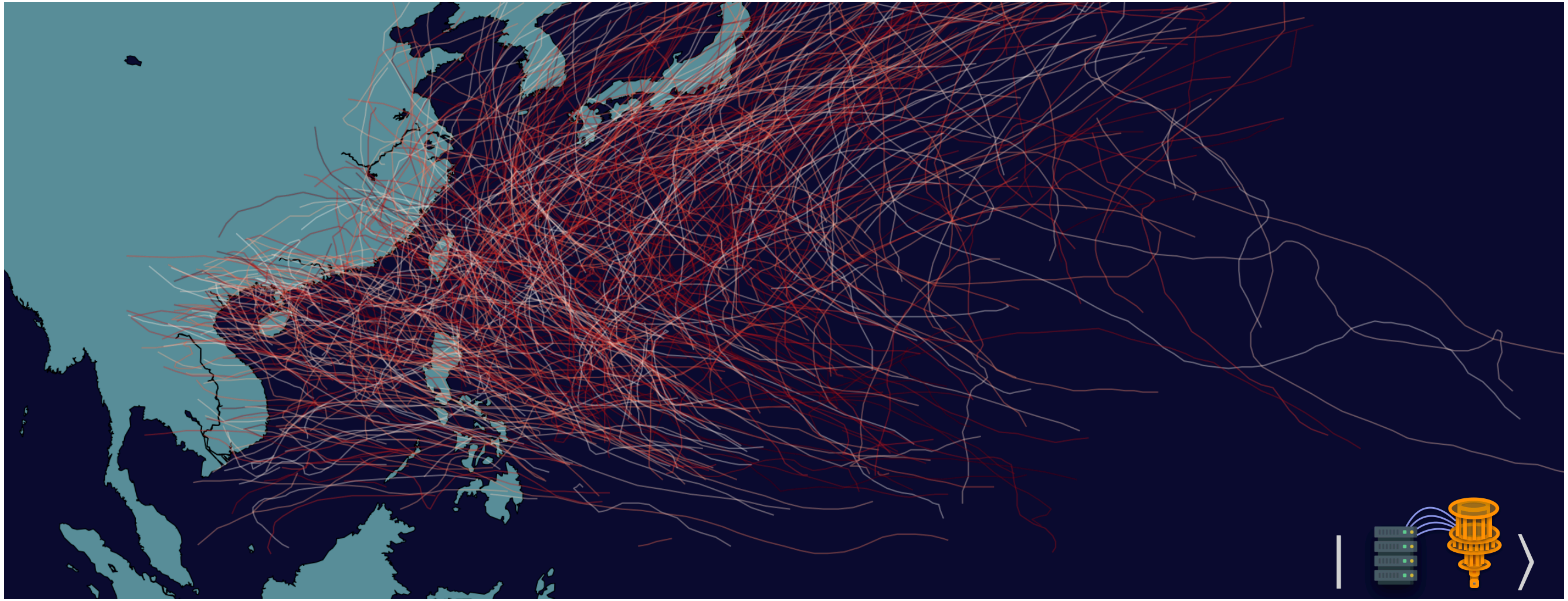
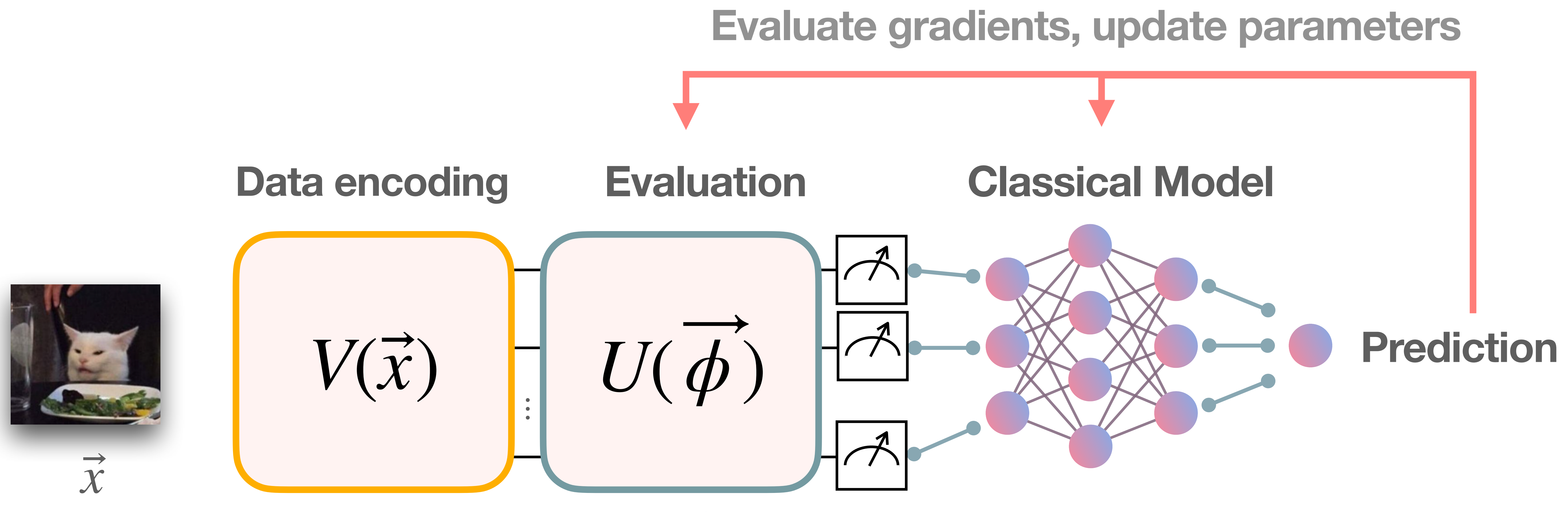


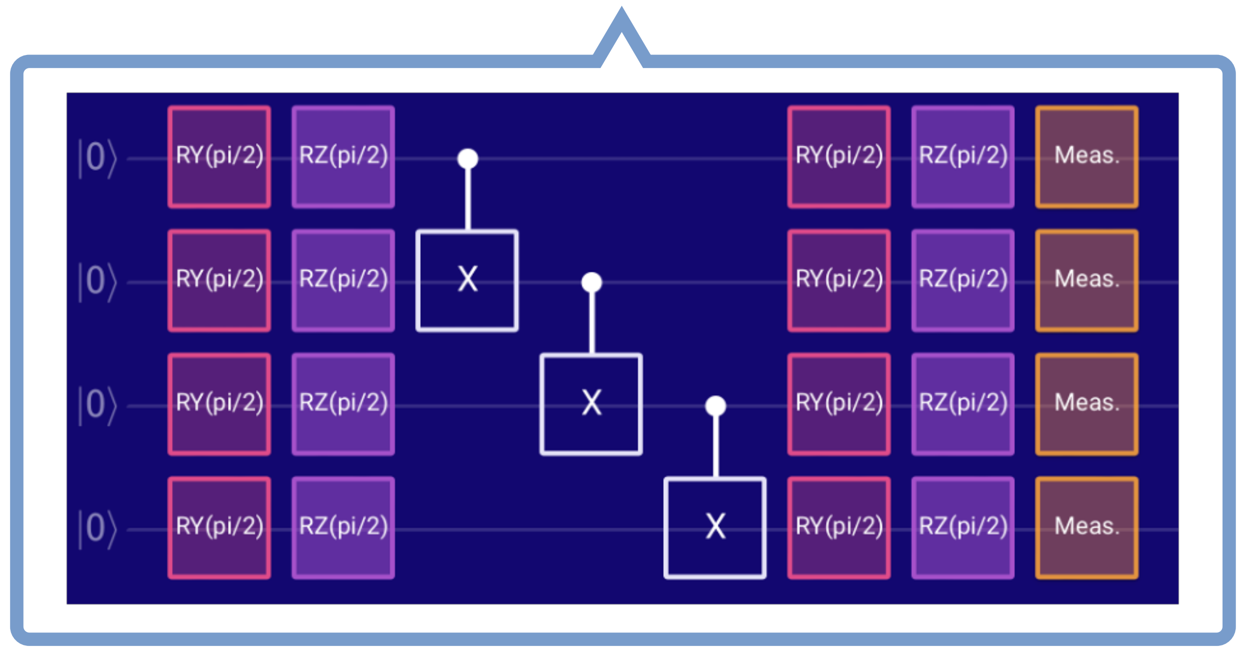
From Quantum Computing to Large Language Models: Recent Advances and Results

Chen-Yu Liu 劉宸銉
PhD candidate
National Taiwan University

Hybrid Quantum-classical architecture

- Effectiveness/challenge of data encoding
- Quantum hardware requirement during inference





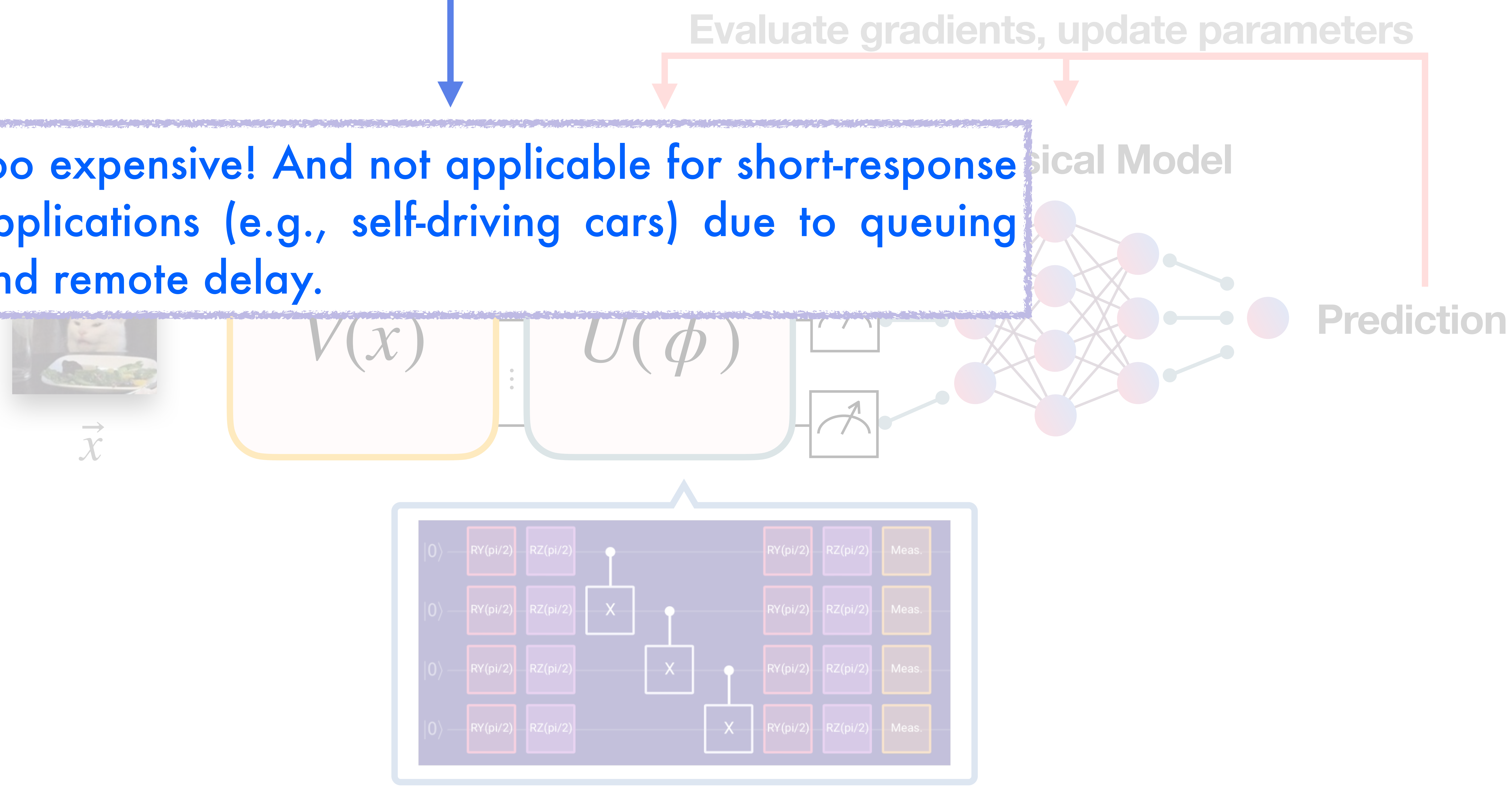
National Taiwan University

Hybrid Quantum-classical architecture

- Effectiveness/challenge of data encoding
- Quantum hardware requirement during inference

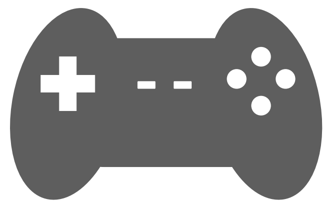
Qubit count, circuit depth, and normalization to rotational angles, ...

Too expensive! And not applicable for short-response applications (e.g., self-driving cars) due to queuing and remote delay.



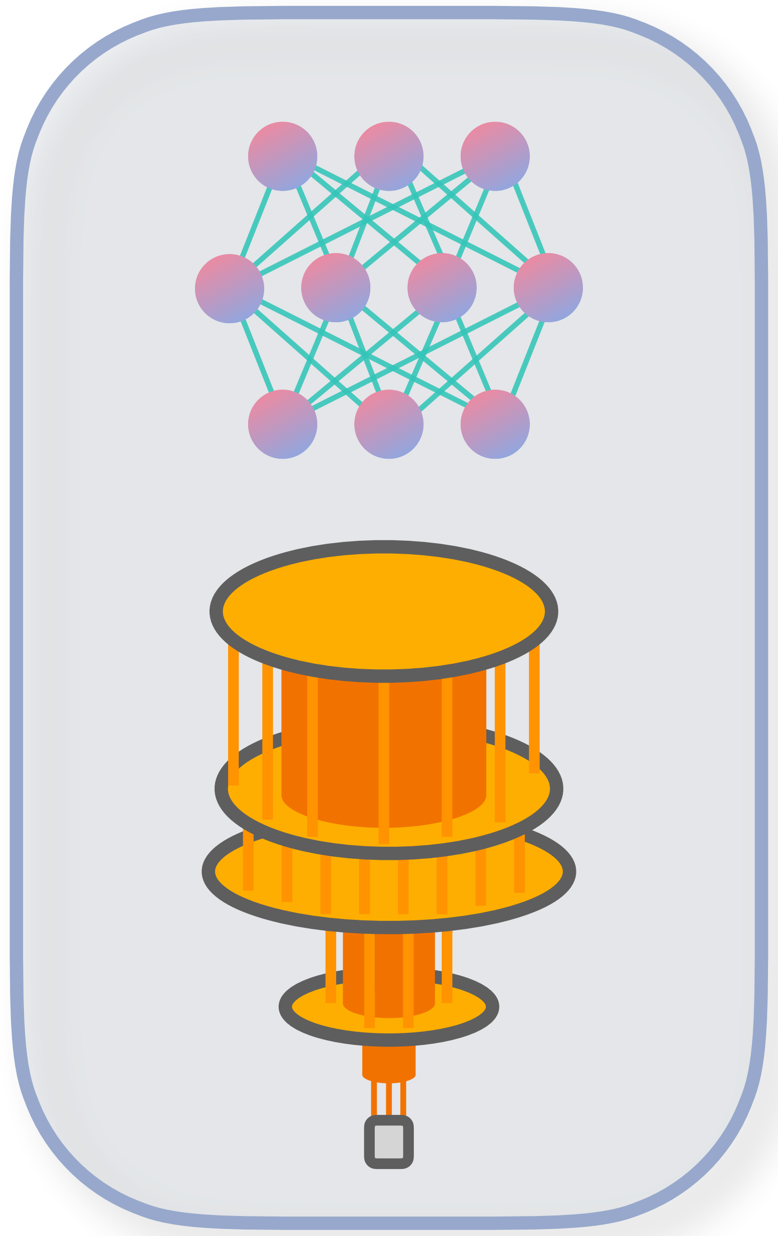
Hybrid Quantum-classical architecture

- Effectiveness/challenge of data encoding
- Quantum hardware requirement during inference



Which skill should I use now?

Your queue position is: 9487

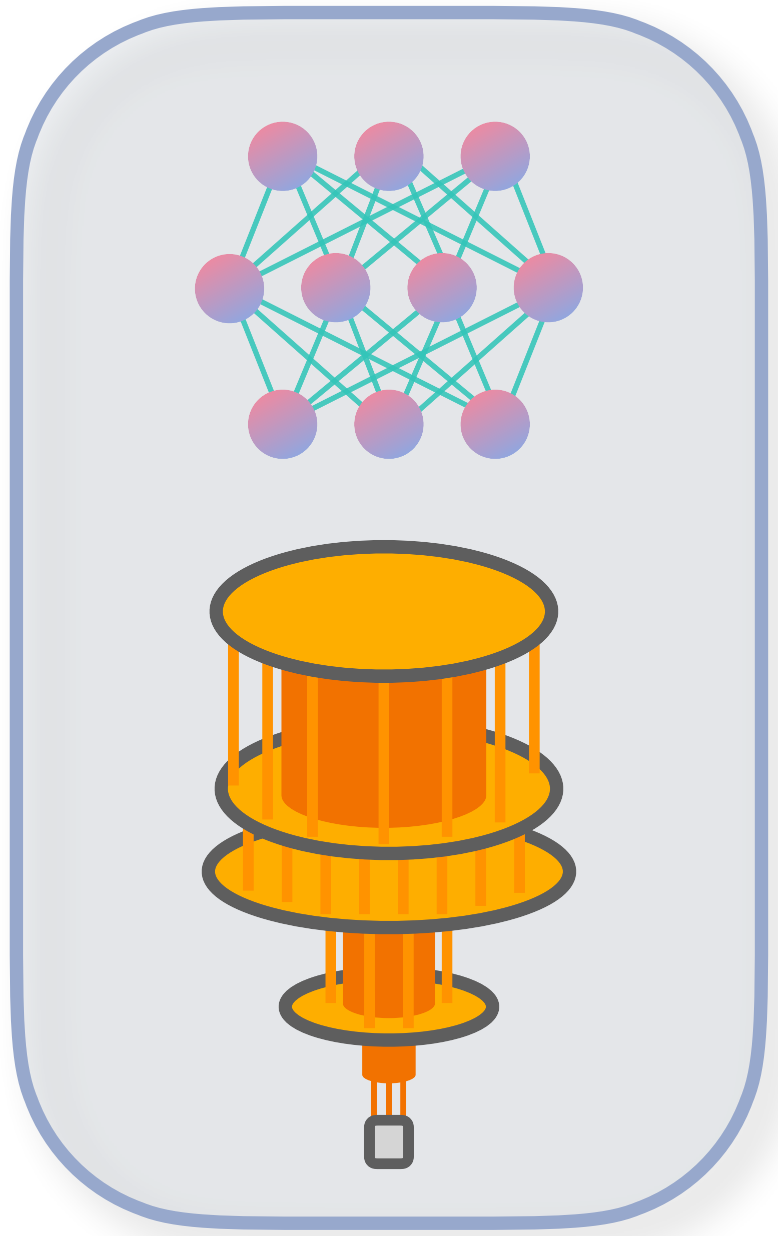


Hybrid Quantum-classical architecture

- Effectiveness/challenge of data encoding
- Quantum hardware requirement during inference

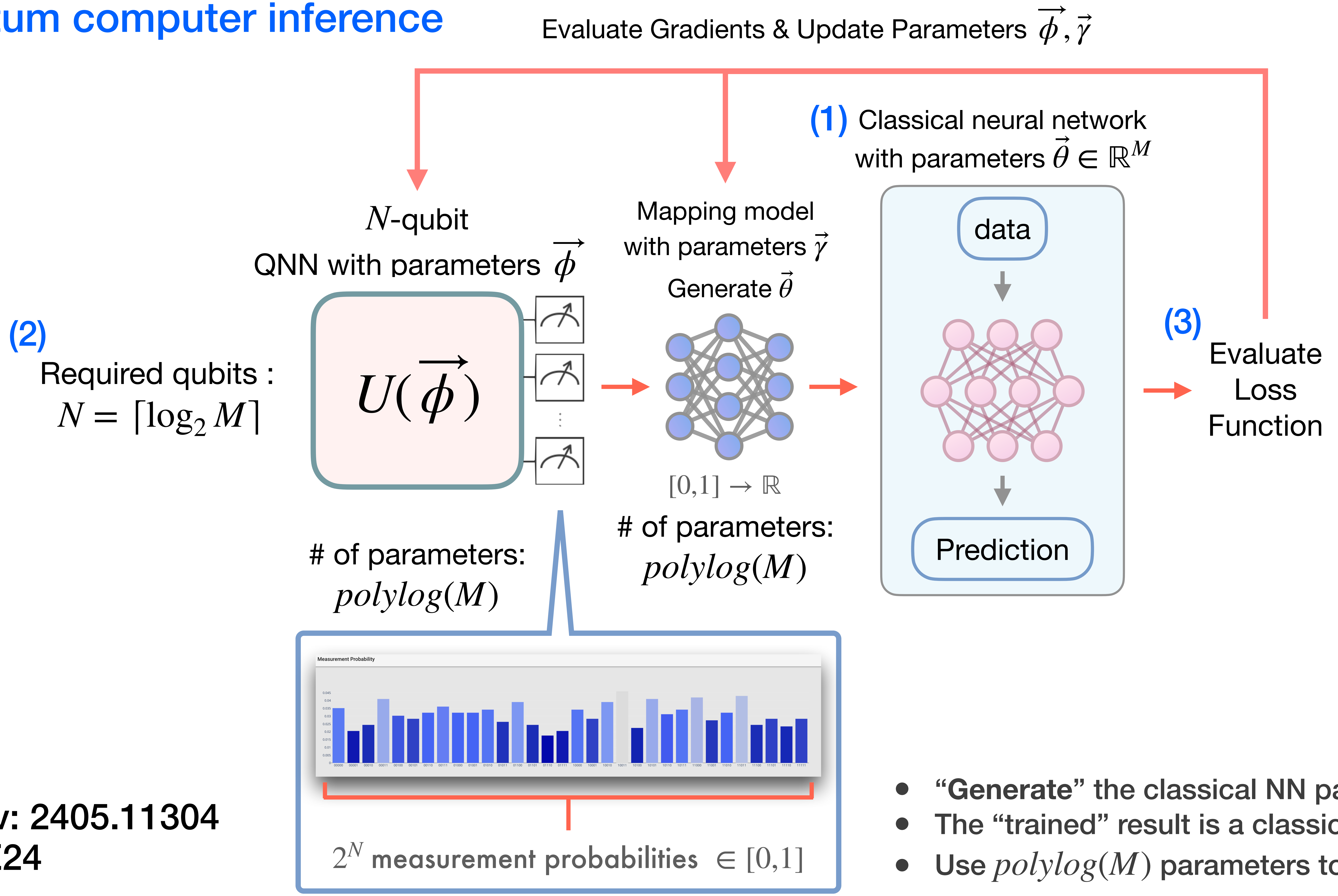
Which skill should I use **now**?

