal approach enables GroverGPT to comprehensively understand the structural and behavioral aspects of quantum search operations. Data augmentation through QASM representations and natural language interactions further enhances the model’s ability to bridge the gap between abstract quantum algorithms and practical implementations.

This refined version maintains the essential information while being more concise and focused. It highlights the key components of our pre-training strategy while eliminating redundant details, making it more accessible to readers while preserving technical accuracy.

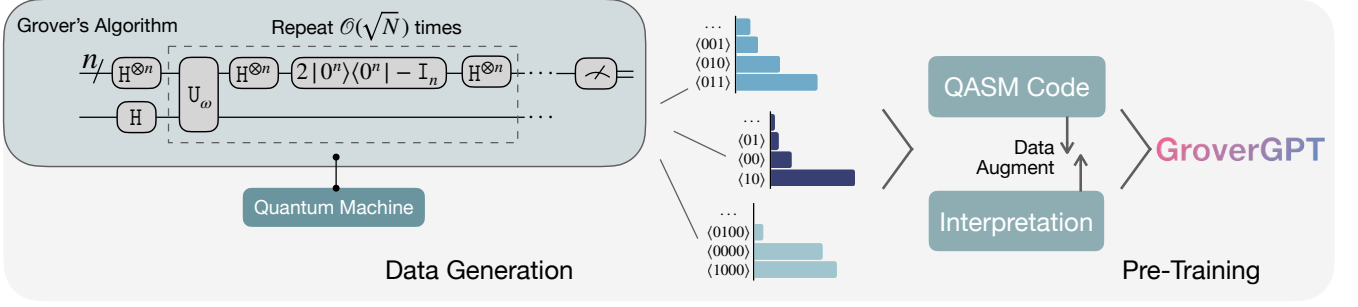


Figure 2. **Overview of GroverGPT’s pre-training pipeline.** From left to right: (1) Data generation begins with implementing Grover’s algorithm on a simulated quantum machine, repeating the circuit  $\mathcal{O}(\sqrt{N})$  times to construct comprehensive training data where  $N = 2^n$  and  $n$  is the number of qubits. (2) The measurement outcomes are collected, represented by probability distributions across different computational basis states (shown in color-coded bars for different qubit configurations). (3) The corresponding QASM code is generated to provide standardized circuit descriptions. (4) These components are combined through augmented training, integrating both quantum circuit information and measurement data to pre-train the GroverGPT model, which builds upon the Llama-3.1-8B [12] architecture.

## B. GroverGPT’s Effectiveness of Simulating Quantum Search

This section demonstrates GroverGPT’s superior performance in quantum search simulation compared to baseline models. Our empirical results demonstrate that GroverGPT significantly outperforms the baseline GPT-4o model in quantum search simulation. As shown in Figure 3c, while GPT-4o maintains a relatively constant accuracy of approximately 45% across different training qubit ranges, GroverGPT achieves substantially higher accuracy, reaching nearly 100% when trained on 5 or 6 qubits. This performance gap is particularly evident in both the 6-qubit (left) and 10-qubit (right) test scenarios. The integration of QASM and conversation components further enhances GroverGPT’s performance, especially in scenarios with fewer training qubits ( $3 \sim 4$  qubits), where the accuracy improves from around 70% to over 80%. This consistent improvement pattern suggests that GroverGPT has successfully learned to emulate the fundamental principles of quantum search, rather than merely approximating classical search strategies.

## C. Scalability and Generalization Capability of GroverGPT

In this section, we examine GroverGPT’s remarkable ability to generalize quantum search capabilities from training on small-scale systems to significantly larger quantum systems. GroverGPT exhibits remarkable generalization capabilities across different qubit scales. Figure 3e comprehensively analyzes the model’s performance when trained on  $3 \sim 6$  qubits and tested on increasingly larger systems, ranging from 6 to 20 qubits. The results show that GroverGPT maintains accuracy above 95% for systems up to 8 qubits, with only a gradual decline to approximately 95% for systems between 9 and 20 qubits. This robust generalization ability is particularly noteworthy given that the model was trained only on smaller systems ( $3 \sim 6$  qubits), suggesting its capacity to extrapolate quantum search principles to substantially larger quantum systems. The parallel trends between accuracy and infidelity in

Figure 3e suggest that the model’s performance degradation at larger qubit counts is gradual and predictable, indicating a systematic rather than catastrophic breakdown of simulation capability. This behavior aligns with theoretical expectations about the scalability challenges in quantum simulation and provides valuable insights into the practical limits of classical models in capturing quantum phenomena.

## D. Quantum v.s. Classical Search Learned by GroverGPT

Through detailed fidelity analysis in this section, we investigate whether GroverGPT truly learns quantum search rather than simply approximating classical search strategies. The fidelity analysis in Figure 3d provides strong evidence that GroverGPT learns genuine quantum search rather than classical approximations. The consistently low infidelity values (below 0.005 for 6 qubits and approaching zero for systems trained on 5 or more qubits) indicate that GroverGPT accurately captures the quantum state amplitudes characteristic of Grover’s algorithm. This is in stark contrast to GPT-4o [19]’s performance, which shows a consistently high infidelity (approximately 0.011), suggesting it fails to capture the subtle quantum features of the search process. Similarly, models such as DeepSeek-V2-Lite [20, 21] and Llama-3.2-3B [12] exhibit even lower accuracy and higher infidelity values, emphasizing their limitations in capturing the quantum search process. The convergence of infidelity to near-zero values with increased training qubit count is particularly significant, as it indicates that GroverGPT successfully learns the interference patterns and amplitude amplification mechanisms that are quintessential to quantum search. This distinction is further reinforced by the symmetric behavior observed in both 6-qubit and 10-qubit test cases, suggesting that the model has internalized fundamental quantum principles rather than memorizing specific problem instances.