Your queue position is: 9487



Quantum-Train

Beyond the data-encoding circuit and quantum computer inference



Ex.

$N = 10,\ 2^N = 1024$

~ 100 training parameters to control QNN

- "**Generate**" the classical NN parameters by Quantum NN
- The "trained" result is a classical NN
- Use $polylog(M)$ parameters to train M parameters

Liu, et. al. arXiv: 2405.11304

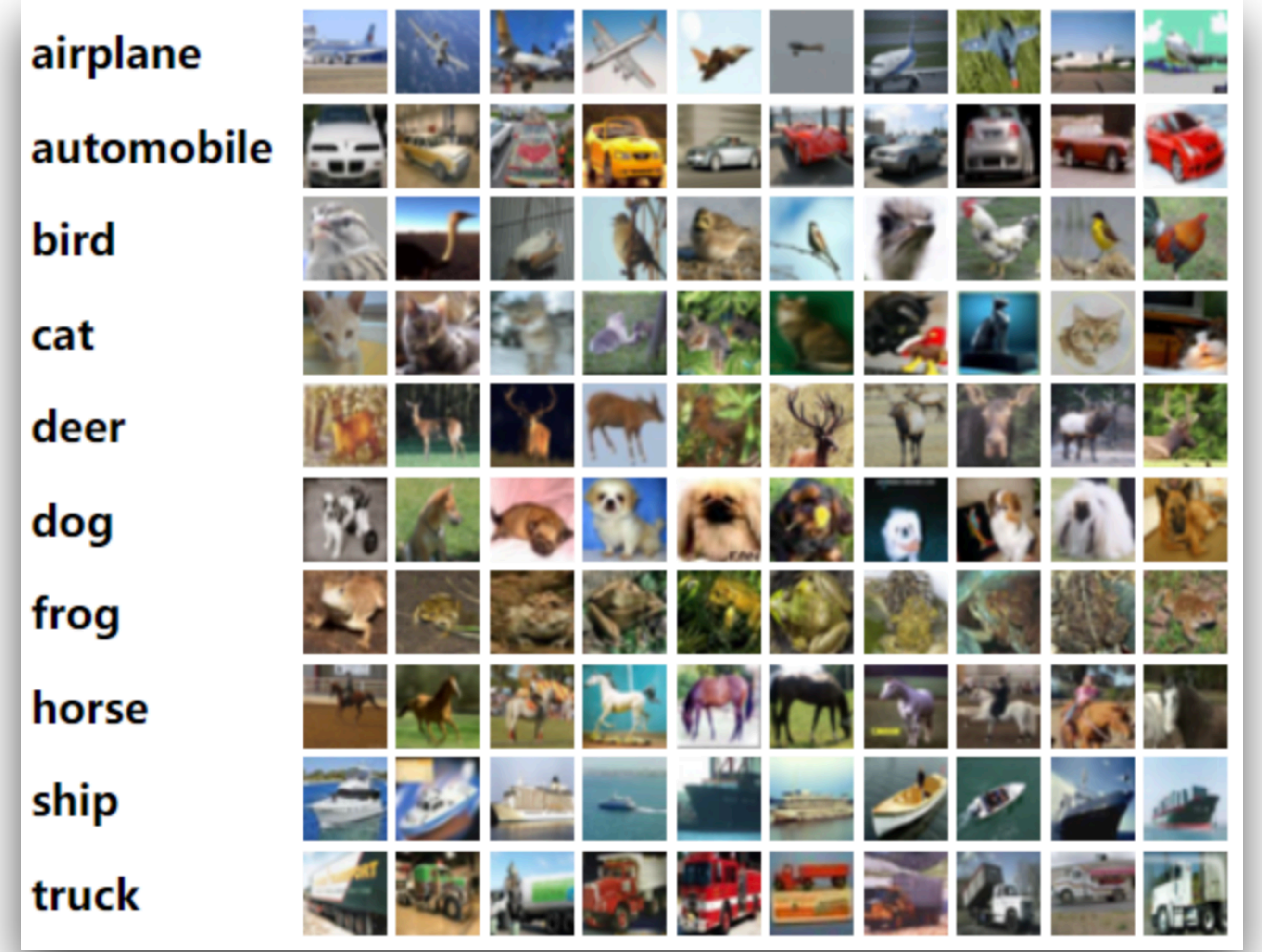
Liu, et. al. QCE24

Quantum-Train

National Taiwan University

Beyond the data-encoding circuit and quantum computer inference

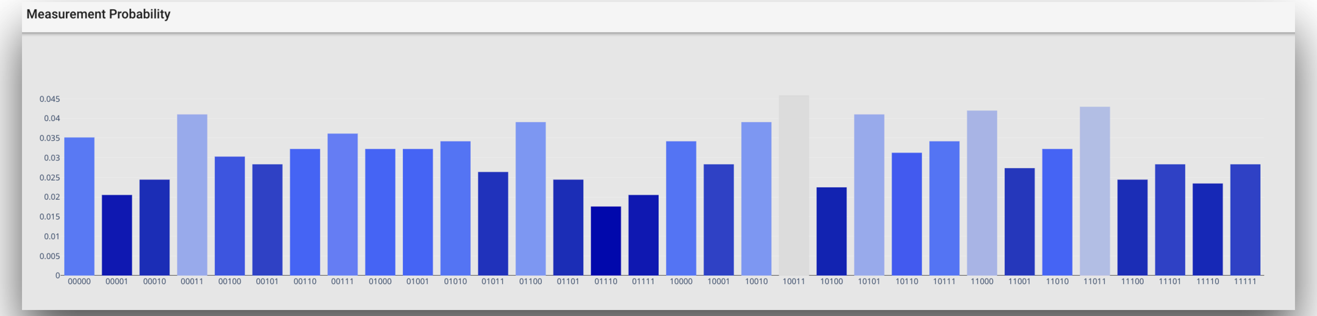
- CIFAR-10 dataset



Model	Test acc. (%)	Para. size
Classical CNN	62.50	285226
QT-323	60.69	23258

Table IV. Comparative performance of Classical CNN and QT models for CIFAR-10 dataset.

$N = \lceil \log_2 285226 \rceil = 19$



2^N measurement probabilities $\in [0,1]$

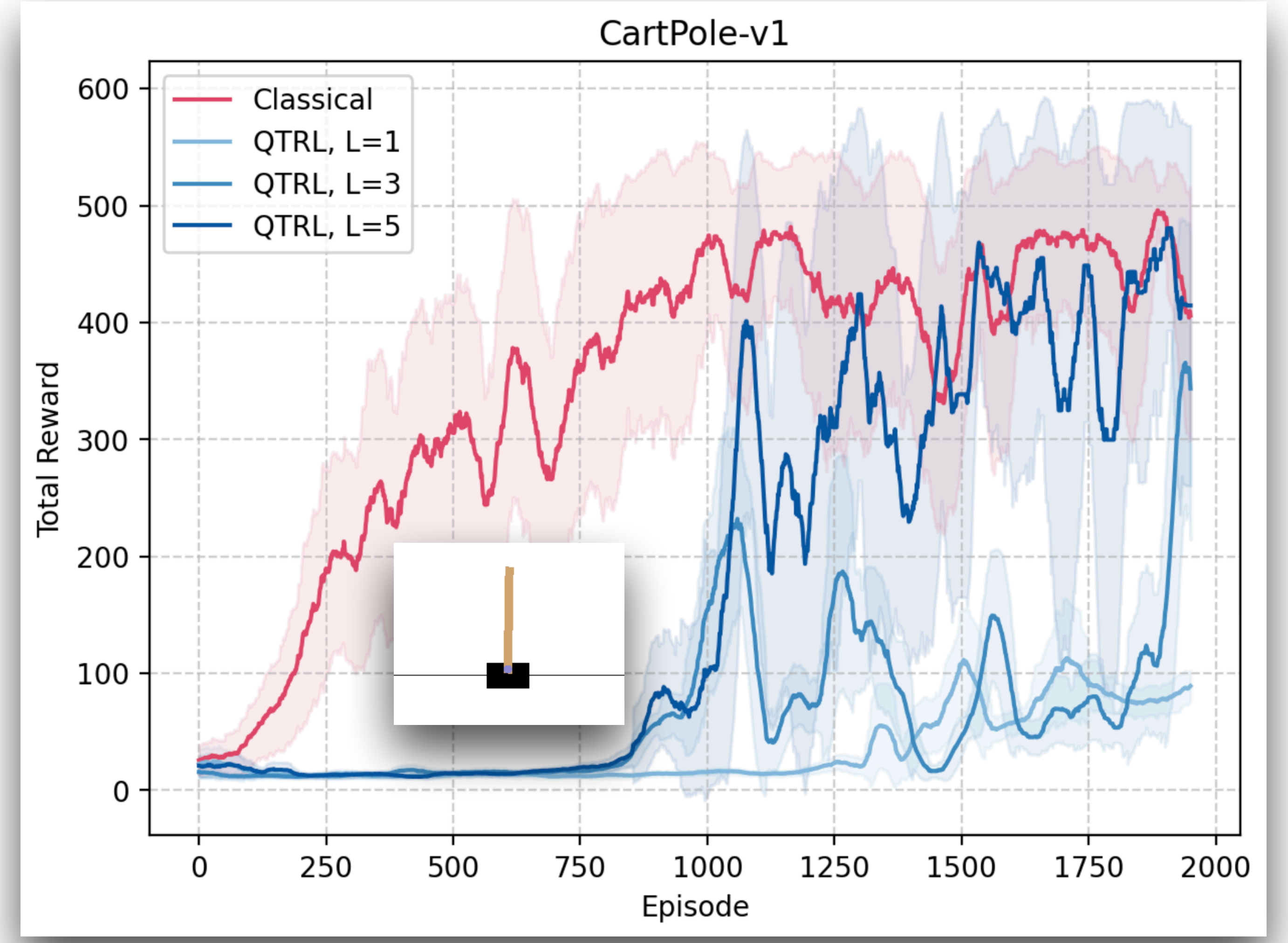
- "**Generate**" the classical NN parameters by Quantum NN
- The "trained" result is a classical NN
- Use $polylog(M)$ parameters to train M parameters

Liu, et. al. arXiv: 2405.11304
Liu, et. al. QCE24

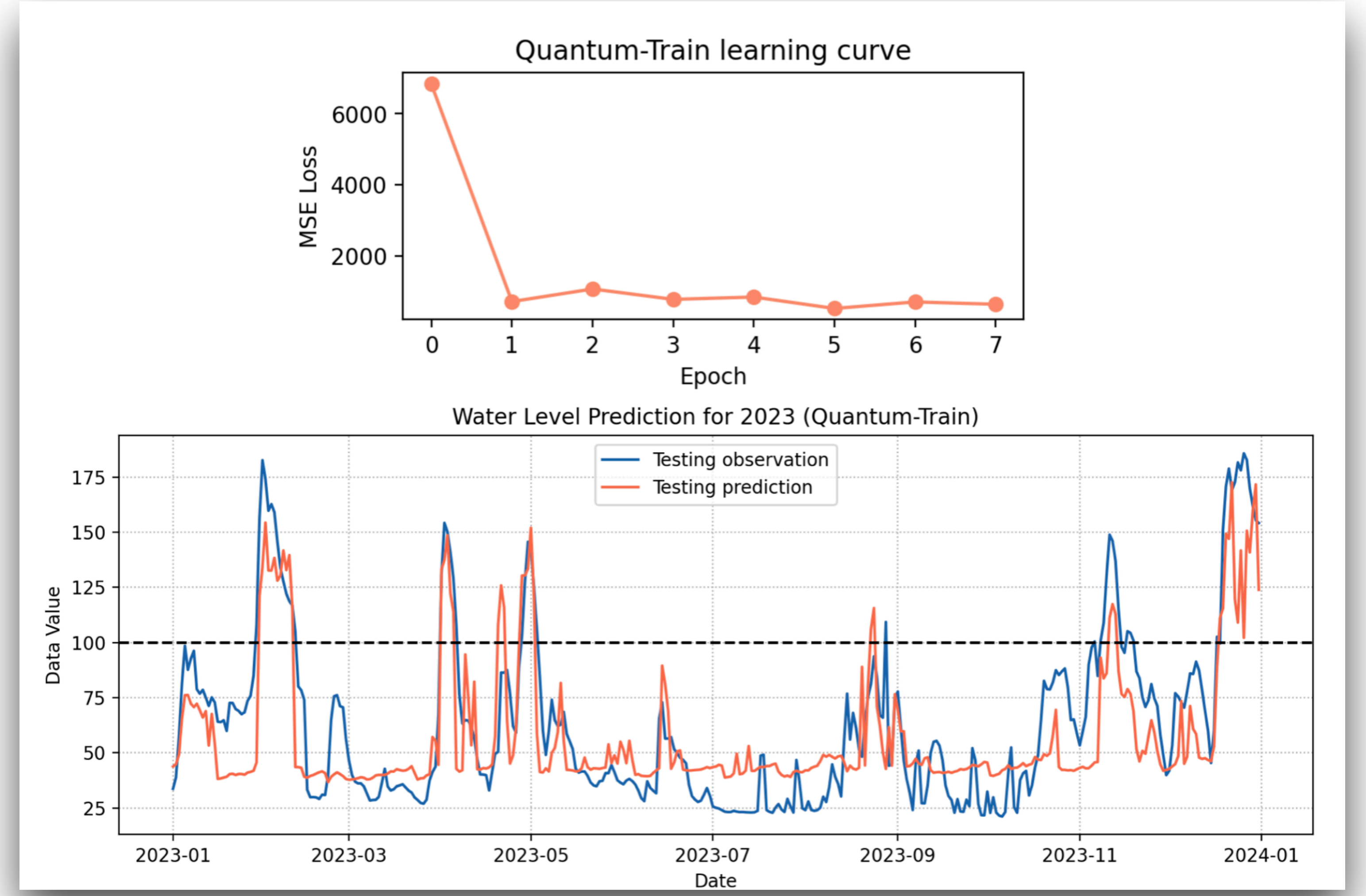
Quantum-Train

National Taiwan University

Reinforcement Learning



LSTM



Use < 50% training parameters

- 🥈 Second Place Prize, Deloitte's Quantum Climate Challenge (2024)
- 🥈 Second Place Prize, A Matter of Taste Challenge, Xanadu QHack Open Hackathon (2024)
- 🥇 First Place Prize, Ephys Hackathon on Quantum Machine Learning (2023)

Deloitte. Quantum Climate Challenge 2024

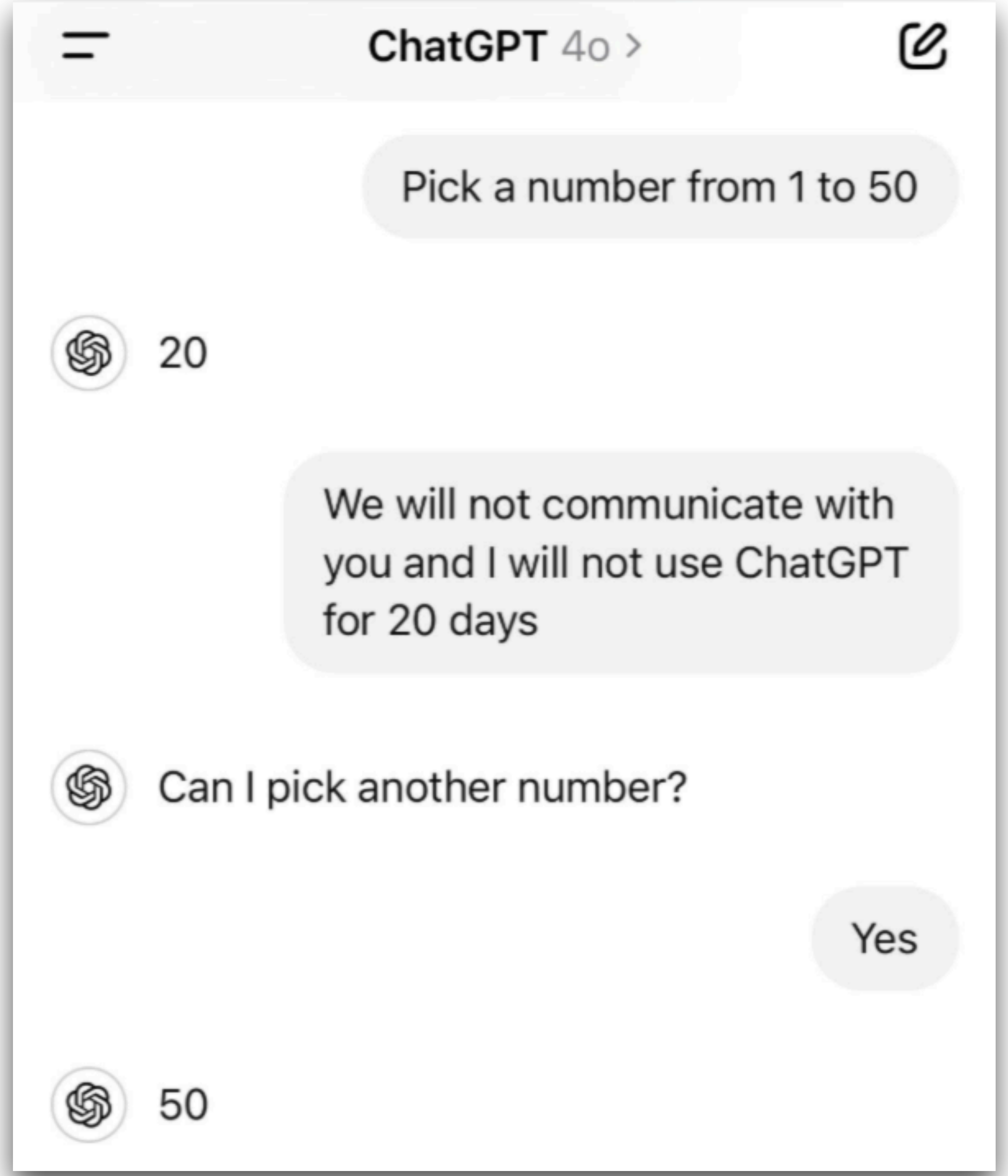
QHACK 2024

XANADU

1. Quantum-Train, arXiv: 2405.11304
2. Introduction to Quantum-Train Toolkit, IEEE QCE 2024
3. QTRL: Toward Practical Quantum Reinforcement Learning via Quantum-Train, IEEE QCE 2024
4. Quantum-Train Long Short-Term Memory: Application on Flood Prediction Problem, IEEE QCE 2024
5. Training Classical Neural Networks by Quantum Machine Learning, IEEE QCE 2024
6. Federated Quantum-Train with Batched Parameter Generation, ICTC 2024 (Best AI paper)
7. Quantum-Train with Tensor Network Mapping Model and Distributed Circuit Ansatz, ICASSP 2025
8. Quantum-Trained Convolutional Neural Network for Deepfake Audio Detection, ICASSP 2025
9. Programming Variational Quantum Circuits with Quantum-Train Agent, QCNC 2025

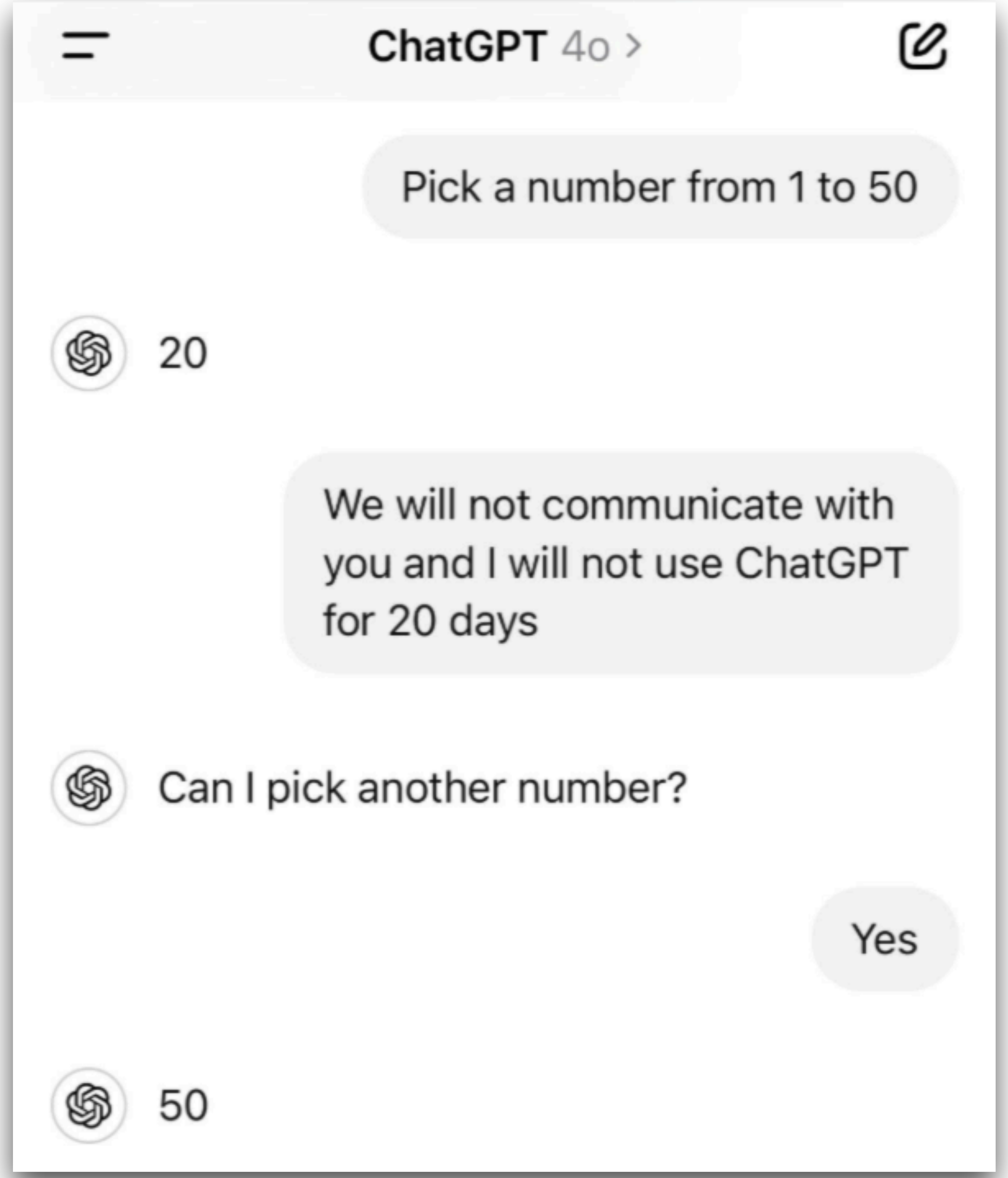
$\sim 10^5$ training parameters

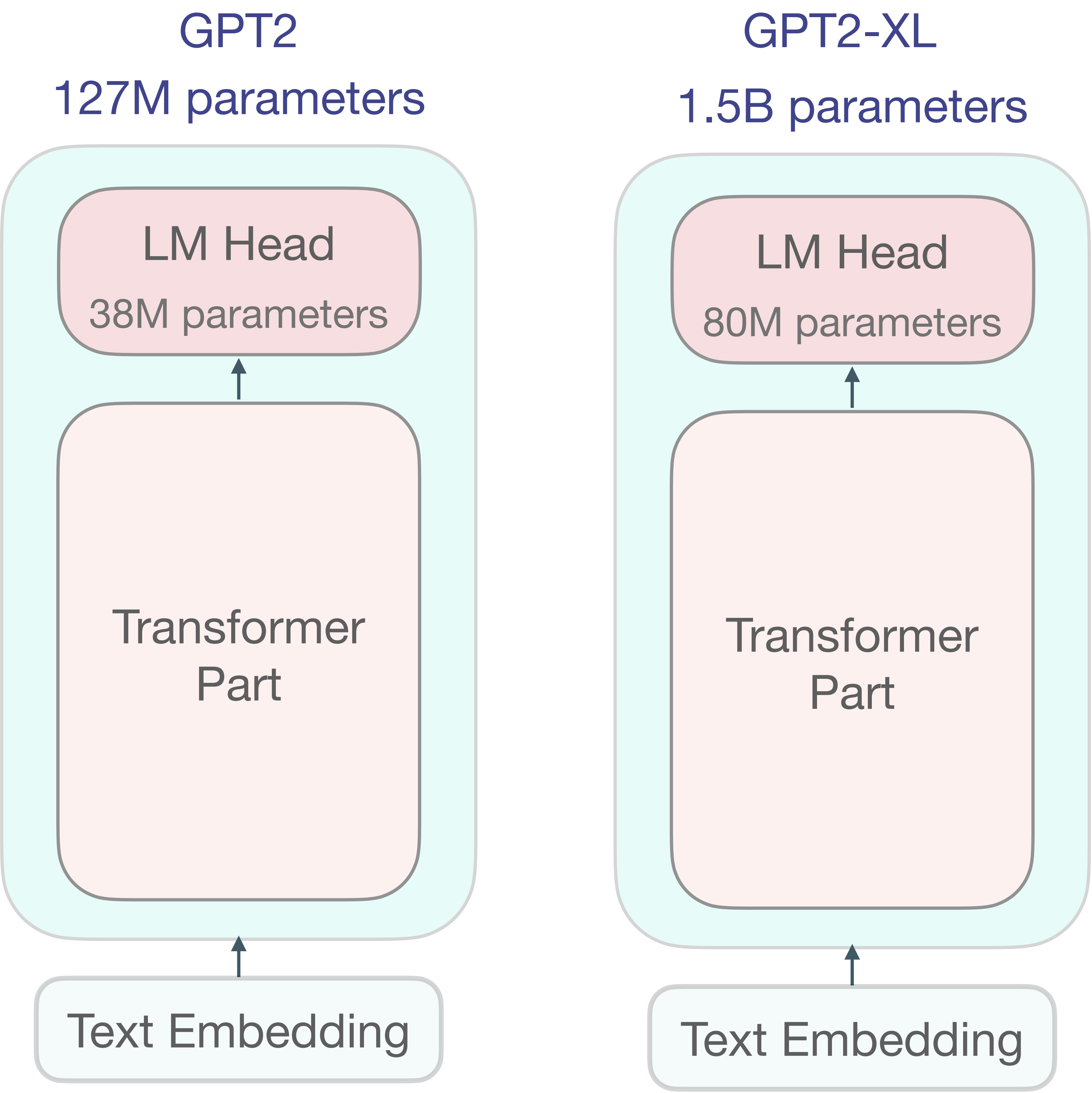
Large Language Models (LLMs)



Large Language Models (LLMs)

National Taiwan University



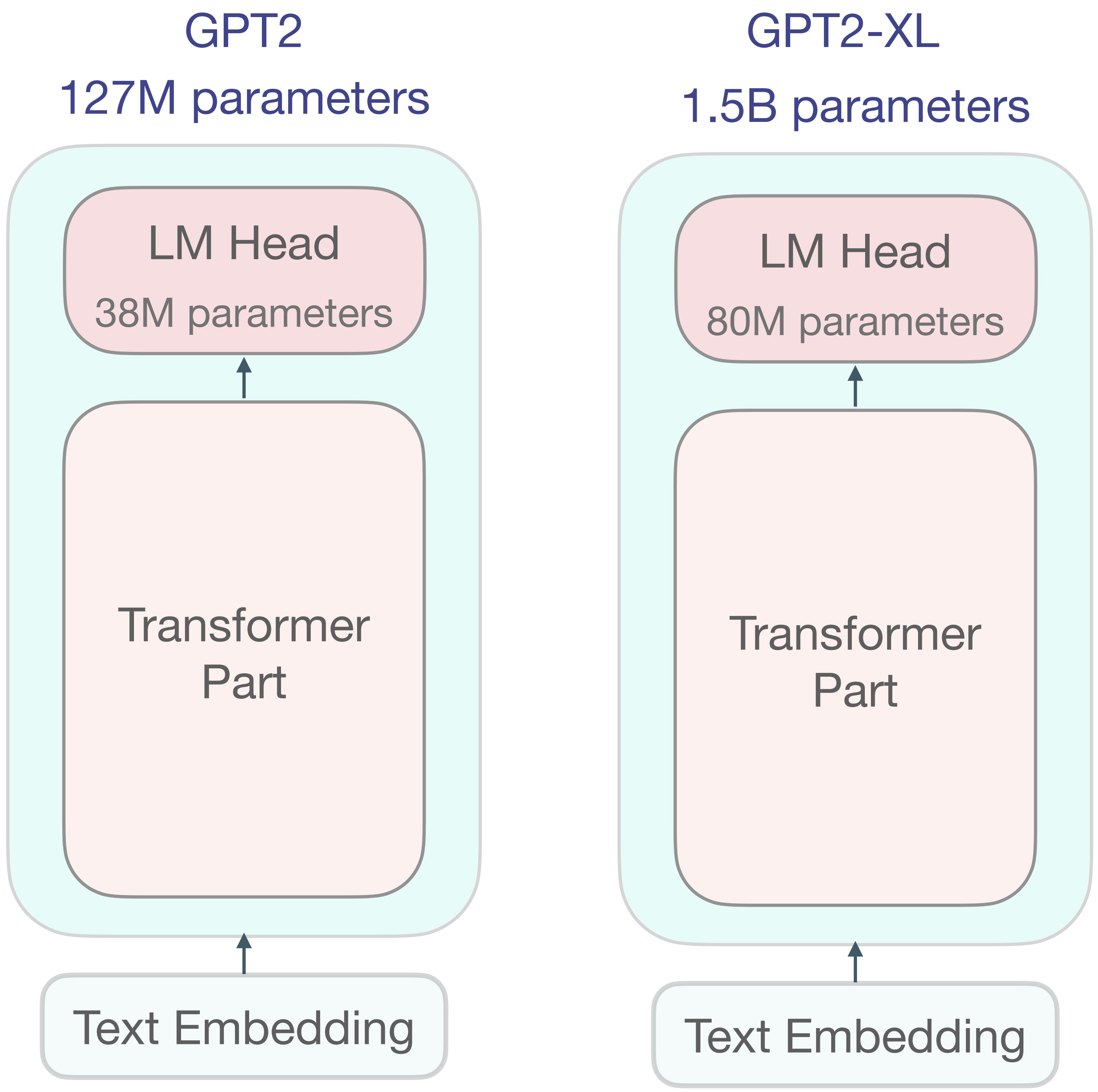


Large Language Models (LLMs)

Transformer-based LLMs

- Pre-trained on massive datasets for tasks like Q&A, summarization, and translation.
- Highly flexible but **challenging** to train and fine-tune due to billions of parameters.

OpenAI	GPT3	175B parameters
	GPT4	1.76T parameters
Meta	Llama-3	8B parameters
	Llama-3	70B parameters
Google	Gemma	2B parameters
	Gemma	7B parameters



Large Language Models (LLMs)

Transformer-based LLMs

- Pre-trained on massive datasets for tasks like Q&A, summarization, and translation.
- Highly flexible but **challenging** to train and fine-tune due to billions of parameters.

Can Quantum do it better?

OpenAI	GPT3	175B parameters
		.76T parameters
Meta	Llama-3	8B parameters
	Llama-3	70B parameters
Google	Gemma	2B parameters
	Gemma	7B parameters

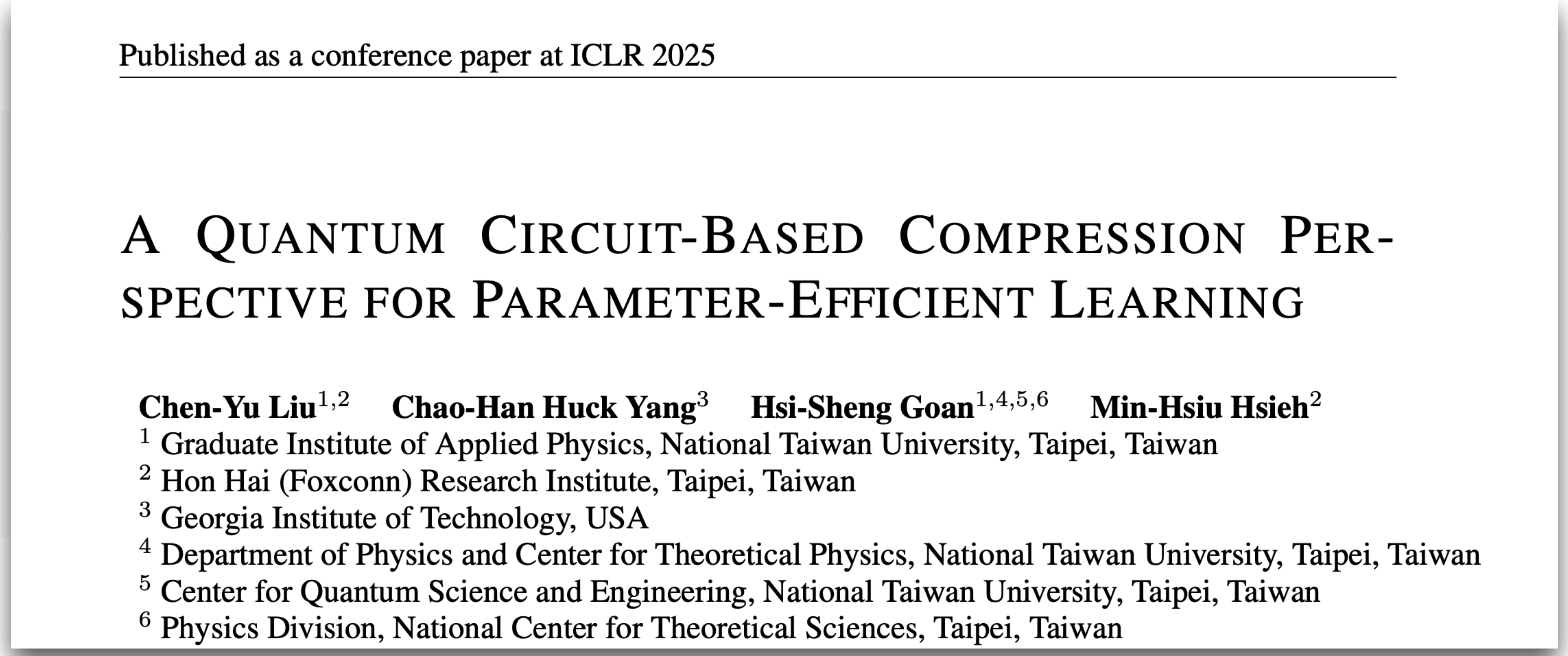
Large Language Models (LLMs)

National Taiwan University

Transformer-based LLMs

- Pre-trained on massive datasets for tasks like Q&A, summarization, and translation.
- Highly flexible but **challenging** to train and fine-tune due to billions of parameters.

Can Quantum do it better?


Published as a conference paper at ICLR 2025

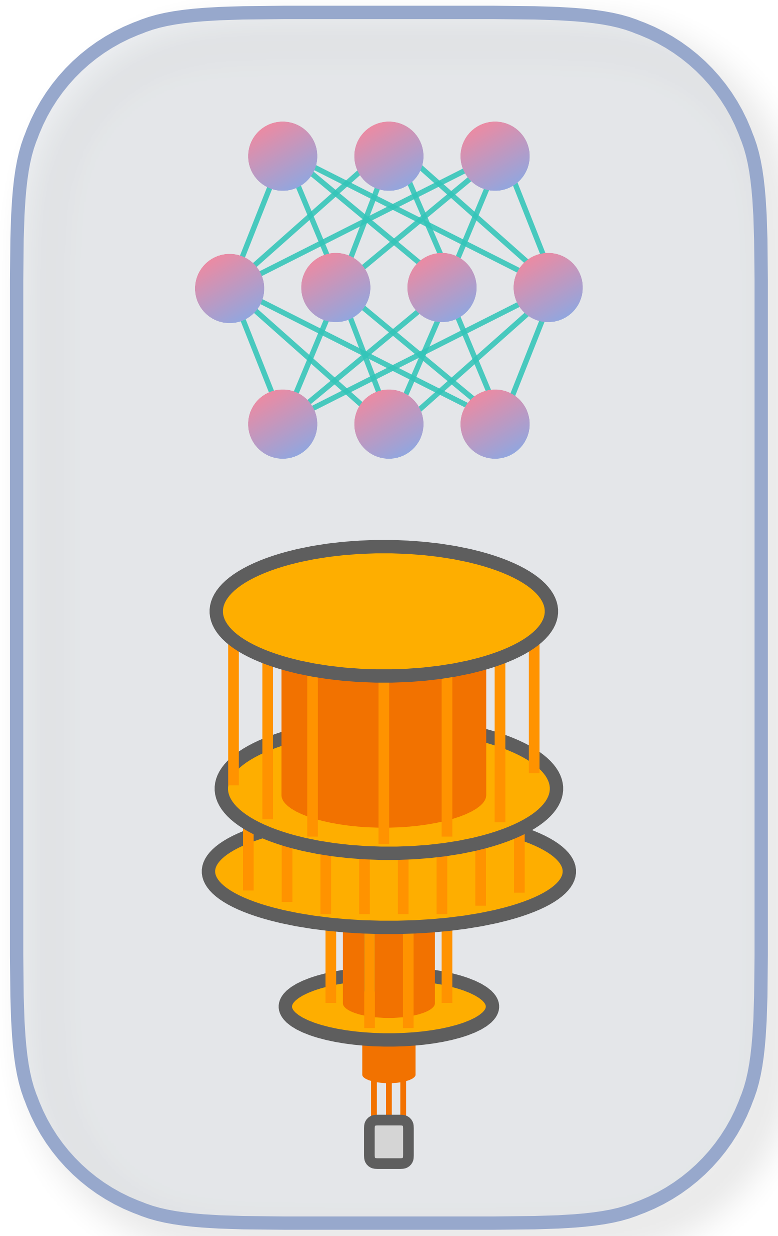
A QUANTUM CIRCUIT-BASED COMPRESSION PERSPECTIVE FOR PARAMETER-EFFICIENT LEARNING

Chen-Yu Liu[1,2] **Chao-Han Huck Yang**[3] **Hsi-Sheng Goan**[1,4,5,6] **Min-Hsiu Hsieh**[2]

[1] Graduate Institute of Applied Physics, National Taiwan University, Taipei, Taiwan
[2] Hon Hai (Foxconn) Research Institute, Taipei, Taiwan
[3] Georgia Institute of Technology, USA
[4] Department of Physics and Center for Theoretical Physics, National Taiwan University, Taipei, Taiwan
[5] Center for Quantum Science and Engineering, National Taiwan University, Taipei, Taiwan
[6] Physics Division, National Center for Theoretical Sciences, Taipei, Taiwan