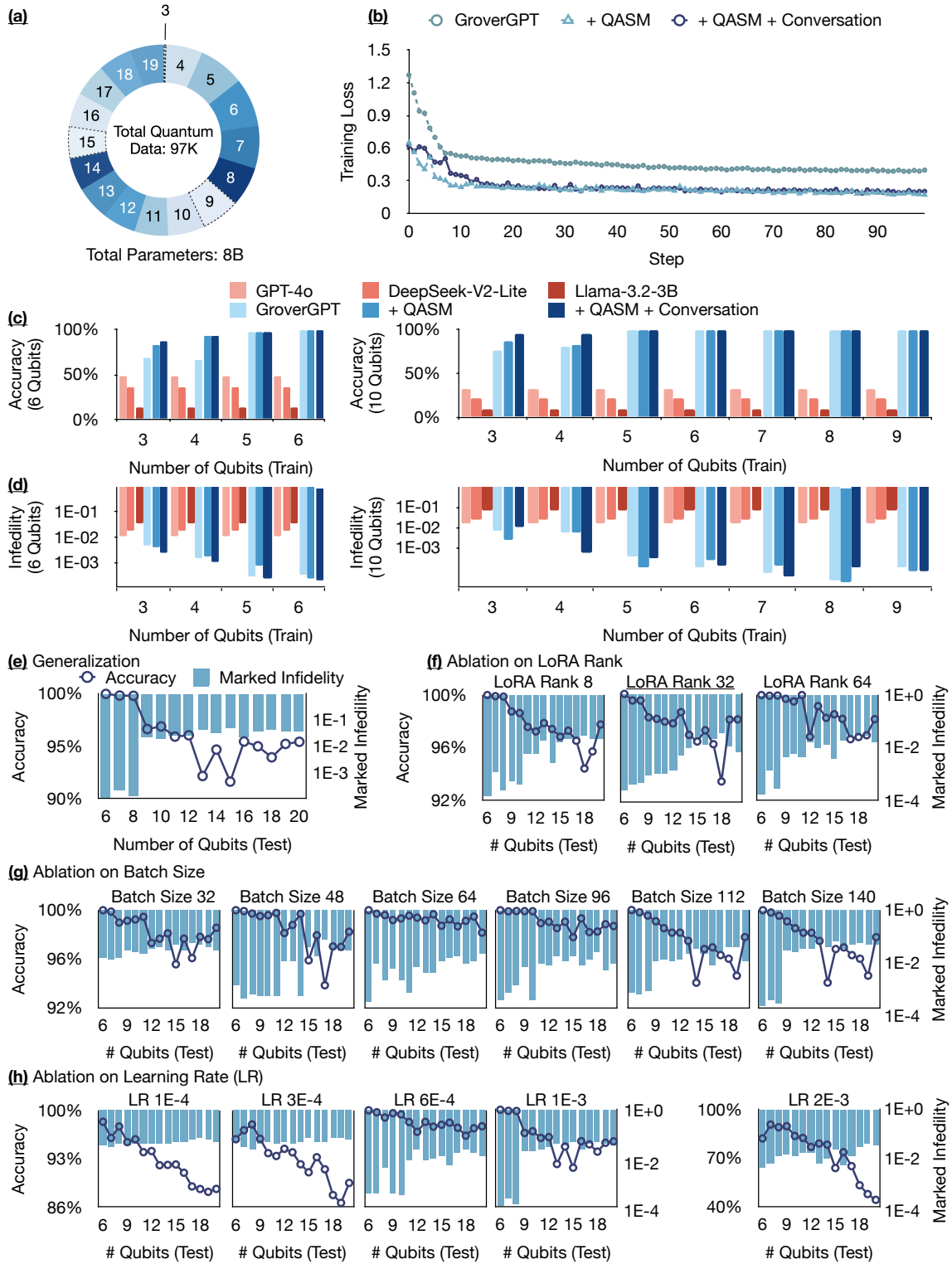


Figure 3. **Performance evaluation and generalization capability of GroverGPT.** (a) Distribution of the pre-training dataset comprising 97K quantum search examples across different qubit sizes from 3 to 20. (b) Training loss curves for different GroverGPT variants, showing convergence behavior during pre-training on 3-6 qubit datasets. (c) Comparative accuracy analysis of GPT-4o, various open-source large models, GroverGPT, and its variants with QASM and conversation components on 6-qubit (left) and 10-qubit (right) test sets across different training qubit ranges. (d) Infidelity  $\epsilon$  comparison between models on 6-qubit (left) and 10-qubit (right) test sets, demonstrating the error reduction as training qubit count increases. (e) Scalability assessment of GroverGPT trained on 3 ~ 6 qubits, showing accuracy (blue line) and Marked Infidelity (blue bars) when tested on larger systems ranging from 6 to 20 qubits, highlighting the LLM’s generalization capabilities beyond its training domain. (f) (g) (h) Model performance across a wide range of hyper-parameters (LoRA rank, batch size, learning rate, respectively), highlighting accuracy as a robust indicator.

### E. Impact of Prompt Design Strategies in GroverGPT

In this section, we evaluate how different prompting strategies, particularly the integration of QASM code and conversational components, influence GroverGPT’s performance in quantum search simulation. The influence of different prompting strategies [22, 23] is clearly illustrated in Figure 3c, where we compare the base GroverGPT model with variants incorporating QASM and conversation components. The addition of QASM prompts improves performance with a substantial margin, particularly evident in the 3 ~ 4 qubit training scenarios where accuracy increases by approximately 10 ~ 15 percentage points. Further enhancement through conversation components yields additional improvements, particularly notable in the 6-qubit test case (left panel). This hierarchical improvement pattern suggests that structured quantum circuit descriptions (QASM) combined with natural language interaction create a more robust framework for quantum search simulation. The synergistic effect of these prompting strategies indicates that the model benefits from both the precision of formal quantum circuit descriptions and the contextual richness of natural language interaction. This finding has important implications for the design of future quantum simulation interfaces and educational tools, suggesting that a multi-modal approach to quantum algorithm description might be optimal for both performance and accessibility.

### F. Robustness of the GroverGPT Training Strategy

In this section, we evaluate the robustness of GroverGPT by examining the limited variability among its metric indicators and investigating how diverse training sets affect its overall performance.

**Ablation Study on Training Configurations.** Our comprehensive ablation studies investigating LoRA rank, batch size, and learning rate, presented in Figure 3(f)(g)(h), demonstrate that GroverGPT exhibits remarkable stability across various hyper-parameter configurations. The model maintains consistent performance in terms of both accuracy and infidelity metrics under standard training conditions. However, performance degradation becomes evident at extreme parameter values, particularly when the learning rate falls below  $3 \times 10^{-4}$  or exceeds  $1 \times 10^{-3}$ . These findings underscore GroverGPT’s robustness within conventional hyper-parameter ranges, suggesting a practical advantage for real-world implementations.