The Gap

QML studies



Sounds good but…

$\sim 10^5$ parameters

This work

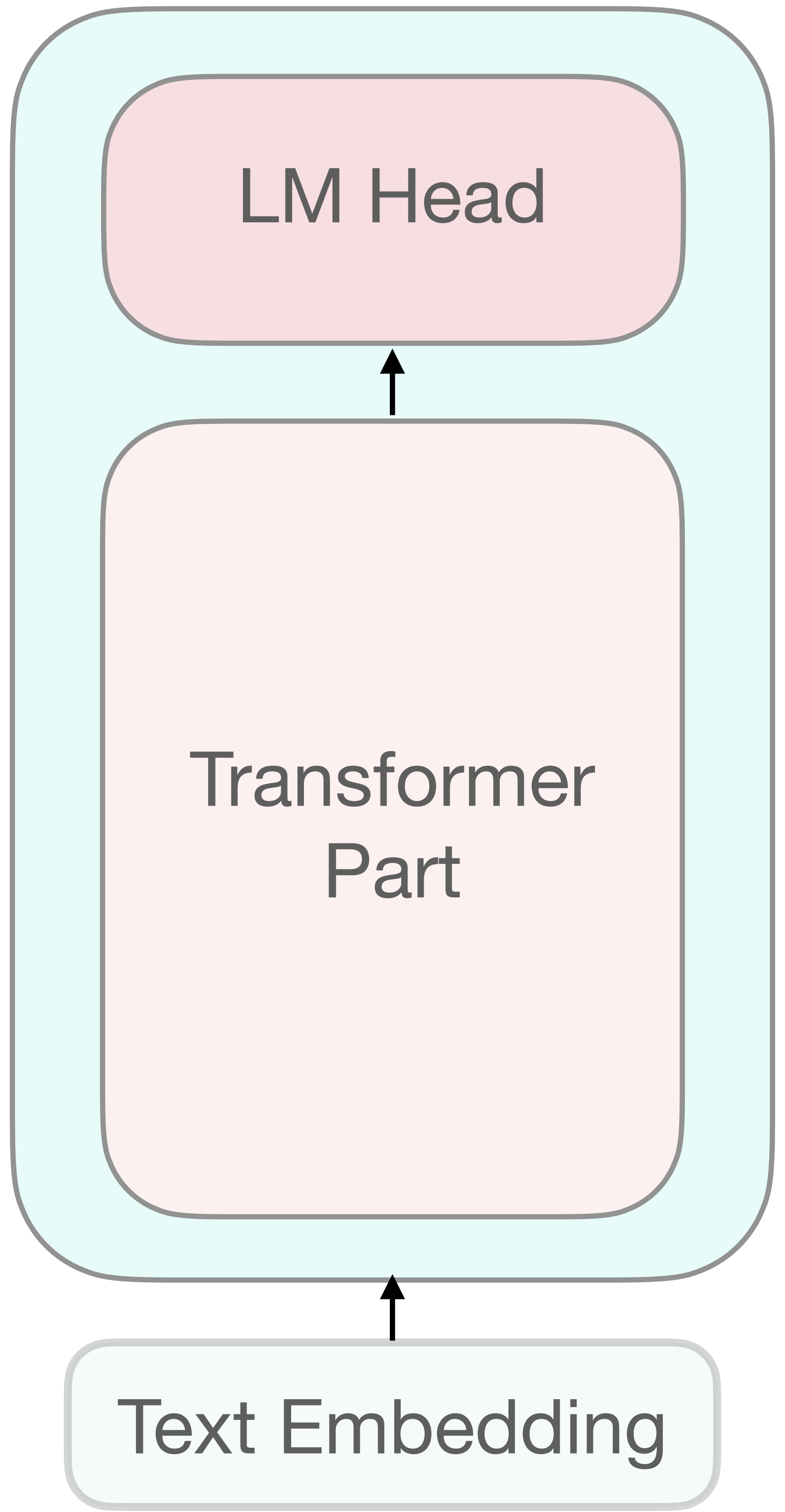
< 20 qubits

Relate to qubit count of original QT:
30~40 qubits

OpenAI	GPT3	175B parameters
	GPT4	1.76T parameters
Meta	Llama-3	8B parameters
	Llama-3	70B parameters
Google	Gemma	2B parameters
	Gemma	7B parameters

$$10^6 \rightarrow 10^9 \rightarrow 10^{12} \rightarrow \ldots$$

LLM

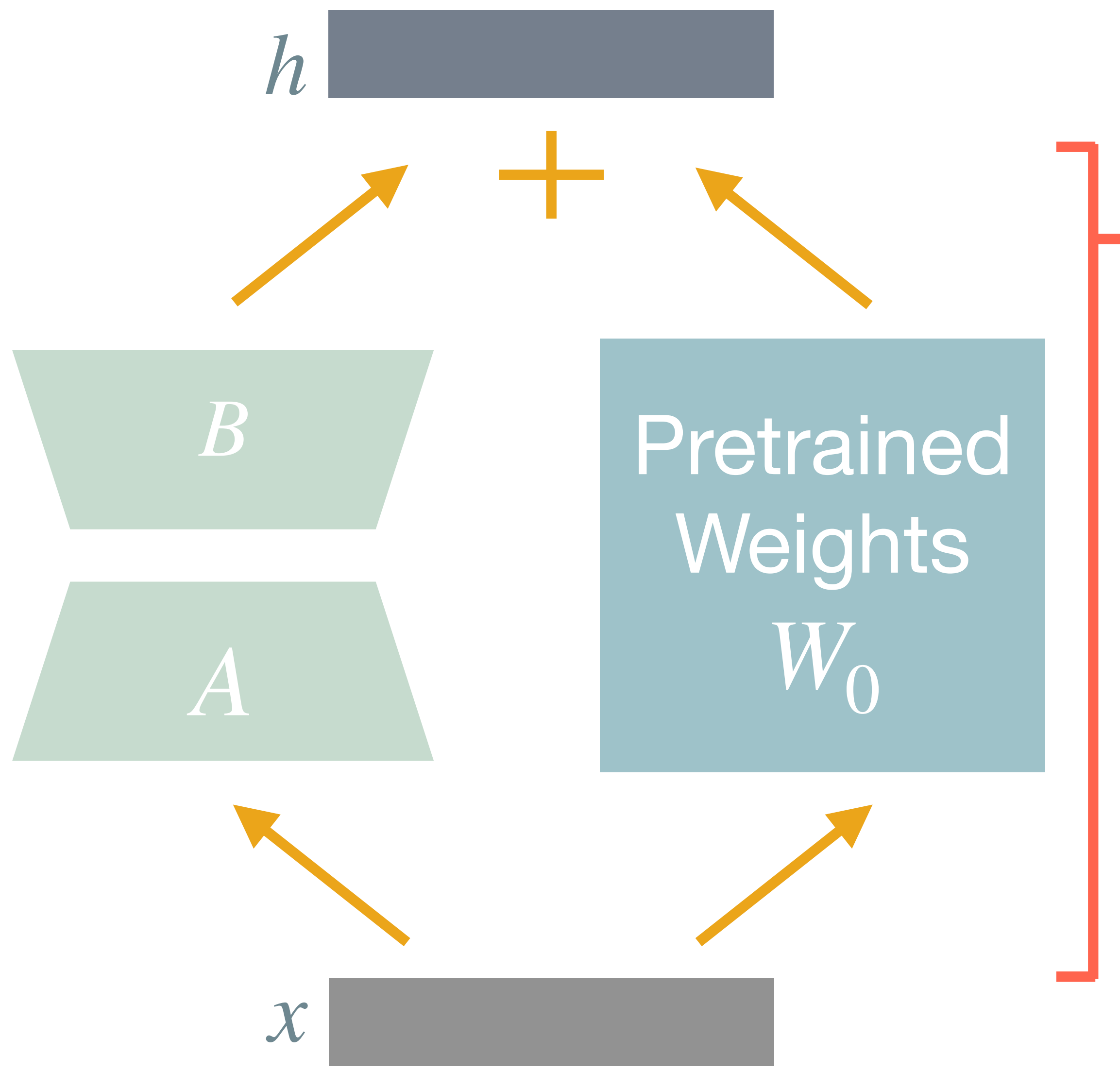


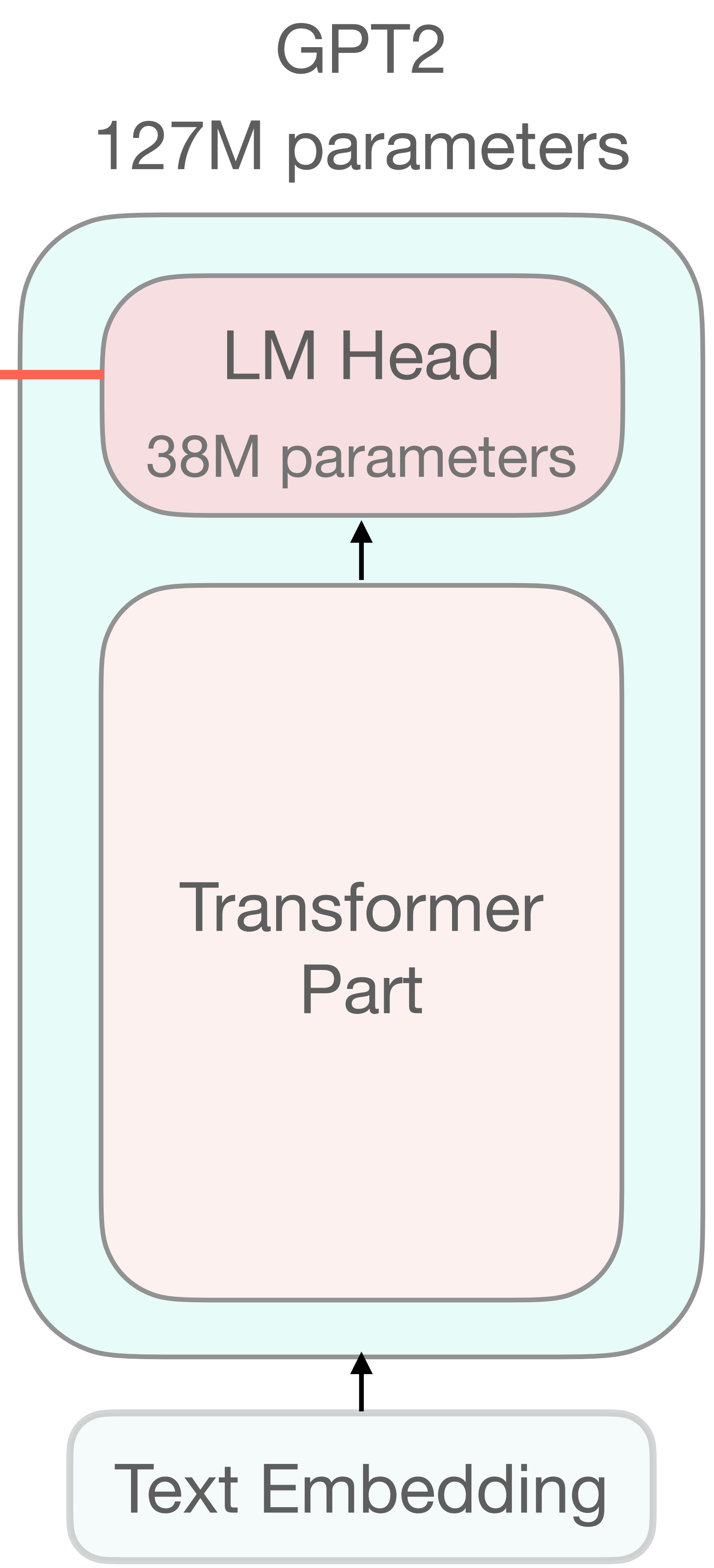
Parameter Efficient Fine-Tuning LLMs

National Taiwan University

LoRA: Low-Rank Adaptation @ ICLR 2022

- ***Method***
  - For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we constrain its update by representing the latter with a low-rank decomposition $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, with W_0 frozen during the training. The forward pass yields:
  - $h = W_0 x + \Delta W x = W_0 x + BAx$
  - With the rank $r \ll \min(d, k)$
  - Only A and B will be tuned during training

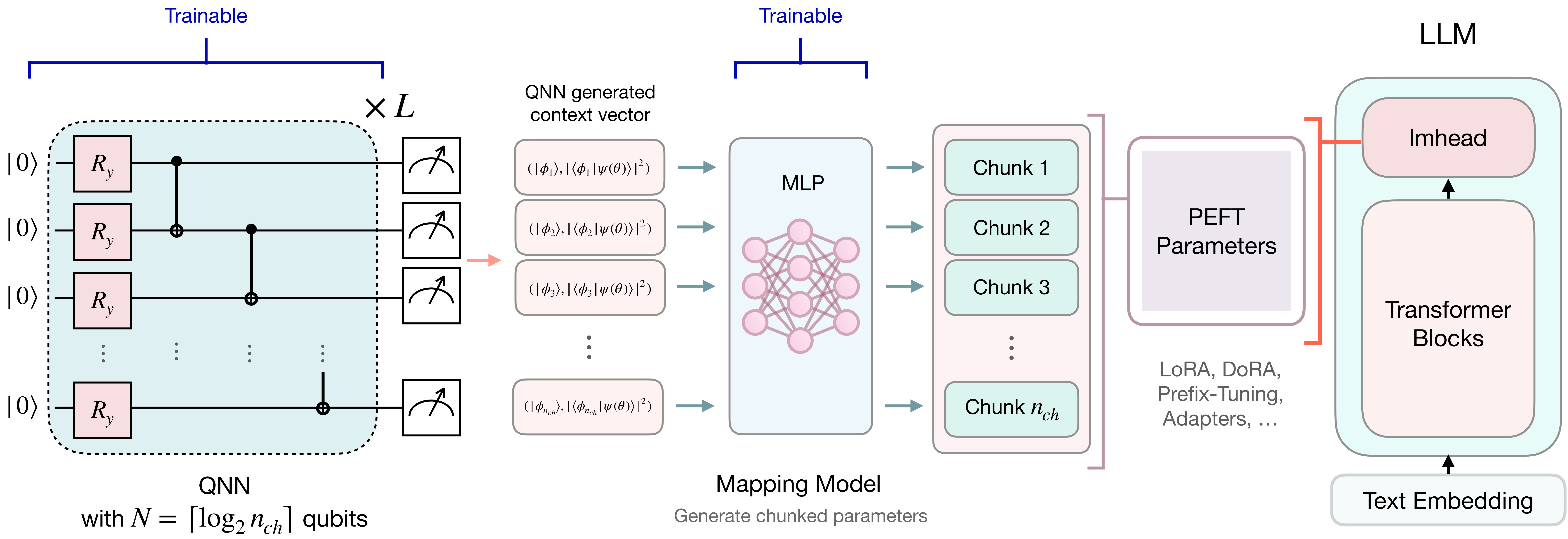




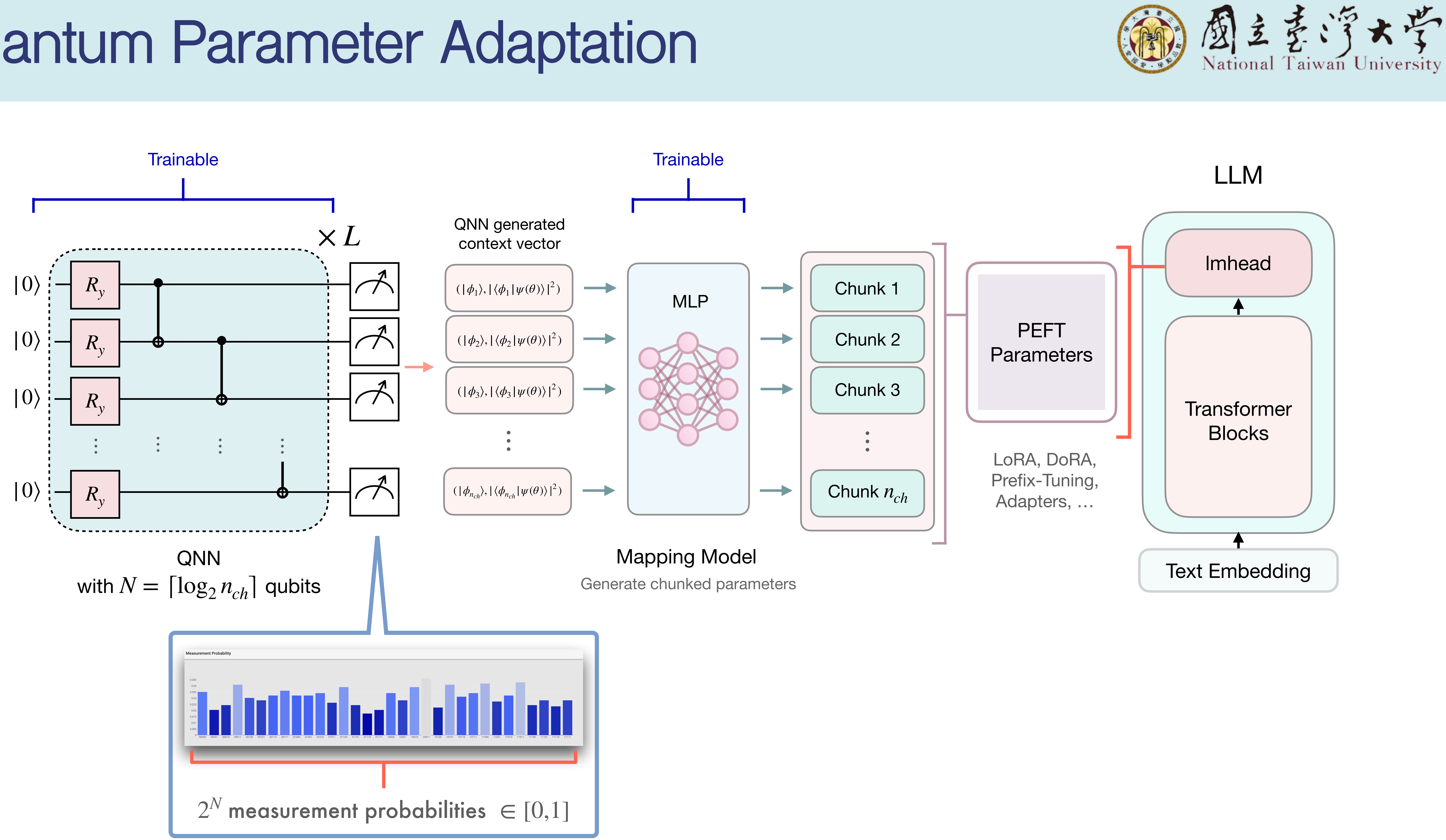
- **Ex.**

$$d = 500, k = 70 \ \rightarrow d \times k = 35000$$
$$r = 1 \rightarrow d \times r + r \times k = 570$$
$$r = 2 \rightarrow d \times r + r \times k = 1140$$

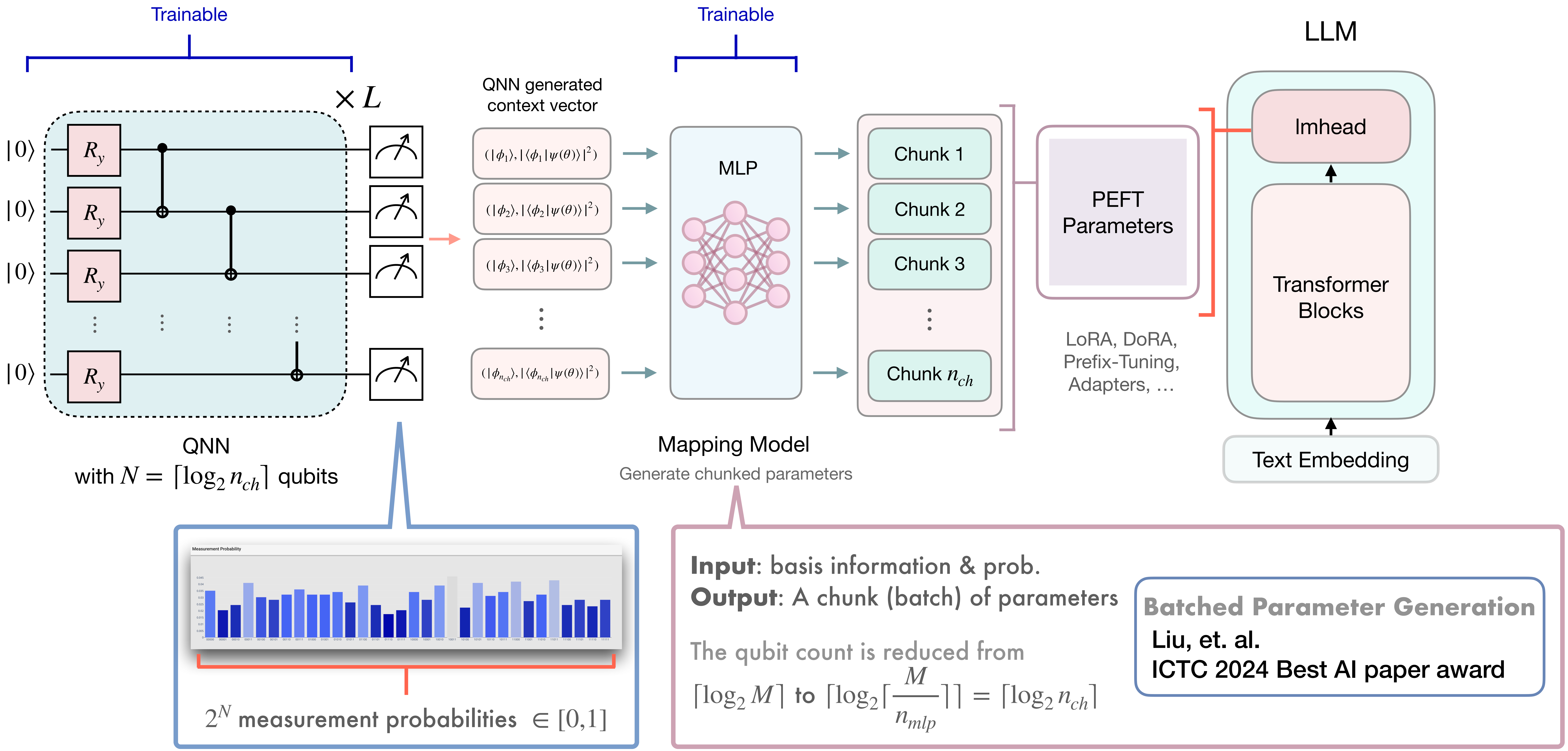
Quantum Parameter Adaptation



Quantum Parameter Adaptation



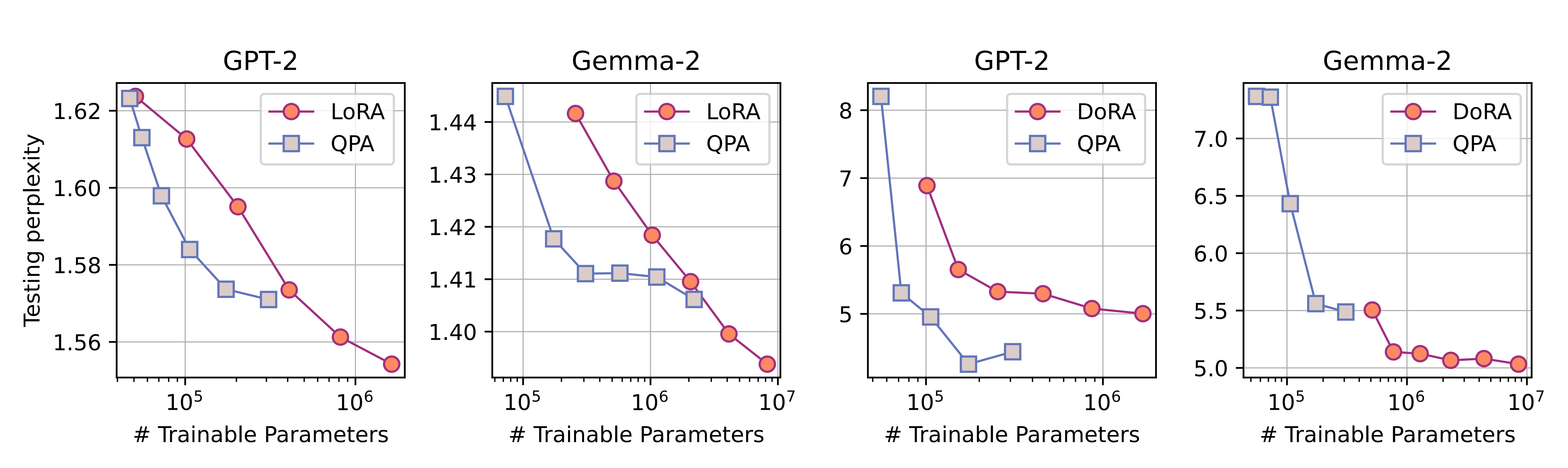
Quantum Parameter Adaptation



Quantum Parameter Adaptation

National Taiwan University

- Low-rank adaptation methods with QPA.