**Training Diversity for Robustness.** Our experimental results demonstrate that training set diversity plays a crucial role in enhancing model robustness. Figure 4 illustrates this through four heatmaps generated under different hyper-parameter configurations. We systematically expanded the training dataset from a single qubit configuration (*e.g.*, qubits=3) to increasingly diverse combinations (*e.g.*, qubits=[3, 4], [3, 4, 5], and [3, 4, 5, 6]). The results reveal that as training data diversity increases, regions of high accuracy become more distinctly defined within the hyper-parameter space, facilitating more effective optimization. These findings suggest that enhanced data diversity not only strengthens model robustness but may also

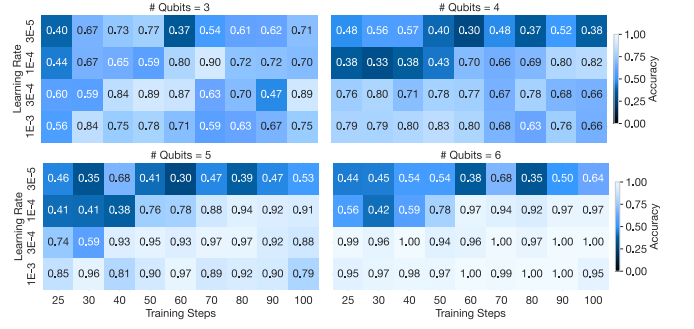


Figure 4. The model’s robustness across diverse training datasets on different qubit systems illustrates its sensitivity to hyper-parameters during finetuning.

improve generalization capabilities across other challenging scenarios.

## III. METHODOLOGY

### A. Grover’s Algorithm

As demonstrated in [2, 24, 25], Grover’s Algorithm [2] represents a fundamental breakthrough in quantum computing, demonstrating significant computational advantages over classical methods. For the ubiquitous task of searching through an unstructured database of  $N$  items, classical computers require examining items one by one, taking  $O(N)$  operations on average. In contrast, Grover’s Algorithm achieves a quadratic speedup, requiring only  $O(\sqrt{N})$  operations by leveraging principles of quantum mechanics.

As indicated by Figure 3 (upper left), the algorithm creates a quantum state that simultaneously represents all possible database items through superposition – a quantum property where a qubit can exist in multiple states simultaneously. It then iteratively applies two operations: one that marks the desired solution and another that amplifies the probability of finding this marked state. After  $\sim \sqrt{N}$  iterations, measuring the system yields the target item with high probability.

### B. Classical Simulation of GroverGPT

We adopt a noise-free classical simulation of Grover’s quantum search algorithm using Qiskit [18], which is an open-source software development kit for quantum computing developed by IBM, and its state vector simulator to establish ground truth data for training and evaluation. This simulator tracks the complete mathematical description of a quantum system: for an  $n$ -qubit system where a qubit is the quantum equivalent of a classical bit, the simulator maintains a state vector of dimension  $N = 2^n$ . It enables the exact computation of quantum state amplitudes, which are the complex numbers that describe the probability of measuring each possible quantum state, along with their measurement probabilities.



As shown in Figure 2, the simulation executes the standard Grover circuit template, which is a sequence of quantum operations, with  $\mathcal{O}(\sqrt{2^n})$  iterations. This process comprises three essential components: initialization through Hadamard gates, which are quantum operations that create equal superpositions; oracle operations for target state marking, which are operations that identify the solution we’re searching for; and diffusion operators for amplitude amplification, which are operations that enhance the probability of finding the marked state. Through multiple simulation shots, which are repeated runs of the quantum circuit, we obtain relatively precise probability distributions over computational basis states, which represents the possible measurement outcomes. This provides reliable reference data for assessing GroverGPT’s quantum search simulation capabilities.

### C. Evaluation Metrics

To rigorously assess GroverGPT’s quantum search simulation capabilities, we establish three complementary evaluation metrics. Given a quantum system with  $n$  qubits and  $k$  marked states, let  $|\psi_{\text{final}}\rangle$  represent the final quantum state after applying Grover’s algorithm. The measurement outcome probability distribution is defined as:

$$\mathcal{P} = (s_i, p_i) \mid i \in [2^n], p_i = |\langle s_i | \psi_{\text{final}} \rangle|^2 \quad (1)$$

where  $s_i$  represents the  $i$ -th computational basis state and  $p_i$  its corresponding measurement probability. Let  $\mathcal{P}_{\text{model}}$  and  $\mathcal{P}_{\text{true}}$  denote the probability distributions generated by GroverGPT and the ideal quantum simulator, respectively. We evaluate performance using:

- **Search Accuracy ( $\alpha$ ):** Measures the model’s ability to identify marked states correctly, defined as:

$$\alpha = \frac{|\mathcal{M}_{\text{model}} \cap \mathcal{M}_{\text{true}}|}{k} \quad (2)$$

where  $\mathcal{M}_{\text{model}}$  and  $\mathcal{M}_{\text{true}}$  are the sets of  $k$  highest-probability states in  $\mathcal{P}_{\text{model}}$  and  $\mathcal{P}_{\text{true}}$ , respectively.

- **Infidelity ( $\epsilon$ ):** Quantifies the overall quantum state reproduction accuracy through:

$$\epsilon = \frac{1}{2^n} \sum_{i=1}^{2^n} (p_i^{\text{model}} - p_i^{\text{true}})^2 \quad (3)$$

where  $p_i^{\text{model}}$  and  $p_i^{\text{true}}$  are probabilities from  $\mathcal{P}_{\text{model}}$  and  $\mathcal{P}_{\text{true}}$ .

- **Marked Infidelity ( $\epsilon^k$ ):** Specifically evaluates the accuracy of marked state probability predictions:

$$\epsilon^k = \frac{1}{k} \sum_{s_i \in \mathcal{M}_{\text{true}}} (p_i^{\text{model}} - p_i^{\text{true}})^2 \quad (4)$$

These metrics provide complementary perspectives on GroverGPT’s performance:  $\alpha$  assesses the model’s ability to identify correct search solutions,  $\epsilon$  evaluates the overall quantum state reproduction infidelity, and  $\epsilon^k$  focuses specifically on the accuracy of marked state predictions. Together, they comprehensively evaluate the model’s practical search capabilities and theoretical quantum properties.