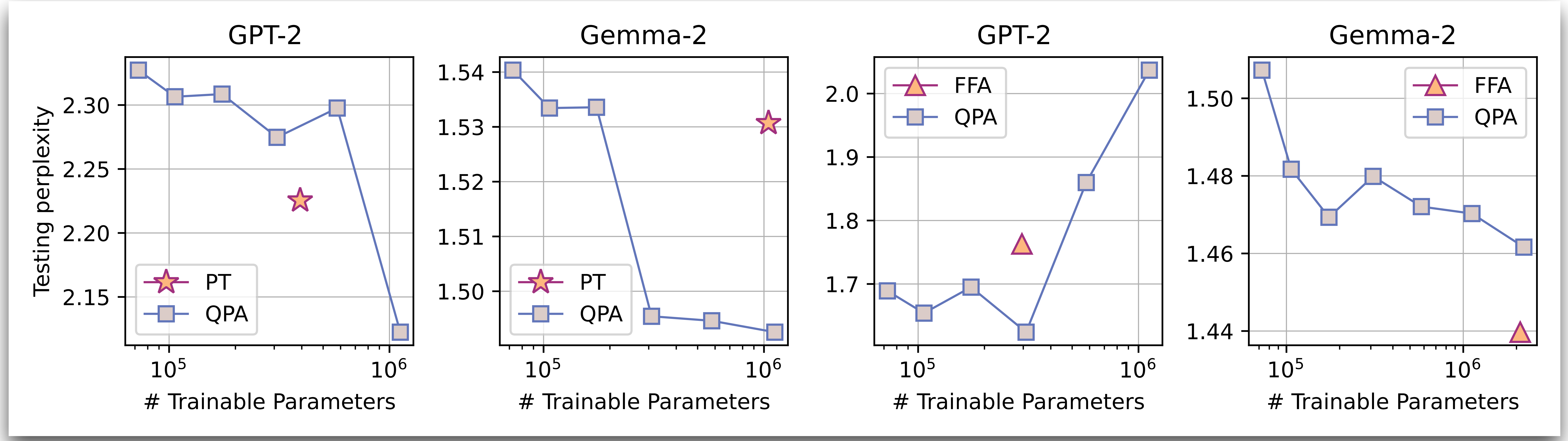
- Testing perplexity of GPT-2 (80M) and Gemma-2 (2B) models compared to the number of trainable parameters for LoRA, DoRA, and QPA on the WikiText-2 dataset.

* On ideal simulator
* All experiments were conducted on NVIDIA V100S and NVIDIA H100 GPUs.

Quantum Parameter Adaptation

National Taiwan University

- QPA on Prefix-Tuning and Feed-forward adapter.



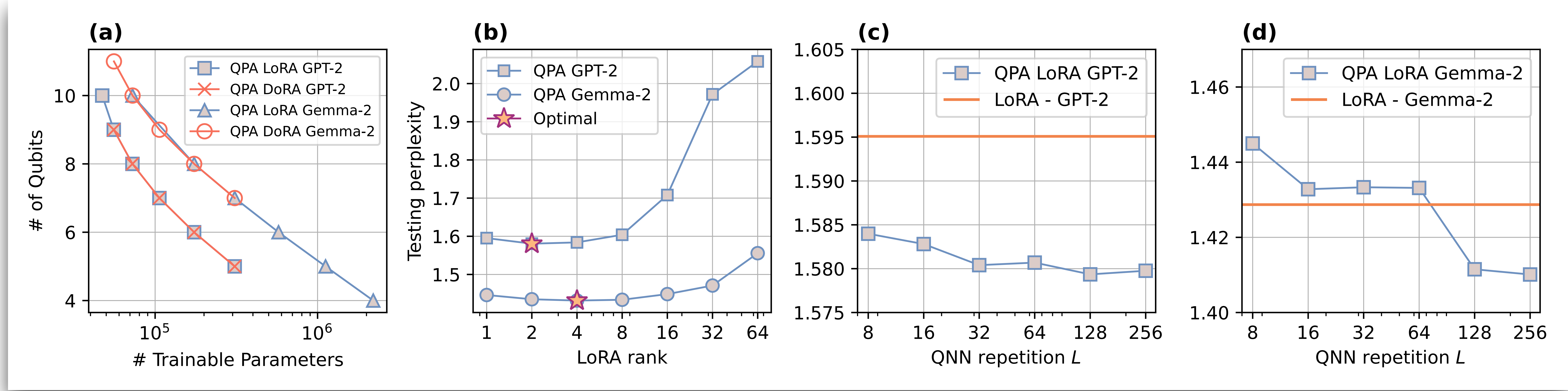
- Testing perplexity of GPT-2 and Gemma-2 models compared to the number of trainable parameters for prefix-tuning (PT), feed-forward adapter (FFA), and QPA on the WikiText-2 dataset.

* On ideal simulator
* All experiments were conducted on NVIDIA V100S and NVIDIA H100 GPUs.

Quantum Parameter Adaptation

National Taiwan University

Effects of QPA settings



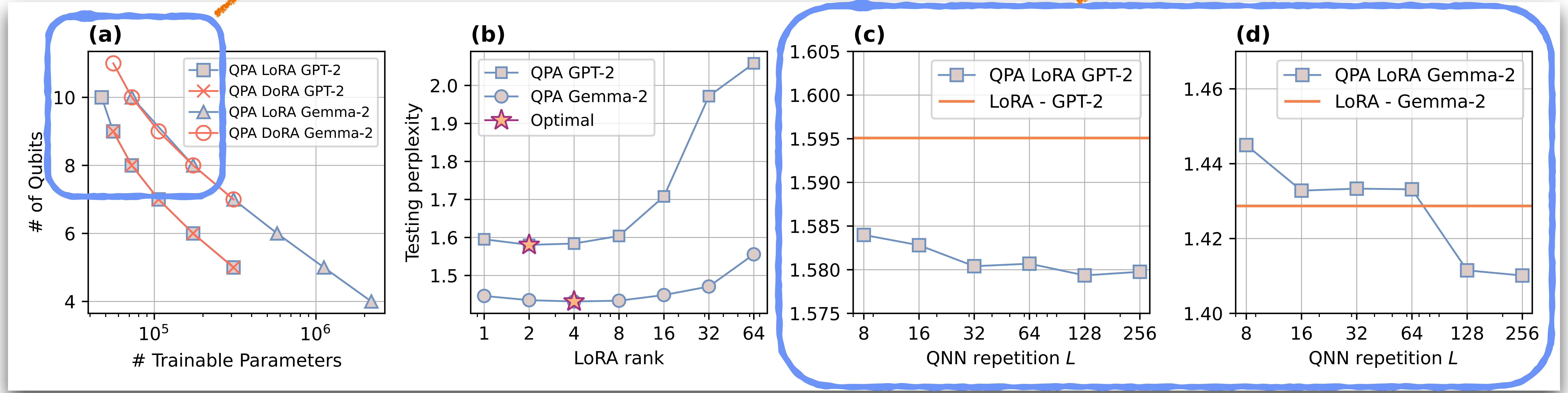
- **(a)** Qubit usage versus the number of trainable parameters for QPA applied to LoRA and DoRA on GPT-2 and Gemma-2 models. **(b)** The relationship between testing perplexity and LoRA rank for QPA applied to GPT-2 and Gemma-2. **(c)** and **(d)** Testing perplexity depending on the QNN repetition L for QPA applied to LoRA on GPT-2 and Gemma-2.

On ideal simulator

Quantum Parameter Adaptation

Reduced qubit count due to batched parameter generation (arXiv:2409.02763), Low-rank adaptation, ...

More expressive circuits have better QPA performance



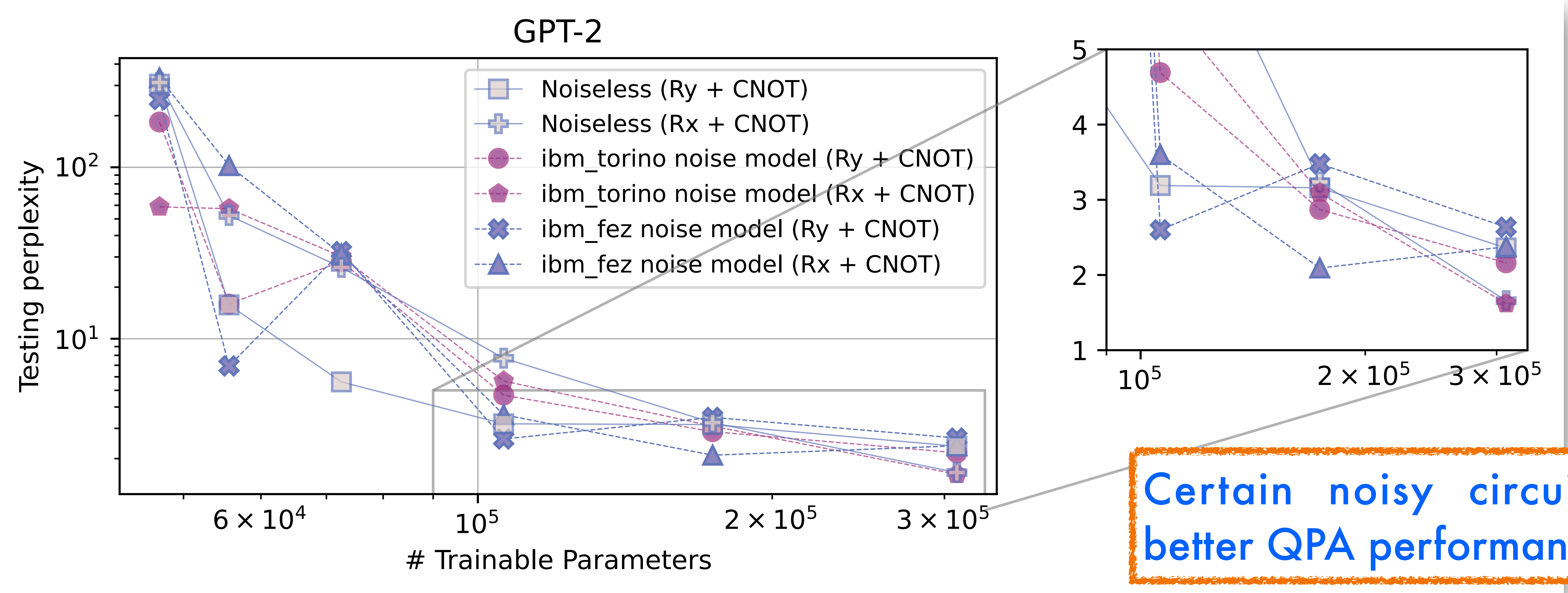
- **(a)** Qubit usage versus the number of trainable parameters for QPA applied to LoRA and DoRA on GPT-2 and Gemma-2 models. **(b)** The relationship between testing perplexity and LoRA rank for QPA applied to GPT-2 and Gemma-2. **(c)** and **(d)** Testing perplexity depending on the QNN repetition L for QPA applied to LoRA on GPT-2 and Gemma-2.

✱ On ideal simulator

Quantum Parameter Adaptation

National Taiwan University

- Effects of quantum noise model from IBM quantum computers



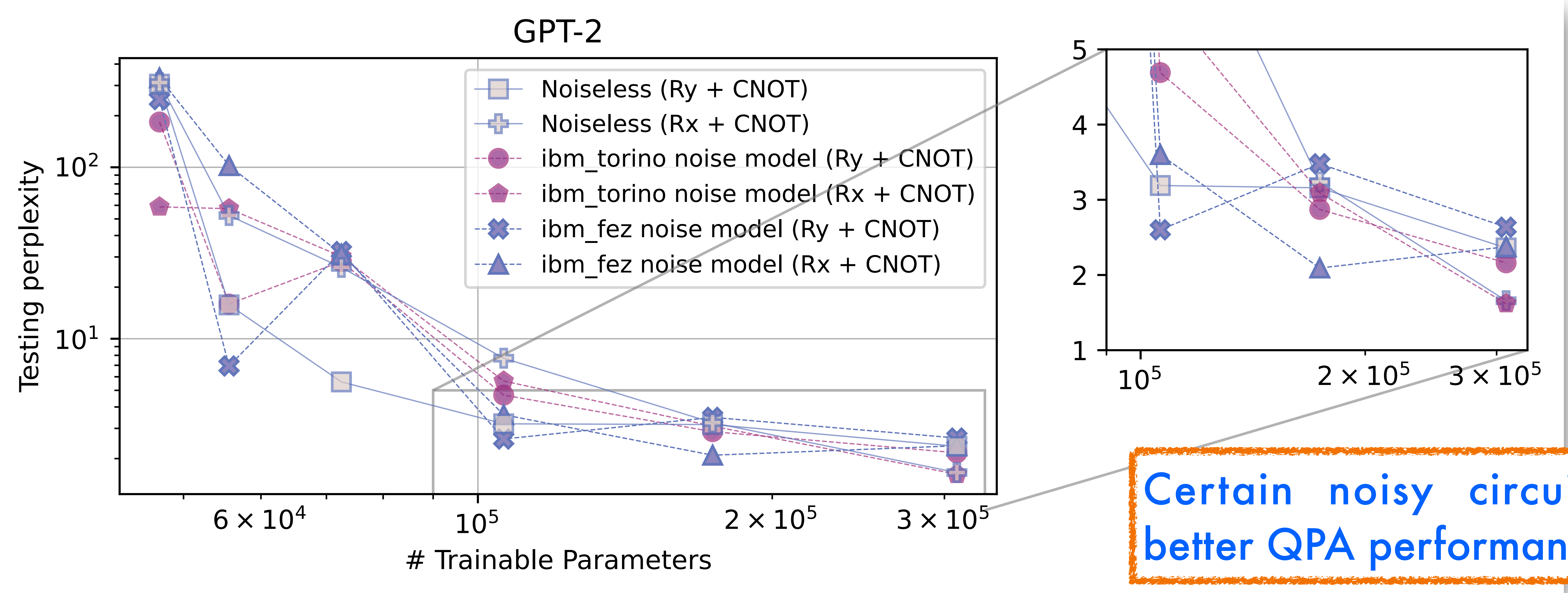
Certain noisy circuits exhibit better QPA performance.

- Testing perplexity of GPT-2 versus the number of trainable parameters on different noise setting with RY + CNOT and RX+ CNOT ansatz. The comparison includes LoRA and QPA applied at LoRA rank (r = 4), where n_{shot} is fixed at $n_{shot} = 20 \times 2^N$.

On noisy simulator

Quantum Parameter Adaptation

- Effects of quantum noise model from IBM quantum computers



Certain noisy circuits exhibit better QPA performance.

NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better

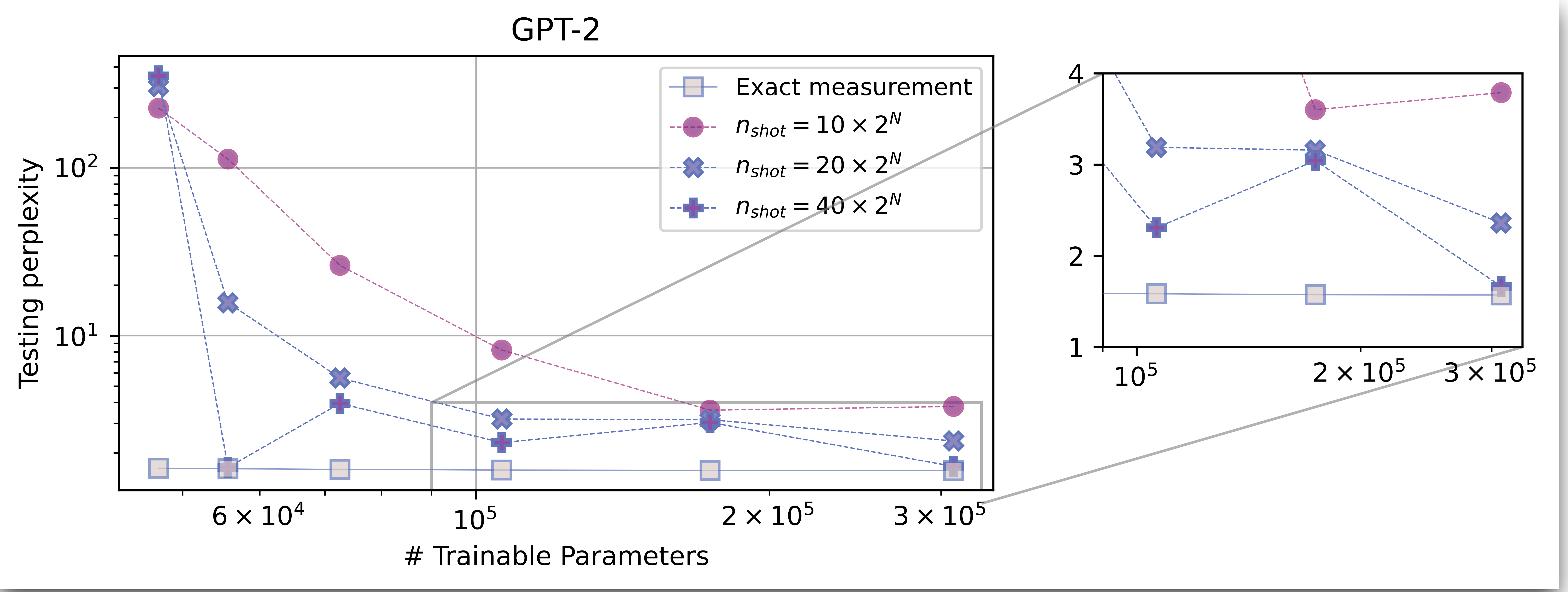
Chuhan Wu[†] **Fangzhao Wu**[‡*] **Tao Qi**[†] **Yongfeng Huang**[†]
[†]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[‡]Microsoft Research Asia, Beijing 100080, China

(ACL2022)

Quantum Parameter Adaptation

國立臺灣大學 National Taiwan University

Effects of finite measurement shots

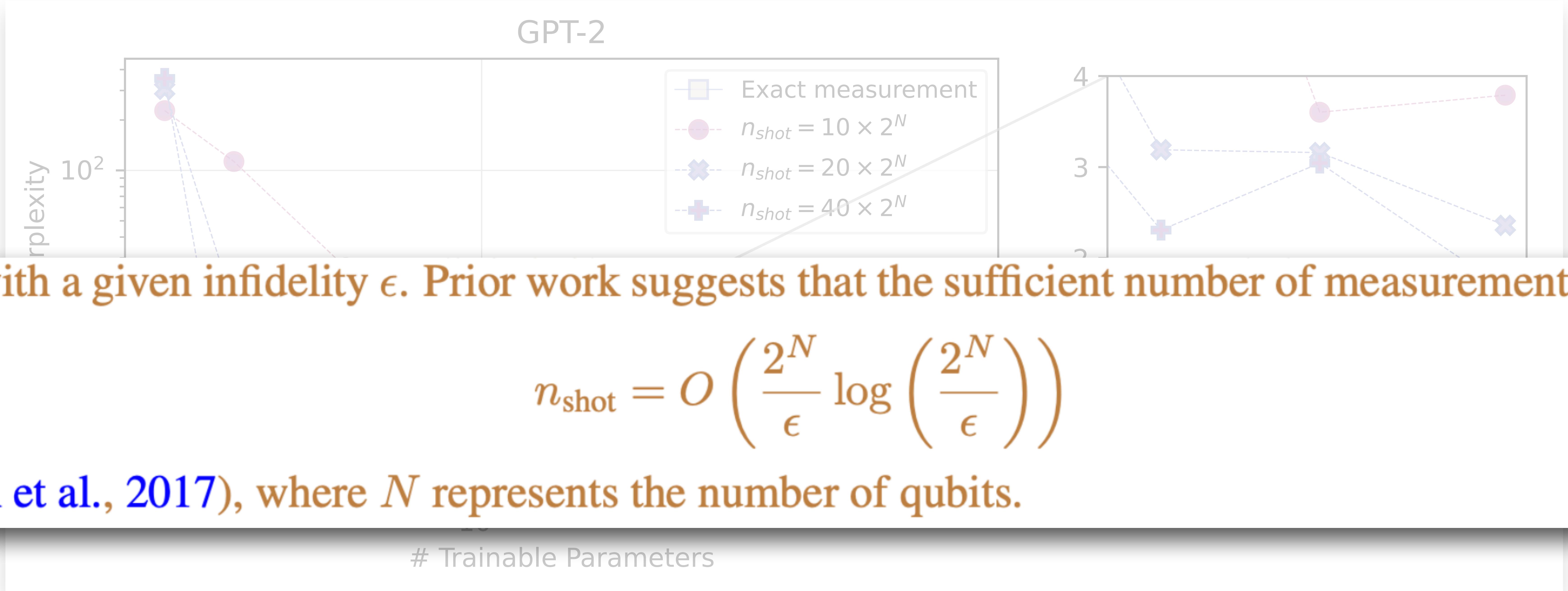


- Testing perplexity of GPT-2 versus the number of trainable parameters on different number of measurement shots with RY + CNOT ansatz. The comparison includes LoRA and QPA applied at LoRA rank (r = 4).