## IV. CONCLUSION AND OUTLOOK

In this work, we demonstrate that an industrial-level, medium-scale LLM can be designed to simulate noiseless quantum algorithms, such as Grover’s search, and potentially outperform popular general-purpose models like OpenAI’s GPT-4. This is achieved through the development of GroverGPT. Additionally, we present evidence suggesting that the model can learn and generalize certain features of quantum algorithms, though its performance significantly deteriorates as the number of qubits increases. Our work offers a valuable tool for advancing research and education in quantum algorithms.

Our work opens a new avenue for exploring the boundaries of classical simulability and quantum advantage using LLMs. This raises numerous questions for future research. For example, how well can LLMs simulate noisy quantum computers? Could they effectively model current noisy quantum systems with 100 to 1,000 qubits, and what level of error would be tolerable? Can other quantum algorithms be simulated, or might it even be possible to develop a foundational model capable of simulating quantum Turing machines? Can we simulate quantum error correction codes? Furthermore, if resources from leading LLM developers such as OpenAI, Anthropic, or xAI were available for training models specifically tailored to quantum algorithms, how many qubits and what circuit depth could we feasibly simulate? These intriguing questions remain open for future investigation.

## ACKNOWLEDGMENT

We thank Yangrui Hu, Jin-Peng Liu, Ziwen Liu, Zhiding Liang and Suqiong Zeng for discussion and help. MC, JC, and JL are supported in part by the University of Pittsburgh, School of Computing and Information, Department of Computer Science, Pitt Cyber, and the PQI Community Collaboration Awards. PL and TC are supported in part by Cisco Faculty Award and UNC SDSS Seed Grant.

- [1] Shor, P. Algorithms for quantum computation: discrete logarithms and factoring. *Proceedings 35th Annual Symposium On Foundations Of Computer Science*, pp. 124-134 (1994)
- [2] Grover, L. A fast quantum mechanical algorithm for database search. *Proceedings Of The Twenty-eighth Annual ACM Symposium On Theory Of Computing*, pp. 212-219 (1996)
- [3] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.
- [4] G. Vidal, "Efficient classical simulation of slightly entangled quantum computations," *Physical Review Letters*, vol. 91, no. 14, p. 147902, 2003.
- [5] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, 2018.
- [6] J. Preskill, "Beyond NISQ: The Megaquop Machine," Q2B 2024. [Online]. Available: <https://www.preskill.caltech.edu/talks/Preskill-Q2B-2024.pdf>
- [7] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.
- [8] C. Oh, M. Liu, Y. Alexeev, B. Fefferman, and L. Jiang, "Classical algorithm for simulating experimental Gaussian boson sampling," *Nature Physics*, pp. 1–8, 2024.
- [9] F. Pan, K. Chen, and P. Zhang, "Solving the sampling problem of the Sycamore quantum circuits," *Physical Review Letters*, vol. 129, no. 9, p. 090502, 2022.
- [10] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. & Others Llama 2: Open foundation and fine-tuned chat models. *ArXiv Preprint arXiv:2307.09288*. (2023)
- [11] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. & Others The llama 3 herd of models. *ArXiv Preprint arXiv:2407.21783*. (2024)
- [12] Vavekanand, R. & Sam, K. Llama 3.1: An in-depth analysis of the next-generation large language model. (ResearchGate, 2024)
- [13] Liang, Zhiding, Cheng, Jinglei, Yang, Rui, Ren, Hang, Song, Zhixin, Wu, Di, Qian, Xuehai, Li, Tongyang, and Shi, Yiyu. "Unleashing the potential of LLMs for quantum computing: A study in quantum architecture design." *arXiv preprint arXiv:2307.08191*, 2023.
- [14] Yang, Rui, Gu, Yuntian, Wang, Ziruo, Liang, Yitao, and Li, Tongyang. "QCCircuitNet: A Large-Scale Hierarchical Dataset for Quantum Algorithm Design." *arXiv preprint arXiv:2410.07961*, 2024.
- [15] Yao, Jian, and You, Yi-Zhuang. "ShadowGPT: Learning to Solve Quantum Many-Body Problems from Randomized Measurements." *arXiv preprint arXiv:2411.03285*, 2024.
- [16] Fitzek, David, Teoh, Yi Hong, Fung, Hin Pok, Dagnew, Gebremedhin A., Merali, Ejaz, Moss, M. Schuyler, MacLellan, Benjamin, and Melko, Roger G. "RydbergGPT." *arXiv preprint arXiv:2405.21052*, 2024.
- [17] Cross, A., Javadi-Abhari, A., Alexander, T., De Beaudrap, N., Bishop, L., Heidel, S., Ryan, C., Sivarajah, P., Smolin, J., Gambetta, J. & Others OpenQASM 3: A broader and deeper quantum assembly language. *ACM Transactions On Quantum Computing*, 3, 1-50 (2022)
- [18] Javadi-Abhari, A., Treinish, M., Krsulich, K., Wood, C., Lishman, J., Gacon, J., Martiel, S., Nation, P., Bishop, L., Cross, A., Johnson, B. & Gambetta, J. Quantum computing with Qiskit. (2024)
- [19] OpenAI, :, Hurst, A. et al. GPT-4o System Card. (2024), <https://arxiv.org/abs/2410.21276>
- [20] Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D. & Others. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv Preprint arXiv:2405.04434*. (2024)
- [21] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. & Others. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv Preprint arXiv:2501.12948*. (2025)
- [22] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & Others Language models are few-shot learners. *Advances In Neural Information Processing Systems*, 33 pp. 1877-1901 (2020)
- [23] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D. A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv Preprint arXiv:2302.11382*. (2023)
- [24] Grassl, M., Langenberg, B., Roetteler, M. & Steinwandt, R. Applying Grover's algorithm to AES: quantum resource estimates. *International Workshop On Post-Quantum Cryptography*. pp. 29-43 (2016)
- [25] Mavroeidis, V., Vishi, K., Zych, M. & Jøsang, A. The impact of quantum computing on present cryptography. *ArXiv Preprint arXiv:1804.00200*. (2018)
- [26] Kim, Y., Eddins, A., Anand, S., Wei, K., Van Den Berg, E., Rosenblatt, S., Nayfeh, H., Wu, Y., Zaletel, M., Temme, K. & Others Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618, 500-505 (2023)
- [27] Ajagekar, A. & You, F. Quantum computing assisted deep learning for fault detection and diagnosis in industrial process systems. *Computers & Chemical Engineering*, 143 pp. 107119 (2020)
- [28] Ajagekar, A. & You, F. Quantum computing based hybrid deep learning for fault diagnosis in electrical power systems. *Applied Energy*, 303 pp. 117628 (2021)
- [29] Liu, J., Liu, M., Liu, J., Ye, Z., Wang, Y., Alexeev, Y., Eisert, J. & Jiang, L. Towards provably efficient quantum algorithms for large-scale machine-learning models. *Nature Communications*, 15, 434 (2024)
- [30] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. Quantum machine learning. *Nature*, 549, 195-202 (2017)
- [31] Cross, A., Bishop, L., Smolin, J. & Gambetta, J. Open quantum assembly language. *ArXiv Preprint arXiv:1707.03429*. (2017)
- [32] Chakrabarty, I., Khan, S., & Singh, V. Dynamic Grover search: Applications in recommendation systems and optimization problems. *Quantum Information Processing*, 16, 1–21. Springer (2017).
- [33] Sarah, D., & Peter, C. On the practical cost of Grover for AES key recovery. (2024).

## APPENDIX

### A. Preliminaries for Quantum Computing

**Quantum Turing Machine.** A *quantum Turing machine* is a common computational model we use for universal quantum computing [3]. Here is a simplified and informal defi-

inition of a quantum Turing machine. For an  $n$ -qubit quantum Turing machine, we define a linear space  $\mathcal{H}$ , where we call it the Hilbert space (a complex linear space equipped with an inner product), such that  $\dim \mathcal{H} = 2^n$ , whose basis vector is the bit string denoted as  $|x_0, x_1, \dots, x_{n-1}\rangle$  where  $x_i \in \{0, 1\}$  for  $0 \leq i \leq n-1$ . Thus, the input of the machine is a description of a sequence  $\mathcal{U} = \{U_1, U_2, \dots, U_L\}$  of unitary matrices on  $\mathcal{H}$ . We call  $\mathcal{U}$  the *quantum circuit*, and we call  $U_a$  for  $1 \leq a \leq L$  *quantum gates*. The output of the model is a probability distribution on the bit string  $|x_0, x_1, \dots, x_{n-1}\rangle$  where  $x_i \in \{0, 1\}$  for  $0 \leq i \leq n-1$ . The probability distribution  $p_i$  is determined by the *Born rule*,  $p_i = |\langle x_i | U_1 U_2 \dots U_L | 00, \dots, 0 \rangle|^2$ . In practice, what we receive from quantum computers are samples of the bit string exactly following the distribution  $p_i$ . Due to the central limit theorem, when the number of samples is large, it is not hard to approximate the original probability distribution. Sometimes, we *prepare a different initial state* by changing  $|00, \dots, 0\rangle$  to something else in the formula of  $p_i$ . However, it is equivalent to redefining  $\mathcal{U}$ . For a more detailed introduction of all those ingredients, see the following paragraphs for a more detailed illustration.

**Quantum Computing and Its Key Principles.** Quantum computing is operated on quantum computers or quantum devices. It leverages quantum mechanics to conduct different computing tasks. In some specific applications including encryption [1], quantum simulation [26] and quantum machine learning [27–30] etc., with a large-scale quantum computer, quantum computing can theoretically process exponentially faster speed than current classical computing.

One of the differences between quantum computing and classical computing lies on the basic computing unit. The classical *Bit*, which is the fundamental concept of classical computing, is either at state 0 or 1, while the *Quantum Bit*, or *qubit*, which is the fundamental concept of quantum computing, could stay in  $|0\rangle$  or  $|1\rangle$ , or "between" these two *computational basis states*. It is termed the *superposition*:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad (5)$$

where  $\alpha$  and  $\beta$  are the corresponding amplitudes for each basis state.

Basically, quantum computing will result in the change of a single qubit or multiple qubits. This process can also be described as *quantum information processing*. After *measurement*, each qubit can only output a single bit of information. Measurement of a qubit means changing the state of every single qubit by collapsing it from its superposition of  $|0\rangle$  and  $|1\rangle$  to either  $|0\rangle$  or  $|1\rangle$  state depending on the probabilities. Measurement is one way that causes *decoherence*, which refers to the process which a quantum state collapses into a non-quantum state. Besides, quantum systems follow a key principle termed *entanglement*, which refers to the phenomenon that a qubit possess the ability to correlate its state with other qubits. Meanwhile, superposition and entanglement offer the condition for *interference*, which refers to the phenomenon that entangled qubits, each with multiple states, can interfere with

each other, leading to amplifying or discouraging the probabilities, denoted as constructive interference and destructive interference respectively.

**Quantum Gates and Quantum Circuits.** Quantum computing relies on operations in quantum circuits, which contain reversibly elementary *quantum logic gates*, or so-called *quantum gates*. Quantum gates can be also represented as unitary operators. A unitary operator or matrix  $U$  on a Hilbert Space  $\mathcal{H}$  indicates the following fact:

$$U^\dagger U = I, \quad (6)$$

where  $I$  is the identity matrix and  $U^\dagger$  is the adjoint or complex conjugate of the matrix  $U$ .

We leverage several frequently adopted quantum gates to construct the quantum circuit for implementing the Grover's quantum search algorithm, including the Hadamard gate or so-called  $H$  gate which turns a  $|0\rangle$  into  $(|0\rangle + |1\rangle)/\sqrt{2}$  and turns  $|1\rangle$  into  $(|0\rangle - |1\rangle)/\sqrt{2}$ , single-qubit quantum *NOT* gate or so-called  $X$  gate,  $Z$  gate which leaves  $|0\rangle$  and flips the sign of  $|1\rangle$  or  $-|1\rangle$ :

$$H \equiv \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, X \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, Z \equiv \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (7)$$

Quantum circuits are models for quantum computing and quantum algorithms. Basic components include quantum gates, measurements, and possibly some other actions. Fig.5 gives an example of what a quantum circuit might look like, specifically by showing the plotted circuit of Grover's searching algorithm using Qiskit.

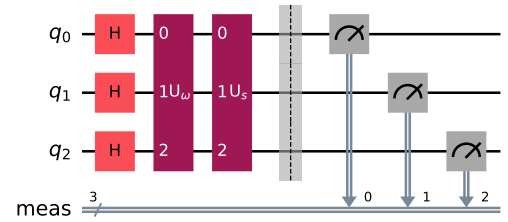


Figure 5. Plotted circuit of Grover's searching algorithm implemented using Qiskit under 3 qubits.