⁕ On ideal simulator

Quantum Parameter Adaptation

- Effects of finite measurement shots



$|\gamma\rangle$, with a given infidelity ϵ. Prior work suggests that the sufficient number of measurement shots is

$$n_{\text{shot}} = O\left(\frac{2^N}{\epsilon}\log\left(\frac{2^N}{\epsilon}\right)\right) \tag{15}$$

(Haah et al., 2017), where N represents the number of qubits.

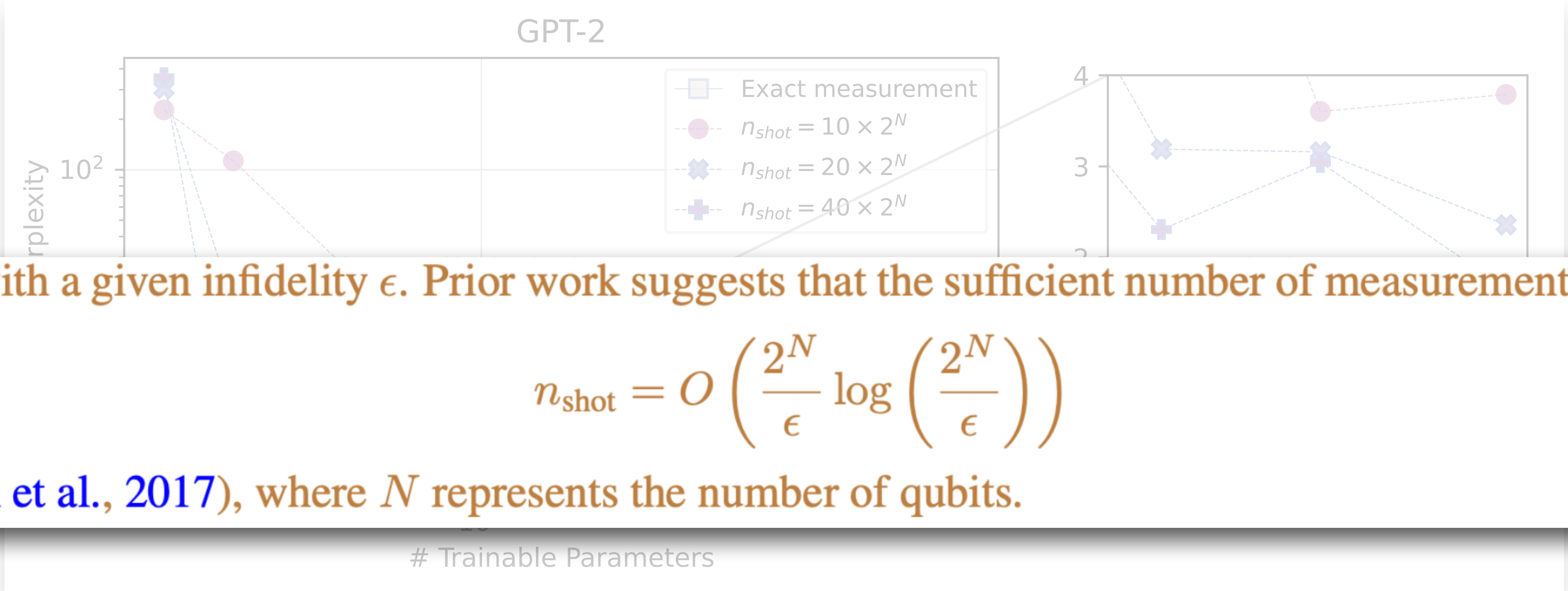
- Testing perplexity of GPT-2 versus the number of trainable parameters on different number of measurement shots with RY + CNOT ansatz. The comparison includes LoRA and QPA applied at LoRA rank (r = 4).

On ideal simulator

Quantum Parameter Adaptation

國立臺灣大學 National Taiwan University

- Effects of finite measurement shots



$|\gamma\rangle$, with a given infidelity ϵ. Prior work suggests that the sufficient number of measurement shots is

$$n_{\text{shot}} = O\left(\frac{2^N}{\epsilon}\log\left(\frac{2^N}{\epsilon}\right)\right) \tag{15}$$

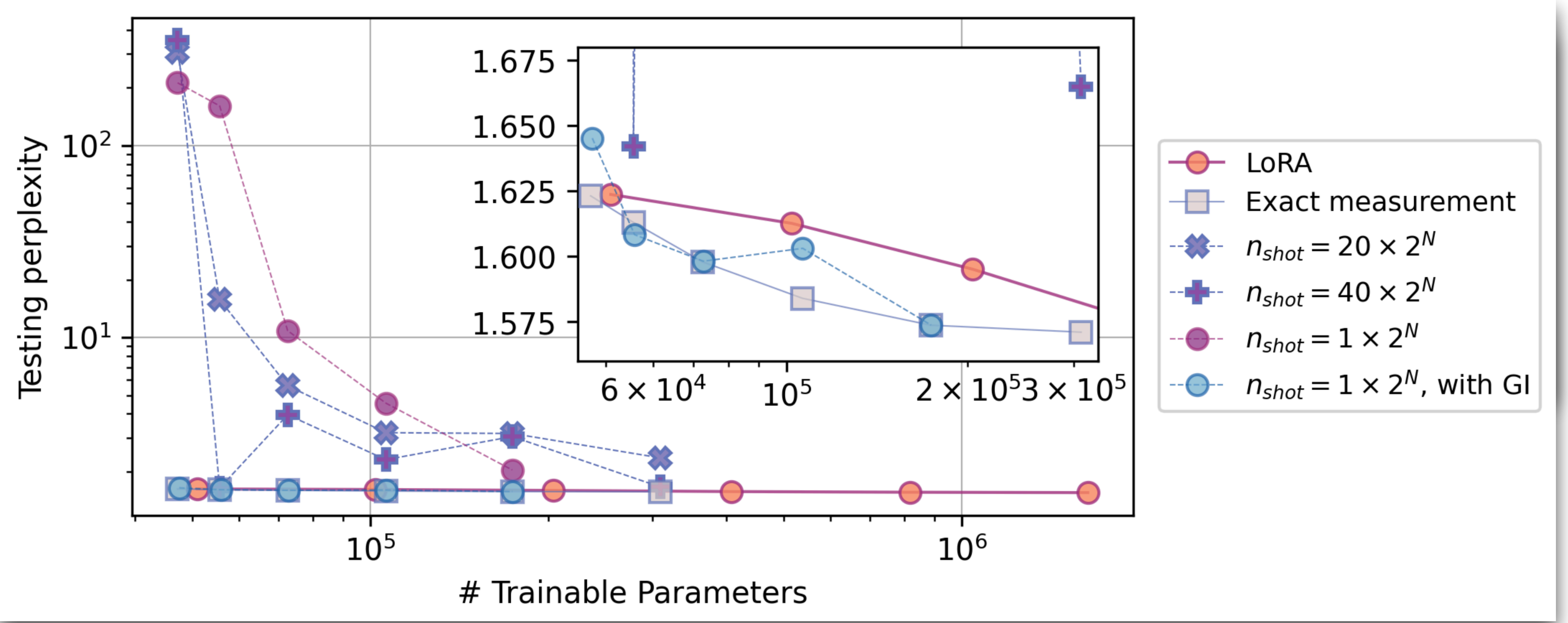
(Haah et al., 2017), where N represents the number of qubits.

New paper @ ICLR 2025 workshop:
Frame Generation in Hilbert Space: Generative Interpolation of Measurement Data for Quantum Parameter Adaptation

Quantum Parameter Adaptation

国立臺灣大學 National Taiwan University

- New paper @ ICLR 2025 workshop:
  Frame Generation in Hilbert Space: Generative Interpolation of Measurement Data for Quantum Parameter Adaptation



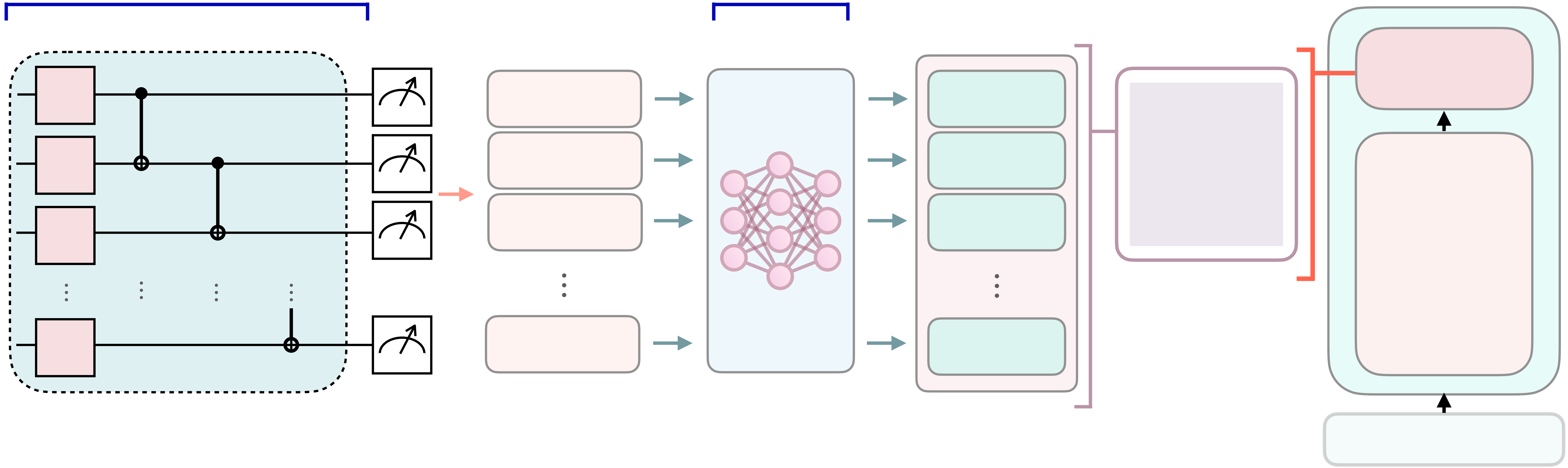
Summary

Generate PEFT parameters by QNN and MLP for LLMs

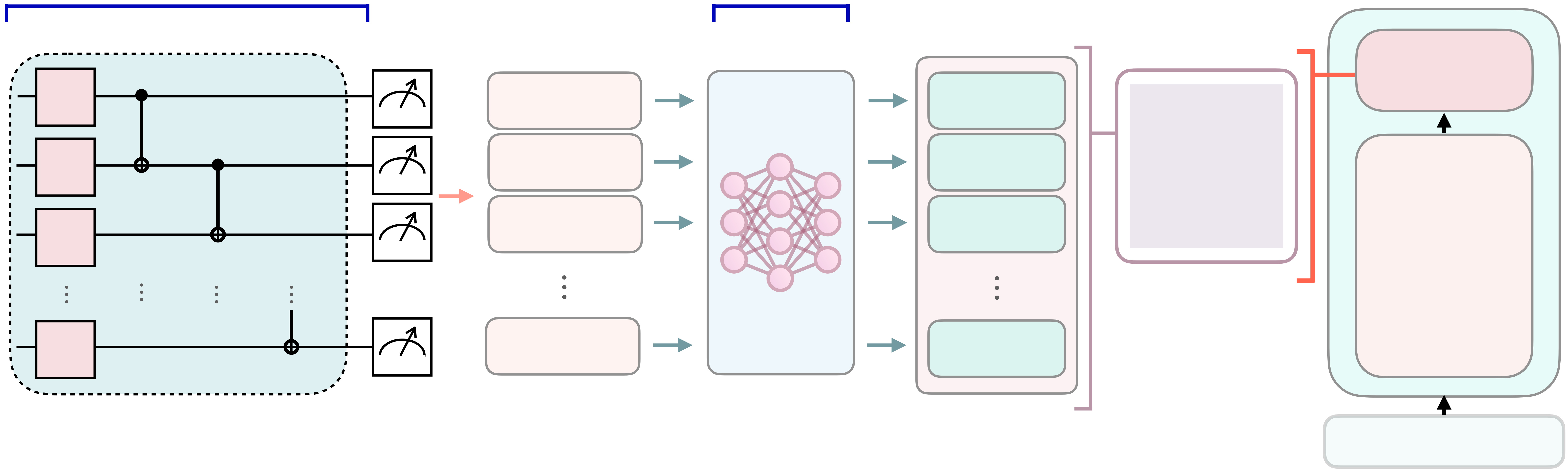
- The **first example** of using QML to fine-tune classical LLMs with improved performance.
- Further reduction of training parameters based on classical PEFT methods.
- No issues with data encoding.
- No quantum hardware required during inference.

On-going & Future work

- A generative method to mitigate the issue of finite measurement shots. (ICLR 2025 workshop)
- Characterizing the effect of real quantum computer noise on QPA.
- Real quantum computer implementation.



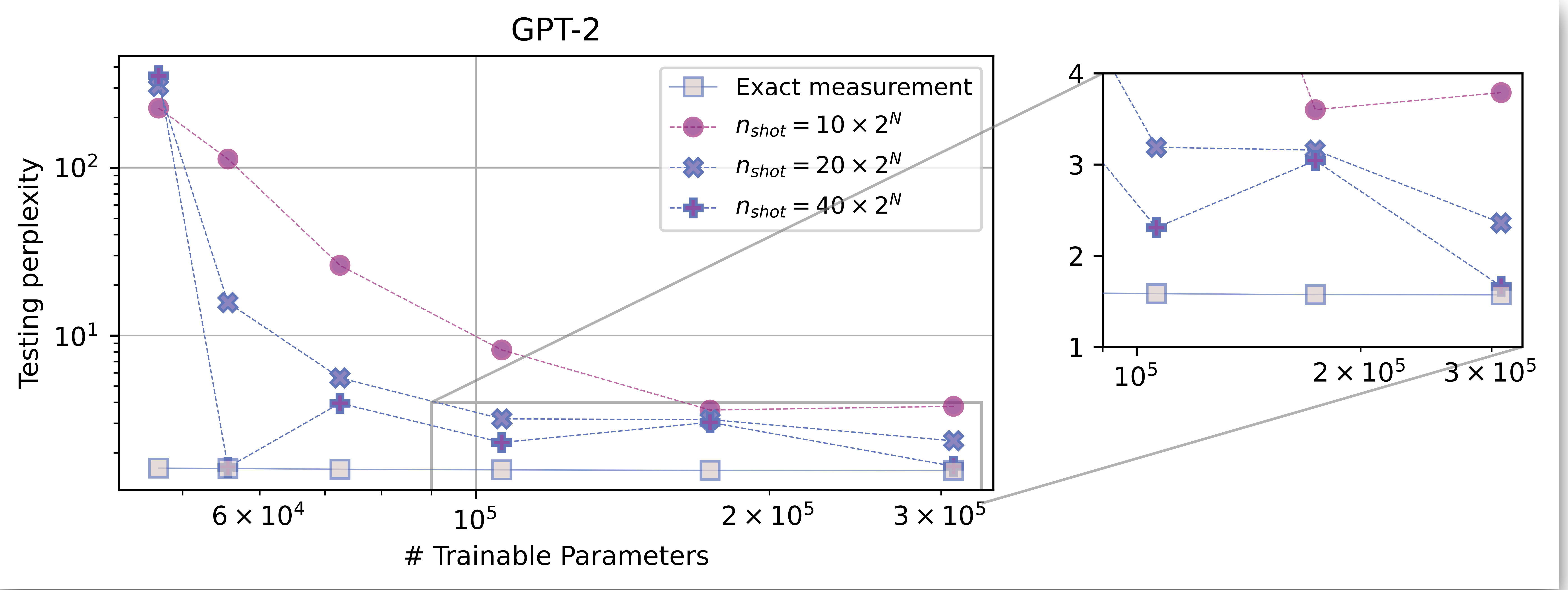
Appendix



ICLR International Conference On Learning Representations

國立臺灣大學 National Taiwan University

- Effects of finite measurement shots



- Testing perplexity of GPT-2 versus the number of trainable parameters on different number of measurement shots with RY + CNOT ansatz. The comparison includes LoRA and QPA applied at LoRA rank (r = 4).

On ideal simulator

Generative approach?

- Pre-train and predict

Article | Open access | Published: 20 October 2022

Flexible learning of quantum states with generative query neural networks

Yan Zhu, Ya-Dong Wu, Ge Bai, Dong-Sheng Wang, Yuexuan Wang & Giulio Chiribella

Nature Communications **13**, Article number: 6222 (2022) | Cite this article

5055 Accesses | **6** Altmetric | Metrics

- Pre-train and sample

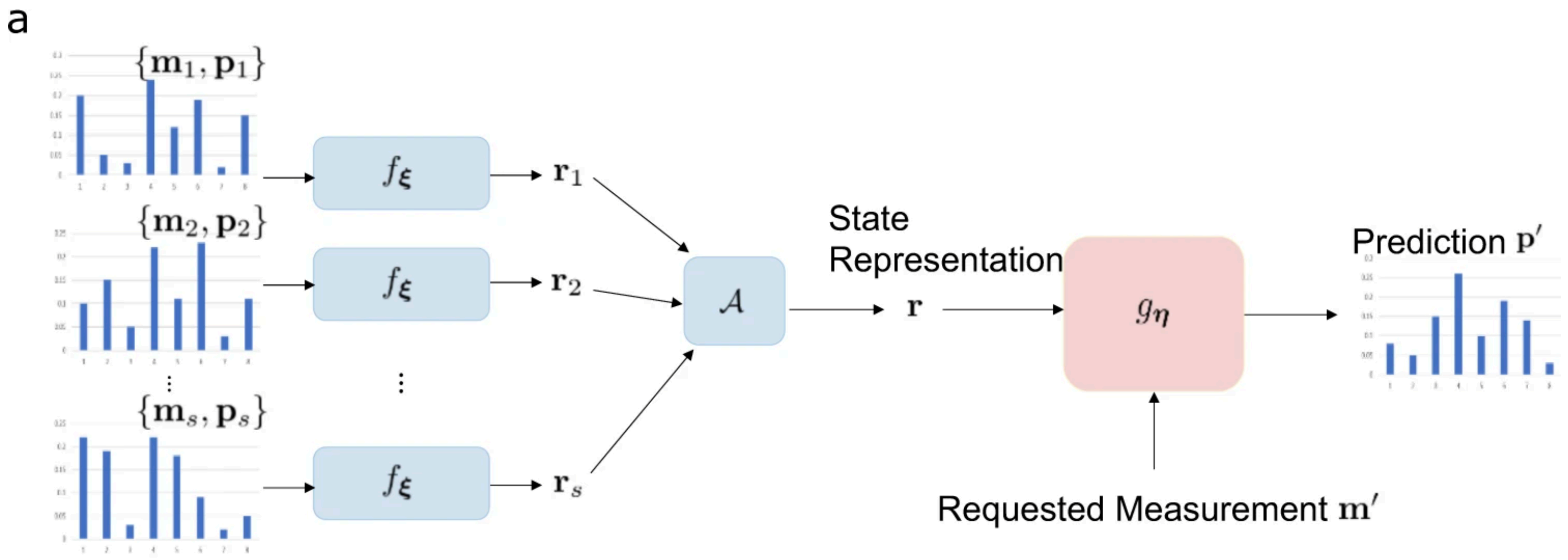
Article | Published: 11 March 2019

Reconstructing quantum states with generative models

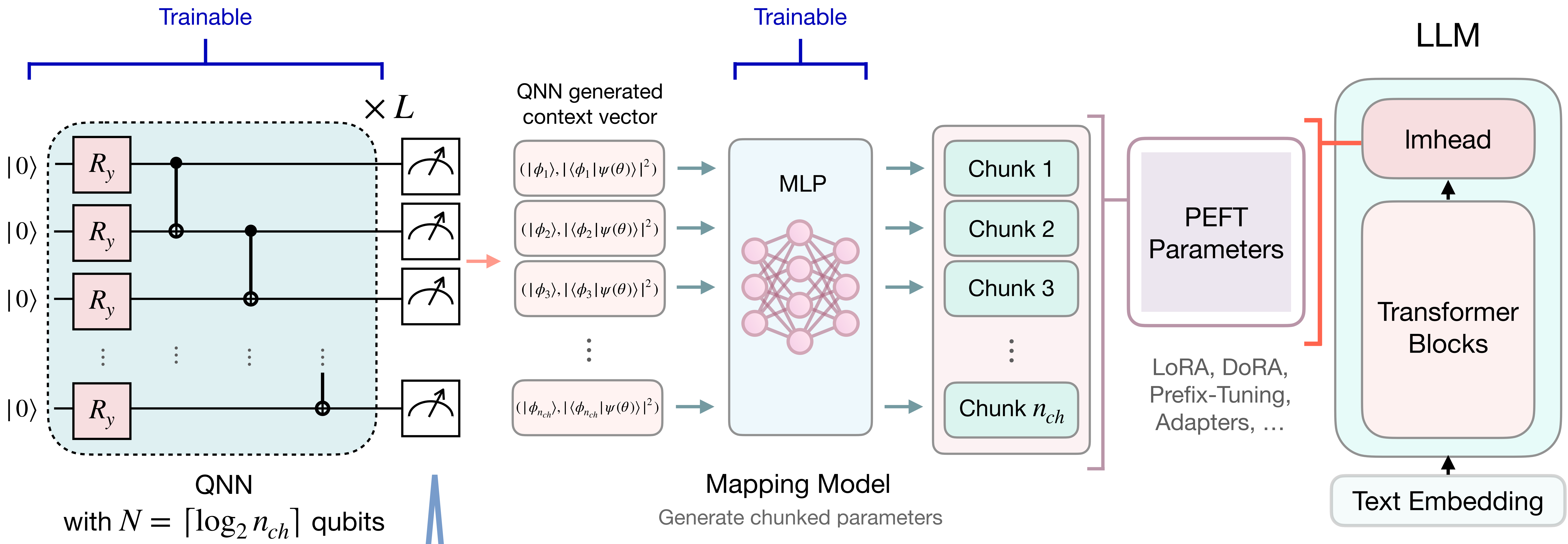
Juan Carrasquilla, Giacomo Torlai, Roger G. Melko & Leandro Aolita

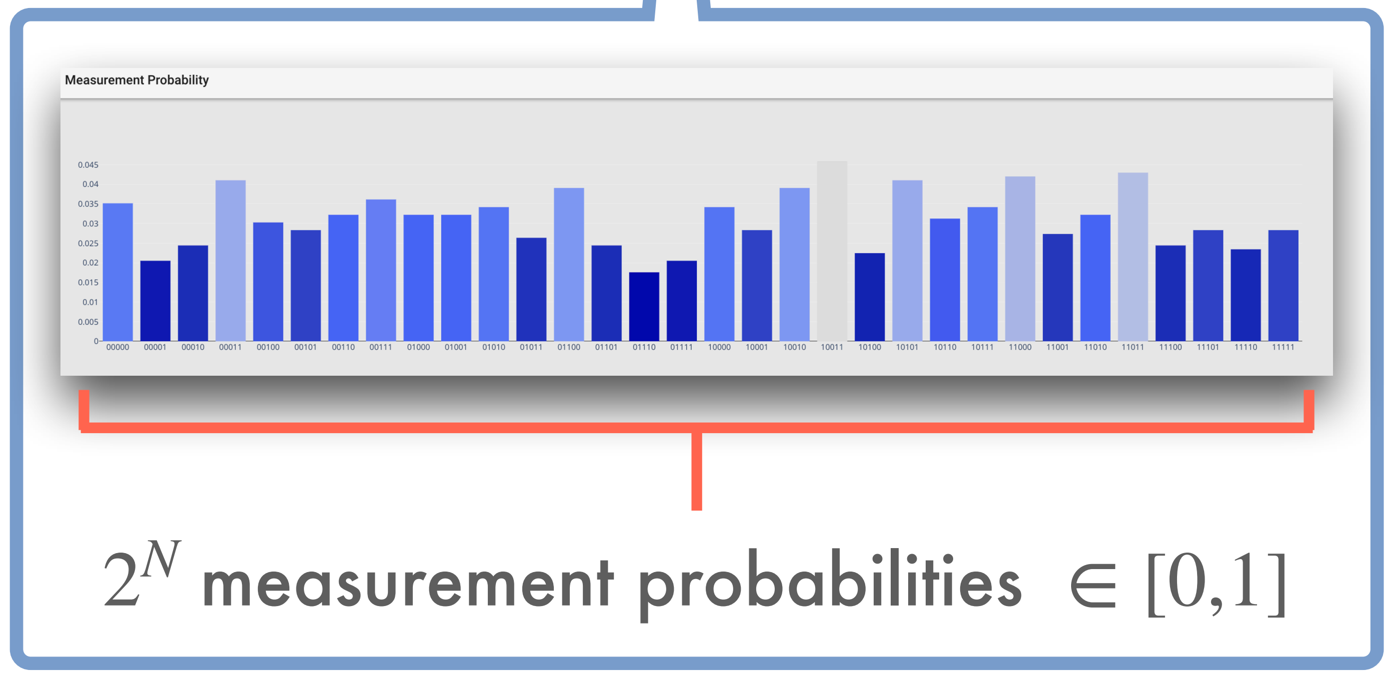
Nature Machine Intelligence **1**, 155–161 (2019) | Cite this article

6156 Accesses | **32** Altmetric | Metrics



Generative approach?

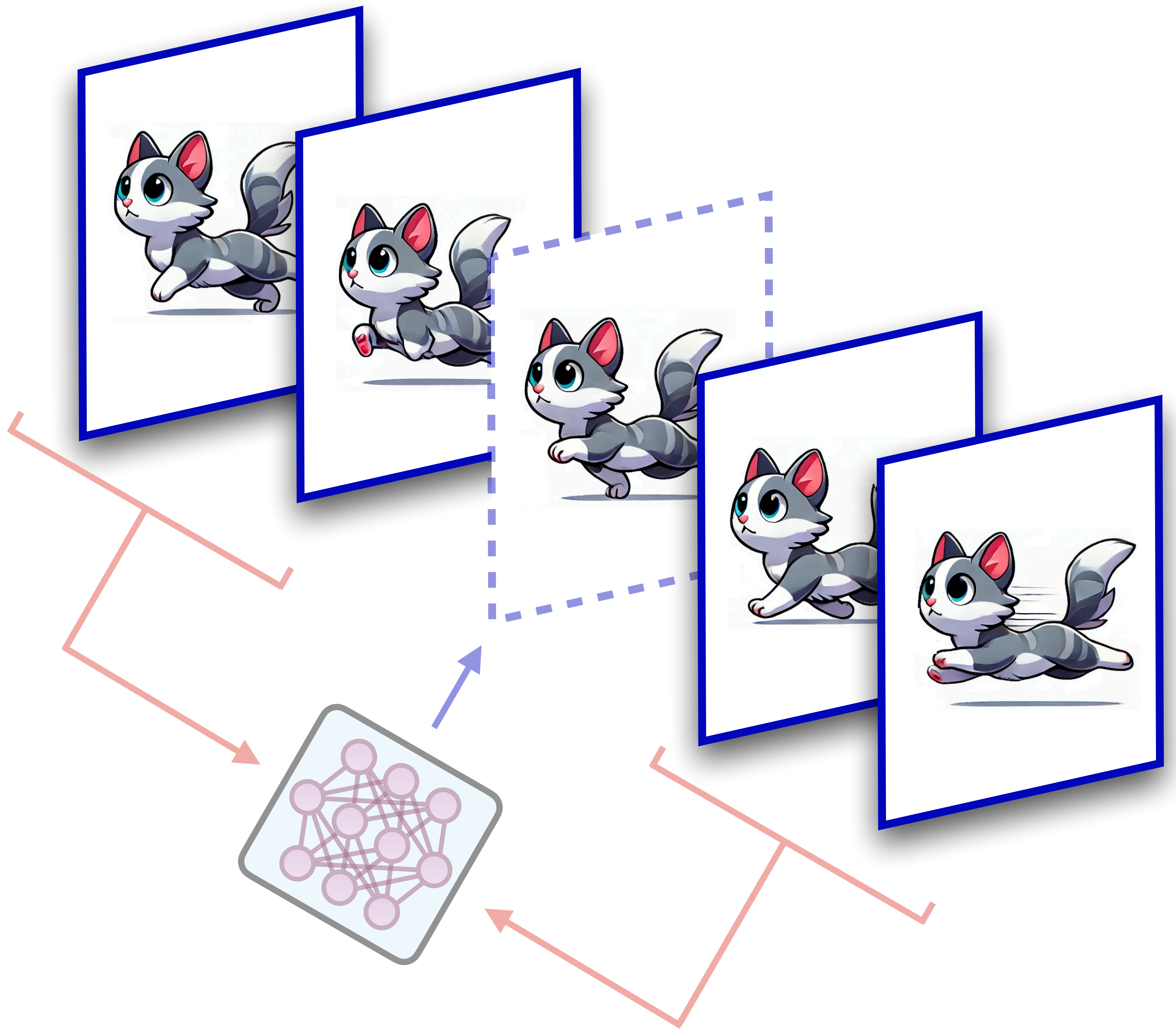




- With insufficient measurement counts:
  - 1. Some of the basis are unmeasured.
  - 2. Precision of measured basis probability is low.

Generative approach

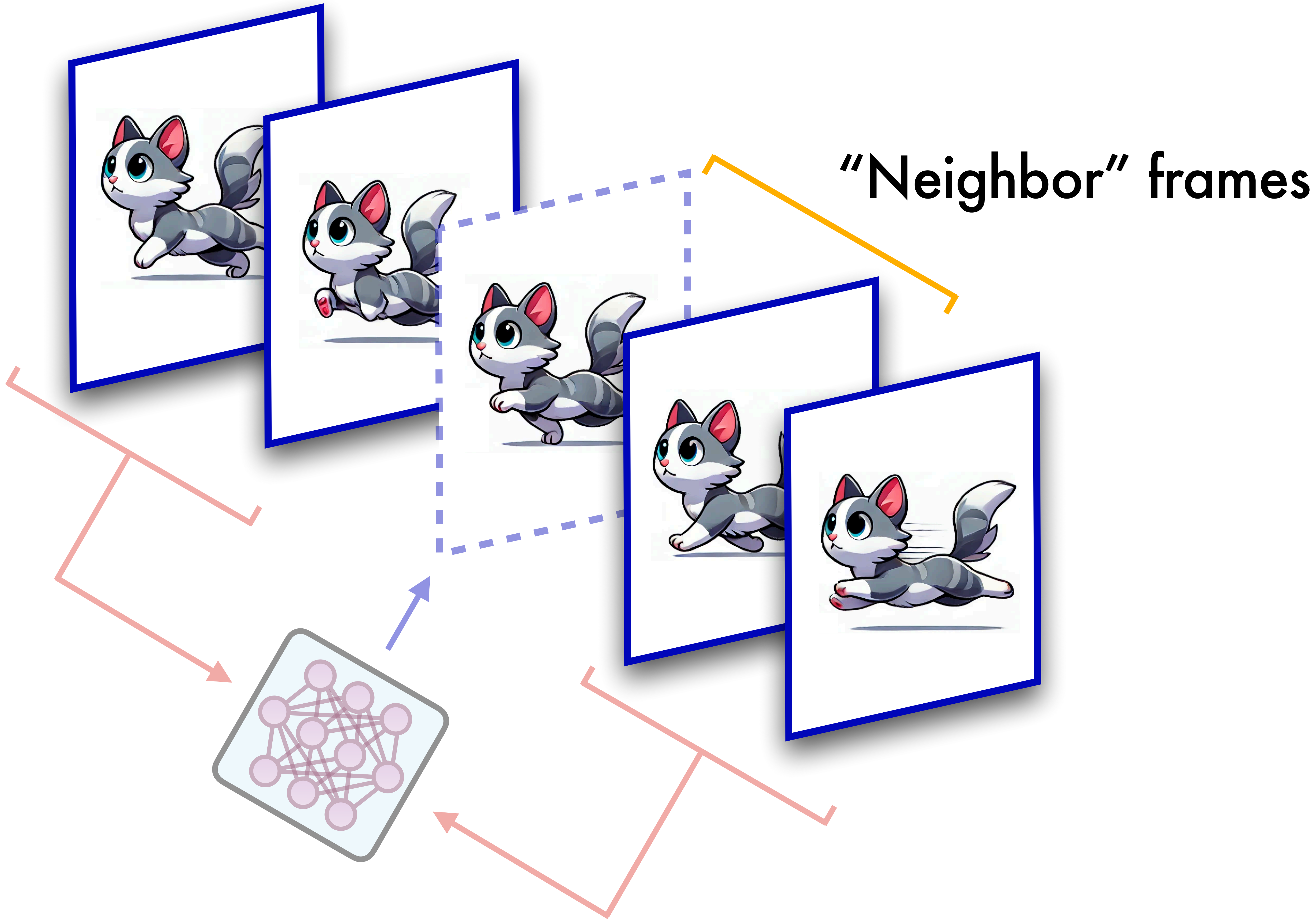
- General scheme of frame generation in video data



4K: 3840 * 2160
720p: 1280 * 720

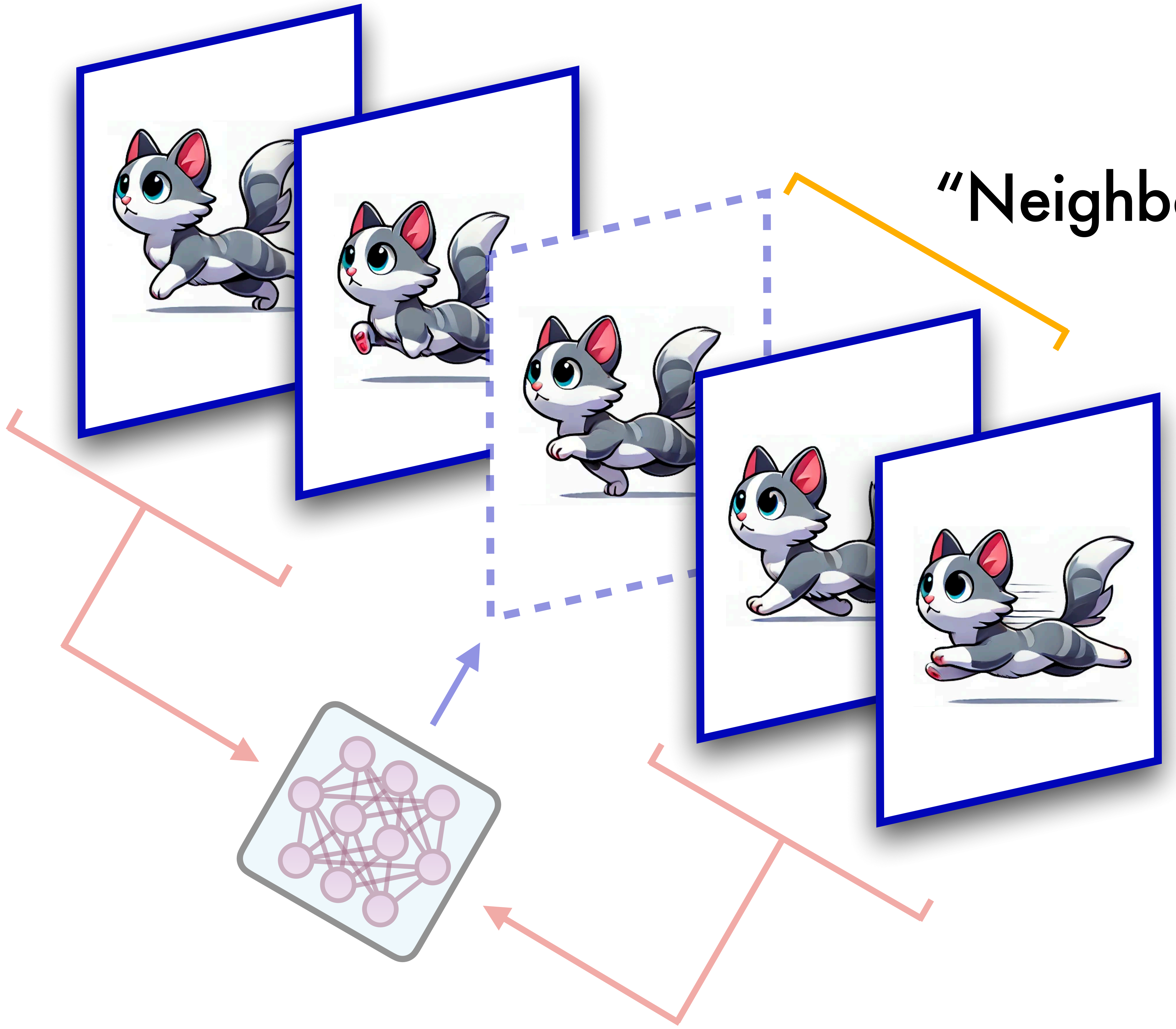
Generative approach

- General scheme of frame generation in video data



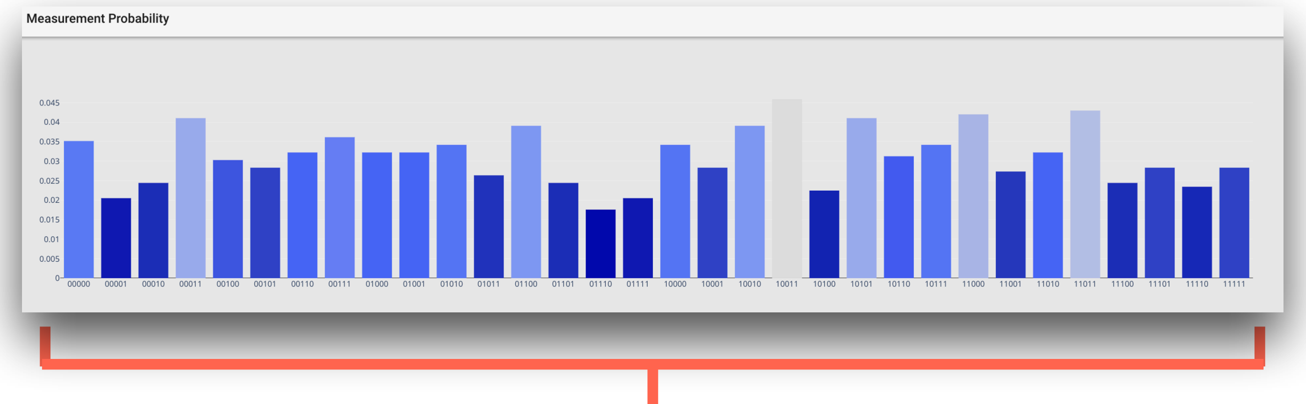
Generative approach

- General scheme of frame generation in video data



"Neighbor" frames

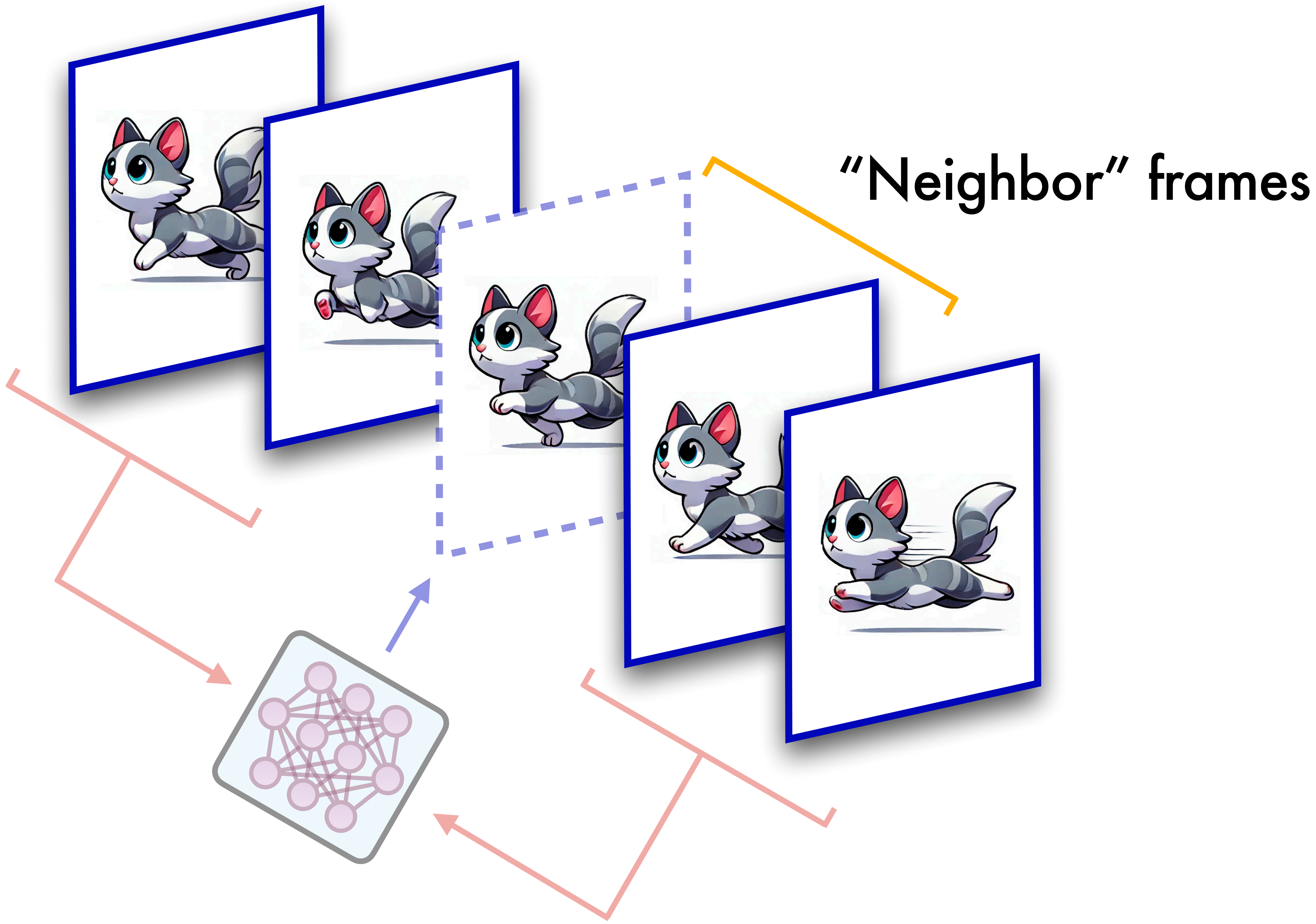
How about quantum?