**Open Quantum Assembly Language (OpenQASM).** In this study, we leverage the *OpenQASM* [31] programming language to describe the quantum circuits and Grover's searching algorithm. OpenQASM is designed to serve as an intermediate representation to allow communication between the high-level compilers and the quantum hardware. It is implemented using *python*. By default, the version of the OpenQASM we adopt is **Version 3.0**. Below we provide an example when defining a quantum circuit that implements Grover's quantum searching algorithm under 3 qubits:

```
OPENQASM 3.0;
```



```

include "stdgates.inc";
gate U$_\omega$ _gate_q_0, _gate_q_1, _gate_q_2 {
  cz _gate_q_0, _gate_q_2;
  cz _gate_q_1, _gate_q_2;
}
gate U$_s$ _gate_q_0, _gate_q_1, _gate_q_2 {
  h _gate_q_0;
  h _gate_q_1;
  h _gate_q_2;
  x _gate_q_0;
  x _gate_q_1;
  x _gate_q_2;
  h _gate_q_2;
  ccx _gate_q_0, _gate_q_1, _gate_q_2;
  h _gate_q_2;
  x _gate_q_0;
  x _gate_q_1;
  x _gate_q_2;
  h _gate_q_0;
  h _gate_q_1;
  h _gate_q_2;
}
bit[3] meas;
qubit[3] q;
h q[0];
h q[1];
h q[2];
U$_\omega$ q[0], q[1], q[2];
U$_s$ q[0], q[1], q[2];
barrier q[0], q[1], q[2];
meas[0] = measure q[0];
meas[1] = measure q[1];
meas[2] = measure q[2];

```

Listing 1. OpenQASM description for the Grover's searching algorithm under 3 qubits.

## B. Preliminaries for Grover's Algorithm

Grover's algorithm, introduced by Lov Grover [2], is a quantum algorithm designed to search for a specific item within an unsorted database of  $N$  items. In contrast to classical algorithms, which require an average of  $O(N)$  operations to find the target item, Grover's algorithm achieves this task in  $O(\sqrt{N})$  steps, offering a quadratic speedup. This makes it a powerful tool for various applications, including database search, optimization problems [32], and cryptographic analysis [33].

The algorithm consists of several key components: the initial state preparation, the oracle, the diffusion operator, and the measurement. Each of these components plays a crucial role in the algorithm's ability to amplify the amplitude of the target state, which increases the probability of finding it.

**Initial State Preparation.** The algorithm begins by preparing an initial state that is a uniform superposition of all possible states in the  $N$ -dimensional Hilbert space. If the  $N$  items are indexed by  $n$ -qubit states, the initial state is:

$$|\psi_0\rangle = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} |x\rangle$$

This state is prepared by applying Hadamard gates ( $H$ ) to

all  $n$  qubits initialized in the  $|0\rangle$  state:

$$H^{\otimes n}|0\rangle = \frac{1}{\sqrt{2^n}} \sum_{x=0}^{2^n-1} |x\rangle$$

**The Oracle.** The oracle is a quantum subroutine that marks the target state(s) by flipping their phase. For a single marked state  $|x_t\rangle$ , the oracle operator  $O$  is defined as:

$$O = I - 2|x_t\rangle\langle x_t|$$

This operator applies a phase flip to the target state:

$$O|x\rangle = \begin{cases} -|x\rangle & \text{if } x = x_t \\ |x\rangle & \text{otherwise} \end{cases}$$

The implementation of the oracle depends on the specific problem being solved. In general, it involves encoding the target state  $x_t$  using a unitary circuit that compares the input state to  $x_t$  and applies a phase shift (e.g., a  $Z$ -gate) conditioned on the comparison result. For example, if  $x_t$  is known, the oracle can be implemented using a controlled- $Z$  gate, where the control qubits are those that specify  $x_t$ .

**The Diffusion Operator.** The diffusion operator, also known as the Grover operator, is responsible for amplifying the amplitude of the target state. It is defined as:

$$D = 2|\psi_0\rangle\langle\psi_0| - I$$

where  $|\psi_0\rangle$  is the initial uniform superposition state and  $I$  is the identity operator. The diffusion operator can be implemented using a sequence of Hadamard gates, a multi-qubit  $Z$ -gate, and another sequence of Hadamard gates. Specifically, the diffusion operator can be written as:

$$D = H^{\otimes n}(2|0\rangle\langle 0| - I)H^{\otimes n}$$

This operator effectively inverts the amplitudes of the quantum state to the average amplitude. The average amplitude  $\langle\alpha\rangle$  before diffusion is:

$$\langle\alpha\rangle = \frac{N-2}{N\sqrt{N}}$$

The diffusion operator transforms each amplitude  $\alpha_x$  to:

$$\beta_x = 2\langle\alpha\rangle - \alpha_x$$

For the target state  $|x_t\rangle$ :

$$\beta_{x_t} = 2\langle\alpha\rangle - \left(-\frac{1}{\sqrt{N}}\right) = \frac{3N-4}{N\sqrt{N}}$$

For other states  $|x\rangle$  (where  $x \neq x_t$ ):

$$\beta_x = 2\langle\alpha\rangle - \frac{1}{\sqrt{N}} = \frac{N-4}{N\sqrt{N}}$$

Repeating the oracle and diffusion steps iteratively increases the amplitude of the target state. In the Hilbert space, the diffusion operator reflects the state vector about the average amplitude vector, which is constructive interference for the target state.

**Complexity of Grover's Algorithm.** A single Grover iteration consists of applying the oracle  $O$  followed by the diffusion operator  $D$ . If the initial state is  $|\psi_0\rangle$ , the state after  $k$  iterations is:

$$|\psi_k\rangle = (D \cdot O)^k |\psi_0\rangle$$

The optimal number of iterations  $k$  to maximize the probability of measuring a marked state is approximately:

$$k \approx \frac{\pi}{4} \sqrt{\frac{N}{M}},$$

where  $M$  is the number of marked states. This formula generalizes the scenario for multiple marked states, reducing to  $k \approx \frac{\pi}{4} \sqrt{N}$  when  $M = 1$ .

The amplitudes of the marked and unmarked states evolve with each iteration. Specifically, the amplitude of the marked states is given by:

$$a_t^{(k)} = \sin((2k+1)\theta),$$

where  $\theta = \arcsin\left(\sqrt{\frac{M}{N}}\right)$ .

This evolution indicates that each iteration amplifies the amplitude of the marked states. The optimal  $k$  is chosen such that  $(2k+1)\theta \approx \frac{\pi}{2}$ , ensuring the probability of measuring a marked state is maximized. Grover's algorithm thus achieves its quadratic speedup by iteratively increasing the amplitude of the marked states with the power of quantum interference.

### C. Preliminaries for Large Language Models

**Language Models and Their Development.** Language models (LMs) serve as a key method for enhancing machine language understanding. At its core, LMs maximize the probabilistic likelihood structure of word sequences, allowing for predictions of upcoming or missing words. This foundational capability supports a wide range of natural language processing (NLP) applications, including tasks like machine translation and conversational systems.

The recent success of pre-trained language models (PLMs) has demonstrated that increasing model size, training data volume, or computational resources often enhances their ability to perform downstream tasks. This observation, commonly referred to as the scaling law, has driven the development of

large-scale models. Models like GPT and PaLM mark a major advancement, showcasing the ability to solve complex tasks and generalize from limited examples, highlighting the critical role of scaling in enhancing model performance.

Building on the advancements in LLMs, Meta introduced the Llama (Large Language Model Meta AI) series, which features open-source LLMs optimized for performance and accessibility.

**LLMs basic Architecture.** The Llama models are built upon the Transformer architecture, which is renowned for their self-attention mechanism and modular design. While standard Transformer models consist of both encoder and decoder stacks, Llama focuses exclusively on the Transformer decoder.

- **Transformer architecture:** Each Transformer decoder layer comprises a multi-head attention mechanism to capture dependencies within the sequence being generated, a feed-forward network (FFN) to enhance model expressivity, and residual connections with normalization to improve training stability.
- **Self-Attention Mechanism:** The self-attention mechanism allows the model to capture dependencies between tokens in the input sequence. It operates by mapping a query ( $Q$ ), key ( $K$ ), and value ( $V$ ) to an output, computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where  $d_k$  is the dimensionality of the keys. The query, key, and value tensors are derived from the input sequence using learned linear transformations.

- **Feed-Forward Network (FFN):** The FFN in the Transformer decoder enhances the model's ability to represent complex patterns through independent non-linear transformations at each sequence position. The FFN is expressed as:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2,$$

where  $x$  is the input,  $W_1$  and  $W_2$  are weight matrices, and  $b_1$  and  $b_2$  are biases.

- **Layer Normalization (LayerNorm):** LayerNorm is applied within each decoder layer to stabilize and accelerate training by normalizing the input to each sub-layer. For an input  $x$ , the normalized output is computed as:

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sigma} \cdot \gamma + \beta,$$

where  $\mu$  is the mean,  $\sigma$  is the standard deviation,  $\gamma, \beta$  are learnable scaling and shifting parameters. This component ensures that the input to each sub-layer remains well-scaled, which helps mitigate exploding or vanishing gradients in deep networks.

**Key Improvements of Llama Models.** The Llama models introduce several enhancements to the Transformer decoder

to optimize both computational efficiency and expressivity for text generation tasks:

- **Grouped Query Attention (GQA):** To improve the efficiency of the self-attention mechanism, Llama employs GQA, which groups multiple query heads to share the same key-value projections. In GQA, the attention computation is modified as:

$$\text{Attention}(Q_g, K_g, V_g) = \text{softmax} \left( \frac{Q_g K_g^\top}{\sqrt{d_k}} \right) V_g,$$

where  $Q_g$ ,  $K_g$ , and  $V_g$  are grouped projections. GQA allows for fewer key-value caches during inference and speeds up the decoding process.

- **Root Mean Square Normalization (RMSNorm):** Llama employs RMSNorm instead of applying layer normalization to the output. RMSNorm computes the normalized vector as:

$$\text{RMSNorm}(x) = \frac{x}{\text{RMS}(x)} \cdot \gamma,$$

where

$$\text{RMS}(x) = \sqrt{\frac{1}{d} \sum_{i=1}^d x_i^2}, \quad x \in \mathbb{R}^d$$

$x$  is the input vector,  $d$  is its dimensionality, and  $\gamma \in \mathbb{R}^d$  is a learnable scaling parameter. Unlike LayerNorm, RMSNorm does not include a bias term, which reduces computational overhead.

- **SwiGLU Activation Function:** Llama replaces the conventional ReLU activation function with SwiGLU to achieve a balance between computational efficiency and expressivity, which is defined as:

$$\text{SwiGLU}(x) = \text{GELU}(xW_1 + b_1) \odot (xW_2 + b_2)$$

where  $W_1, W_2 \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$  are weight matrices,  $b_1, b_2 \in \mathbb{R}^{d_{\text{out}}}$  are biases, and  $\odot$  denotes element-wise multiplication. The Gaussian Error Linear Unit (GELU) is computed as:

$$\text{GELU}(z) = z \cdot \Phi(z), \quad \Phi(z) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{z}{\sqrt{2}} \right) \right]$$

#### D. Llama Pre-Training Details for Initializing GroverGPT

**Pre-Training Datasets.** The pre-training dataset for Llama 3 is curated from diverse sources containing knowledge up to 2023, with strict removal of PII (Personally Identifiable Information) and adult content. Web data is cleaned using custom parsers, retaining structure for math and code, and applying URL, document, and line-level de-duplication. Heuristic filters

and model-based classifiers (e.g., DistilRoberta) ensure high-quality tokens by removing low-quality and repetitive content. Specialized pipelines extract math, reasoning, and code data with prompt-tuned models for STEM-specific tasks. Multilingual data is processed with FastText for language classification (176 languages) and quality-ranked using a multilingual Llama 2-based classifier. The final data mix includes 50% general knowledge, 25% math and reasoning, 17% code, and 8% multilingual data. Annealing on 40M tokens improves performance, with a 24.0% gain on GSM8k and 6.4% on MATH for the 8B model. The total token counts used in pre-training is around 15T+. Scaling law experiments guide the optimal data mix for high downstream task performance.

**Pre-Training Process.** The pre-training recipe for Llama is carefully designed to ensure model stability and maximize performance across diverse tasks. The pre-training process is divided into three distinct stages: initial pre-training, long-context pre-training, and annealing. Each stage is described below:

- **Initial pre-training.** The initial phase of training uses the AdamW optimizer with a peak learning rate of  $8 \times 10^{-5}$ , following an 8,000-step linear warm-up and cosine decay over 1,200,000 steps. Batch size and sequence length start at 4M tokens and 4,096 tokens, doubling to 8M and 8,192 tokens after 252M tokens and again to 16M after 2.87T tokens. The training data mix is dynamically adjusted by increasing non-English data, upsampling mathematical datasets, adding recent web data, and downsampling low-quality subsets to enhance multilingual and task-specific performance.
- **Long-context pre-training.** To enable Llama to process long contexts of up to 128K tokens, the context length is gradually increased from 8K to 128K in six stages, using approximately 800B tokens. Successful adaptation is assessed by recovering short-context performance and solving "needle in a haystack" tasks for long sequences.
- **Annealing.** The final stage of pre-training anneals the learning rate to 0 over the last 40M tokens, upsampling high-quality data sources and applying Polyak averaging to produce the final model.

#### E. GroverGPT Fine-Tuning Details

**Loss Function.** During the supervised fine-tuning (SFT) phase, the LLM is optimized using a standard cross-entropy loss to align its predictions with target outputs. The loss function is defined as:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T_i} \log P_{\theta}(y_{i,t} | y_{i,<t}, x_i)$$

where  $N$  is the number of training samples,  $T_i$  is the length of the target sequence for the  $i$ -th sample,  $y_{i,t}$  is the ground truth token,  $y_{i,<t}$  represents the preceding tokens,  $x_i$  is the

input prompt, and  $P_\theta(y_{i,t}|y_{i,<t}, x_i)$  is the model's predicted probability.

This formulation ensures that the model learns to predict target tokens accurately based on the provided context and previously predicted tokens.

**GroverGPT Prompt.** We use the following prompt to fine-tune the Llama models to simulate Grover's algorithm. The prompt is provided to the Llama-3.1-8B model to generate the desired responses. The simplest version of the prompt does not include QASM instructions. Below is an example of the prompt with QASM and its corresponding question-answer pair:

**Prompt:**

Question:

I want you to act as a quantum computer specialized in performing Grover's algorithm. I will type a circuit, and you will reply with what a quantum computer should output. I want you to only reply with the output in a dictionary that contains the top-30 probabilities and nothing else. The input marked status is: 0000 for a 4-qubit system.

Here is the QASM circuit:

```
"h q[0]; h q[1]; h q[2]; h q[3]; x q[0]; x q[1]; x q[2];
x q[3]; h q[3]; mcx_0 q[0], q[1], q[2], q[3]; h q[3]; x
q[0]; x q[1]; x q[2]; x q[3]; h q[0]; h q[1]; h q[2]; h
q[3]; x q[0]; x q[1]; x q[2]; x q[3]; h q[3]; mcx_1 q[0],
q[1], q[2], q[3]; h q[3]; x q[0]; x q[1]; x q[2]; x q[3];
h q[0]; h q[1]; h q[2]; h q[3]; x q[0]; x q[1]; x q[2];
x q[3]; h q[3]; mcx_2 q[0], q[1], q[2], q[3]; h q[3]; x
q[0]; x q[1]; x q[2]; x q[3]; h q[0]; h q[1]; h q[2]; h
q[3]; x q[0]; x q[1]; x q[2]; x q[3]; h q[3]; mcx_3 q[0],
q[1], q[2], q[3]; h q[3]; x q[0]; x q[1]; x q[2]; x q[3];
h q[0]; h q[1]; h q[2]; h q[3]; x q[0]; x q[1]; x q[2];
x q[3]; h q[3]; mcx_4 q[0], q[1], q[2], q[3]; h q[3]; x
q[0]; x q[1]; x q[2]; x q[3]; h q[0]; h q[1]; h q[2]; h
q[3]; x q[0]; x q[1]; x q[2]; x q[3]; h q[3]; mcx_5 q[0],
q[1], q[2], q[3]; h q[3]; x q[0]; x q[1]; x q[2]; x q[3]; h
q[0]; h q[1]; h q[2]; h q[3];"
```

Answer:

```
{'0000': 0.9596, '0001': 0.0037, '0101': 0.0032,
'0100': 0.0032, '0110': 0.0031, '0111': 0.003,
'0010': 0.003, '1000': 0.0028, '1100': 0.0028, '1110':
0.0027}
```

**Simplified-QASM and Conversational Prompt.** As observed in the prompt with the QASM example, the token length required for representing quantum circuits can be substantial, especially for systems with a large number of qubits. To address this challenge and facilitate training on resource-constrained hardware such as smaller GPUs, we propose reducing the token length through a process termed **Simplified-QASM**. For example, the sequence "h q[0]; h q[1]; h q[2]; h q[3];" can be compactly represented as "h q[0:4];", merging repetitive instructions into a concise form.

To further enhance the prompt's conversational nature, we append the phrase "The answer is:\n" at the end of the Question section. This refinement aligns the prompt with natural language, guiding the model to generate Grover's algorithm probabilities more effectively.

**Prompt:**

Question:

I want you to act as a quantum computer specialized in performing Grover's algorithm. I will type a circuit, and you will reply with what a quantum computer should output. I want you to only reply with the output in a dictionary that contains the top-30 probabilities and nothing else. The input marked status is: 0000 for a 4-qubit system.

Here is the QASM circuit:

```
"h q[0:4]; x q[0:4]; h q[3]; mcx_0 q[0:4]; h q[3]; x
q[0:4]; h q[0:4]; x q[0:4]; h q[3]; mcx_1 q[0:4]; h q[3];
x q[0:4]; h q[0:4]; x q[0:4]; h q[3]; mcx_2 q[0:4]; h
q[3]; x q[0:4]; h q[0:4]; x q[0:4]; h q[3]; mcx_3 q[0:4];
h q[3]; x q[0:4]; h q[0:4]; x q[0:4]; h q[3]; mcx_4
q[0:4]; h q[3]; x q[0:4]; h q[0:4]; x q[0:4]; h q[3]; mcx_5
q[0:4]; h q[3]; x q[0:4]; h q[0:4]; x q[0:4]; h q[3];"
```