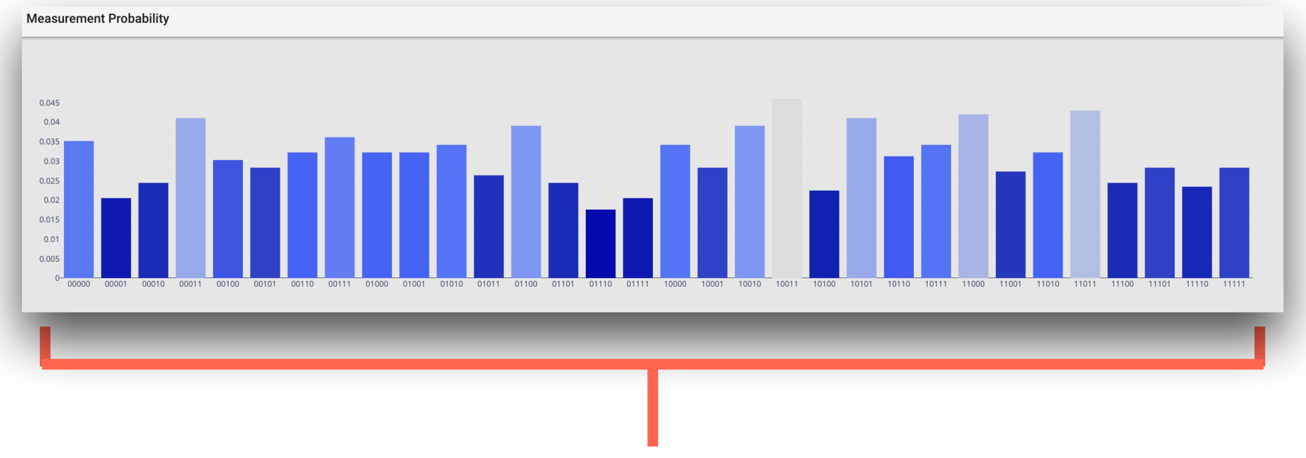
2^N measurement probabilities $\in [0,1]$

Generative approach

- General scheme of frame generation in video data



How about quantum?



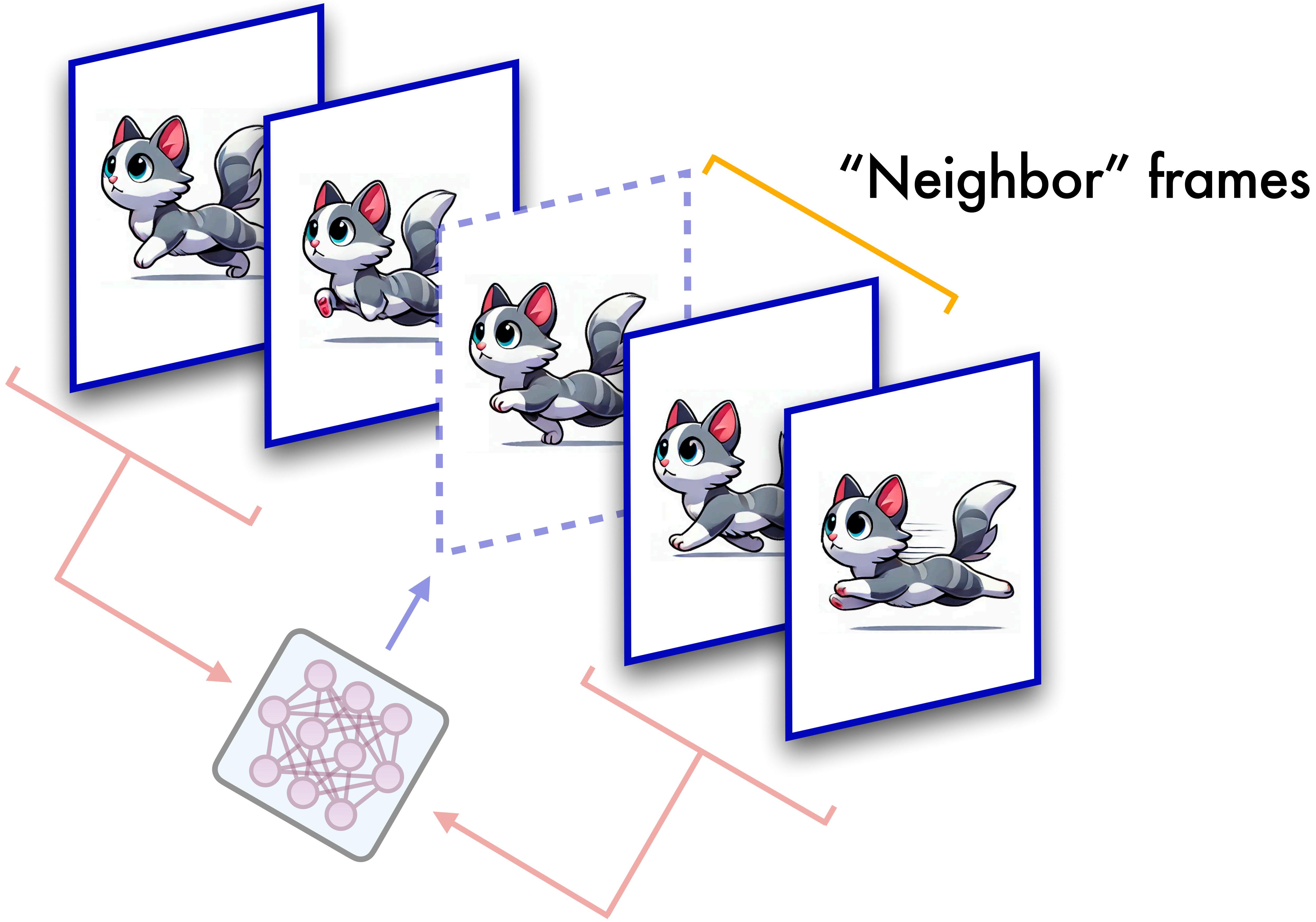
2^N measurement probabilities $\in [0,1]$

What is the similar role of "neighbor" in the Hilbert space (basis data) ?

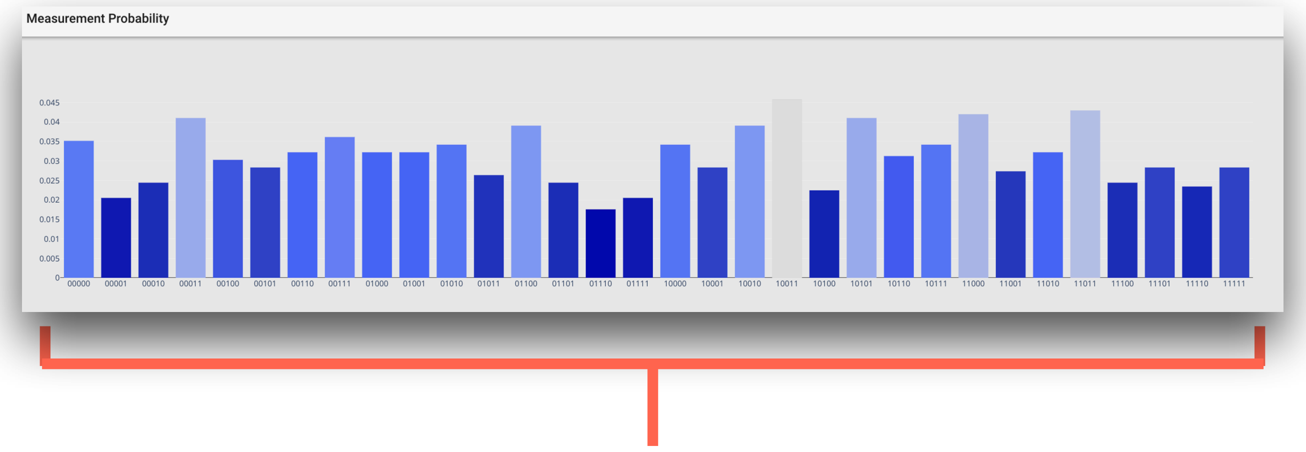
$\rightarrow$ Hamming distance

Generative approach

- General scheme of frame generation in video data



How about quantum?



2^N measurement probabilities $\in [0,1]$

What is the similar role of "neighbor" in the Hilbert space (basis data) ?

→ Hamming distance

Hamming distance = 1

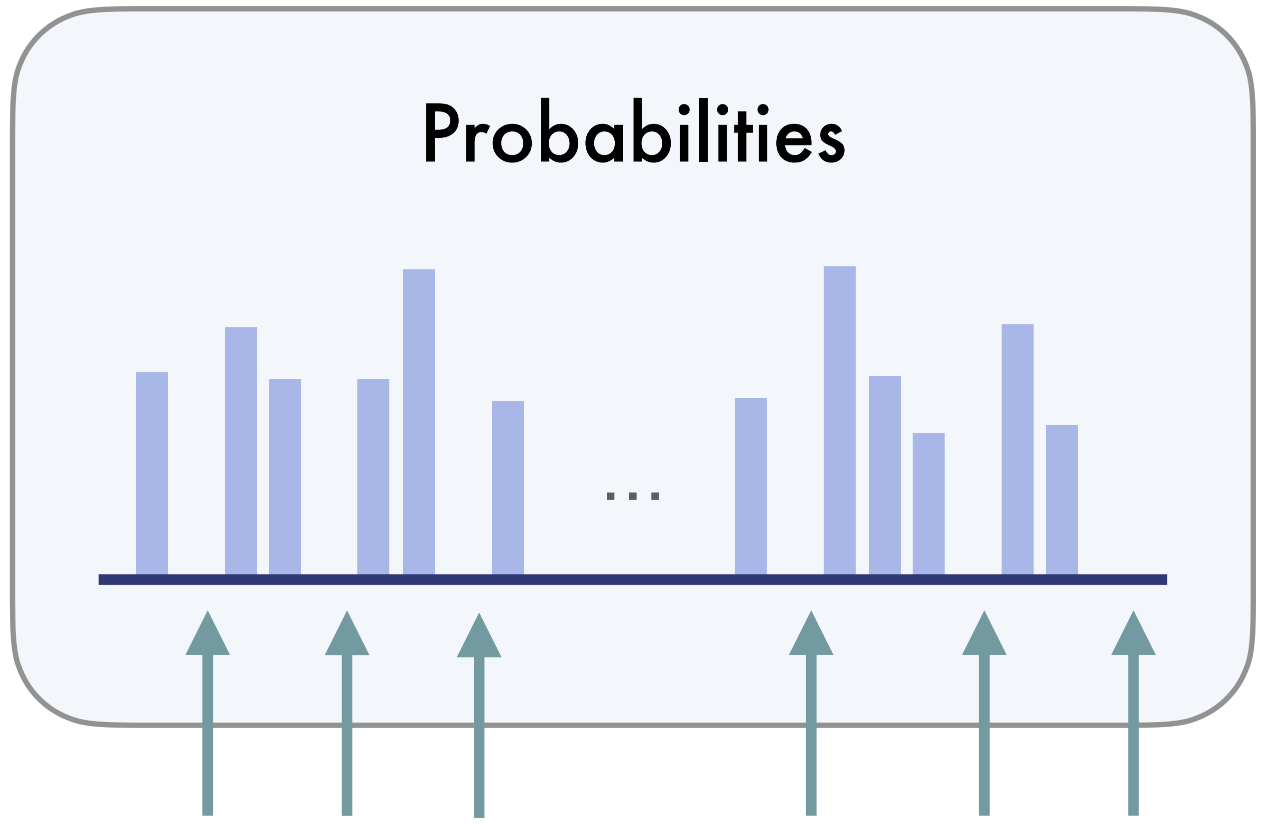
01010 ↔ 01110

Generative Interpolation (GI)

- The target for Generative Interpolation (GI)

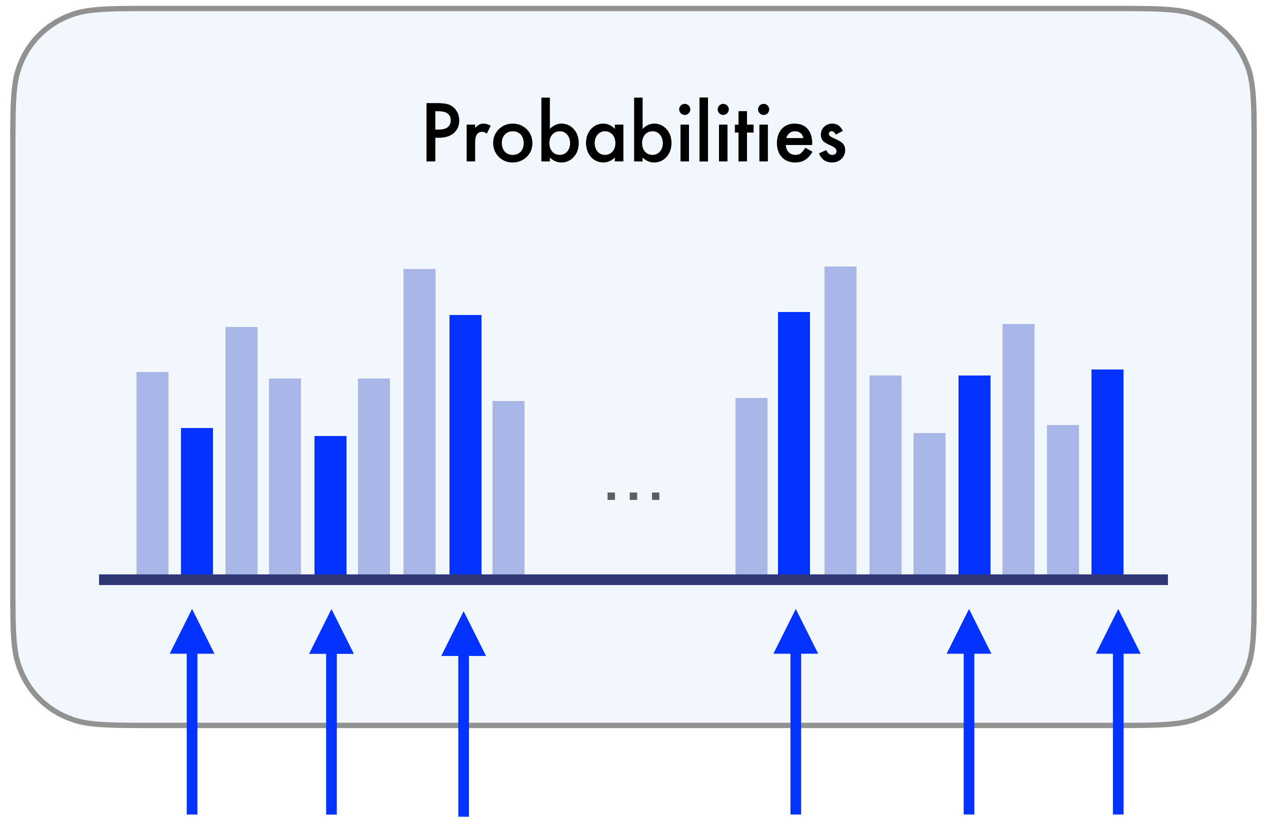
Finite Measurement shots

$(n_{shot} = c \cdot 2^N)$



Unmeasured basis due to insufficient measurement shots

Finite Measurement shots with Generative Interpolation

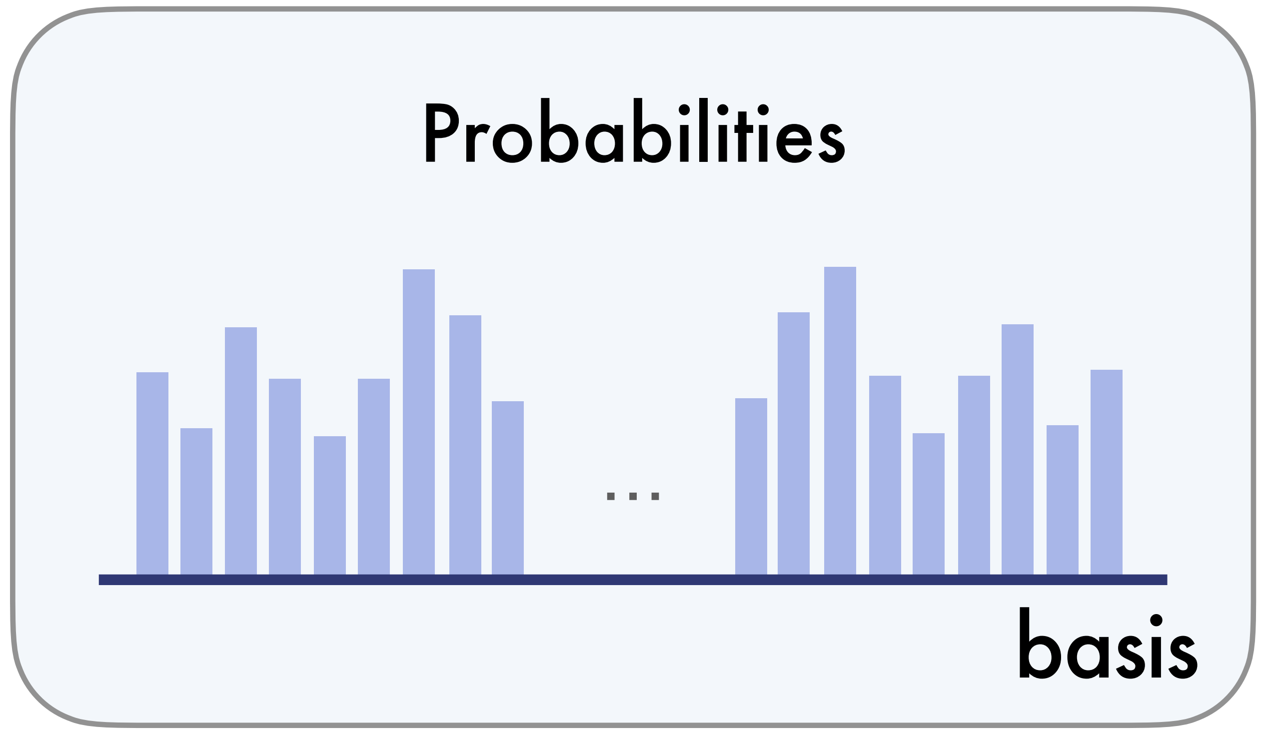


Generated measurement data

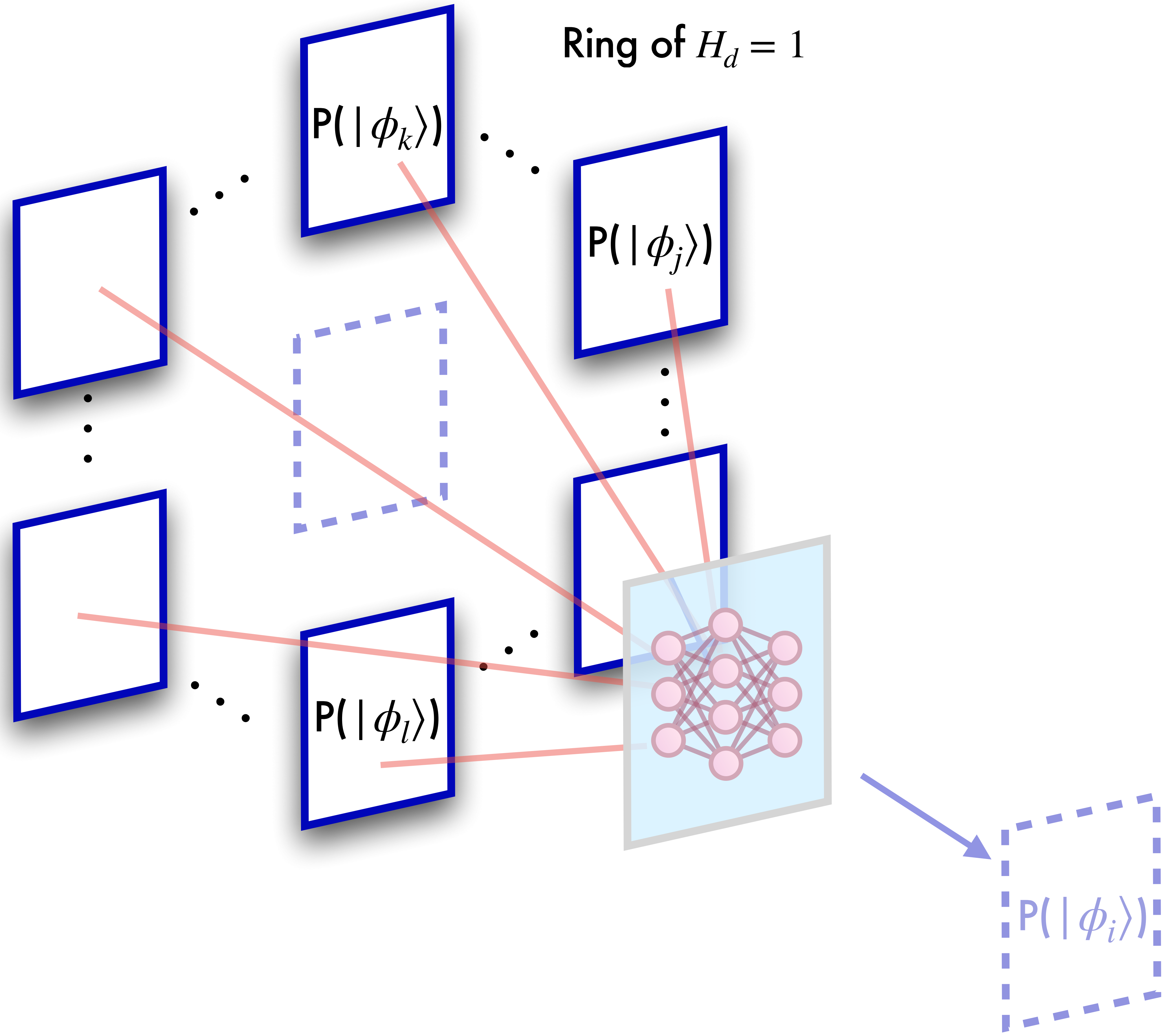
$\approx$

Exact Measurement

$(n_{shot} \to \infty)$



- Generative Interpolation (GI) of close Hamming distance data

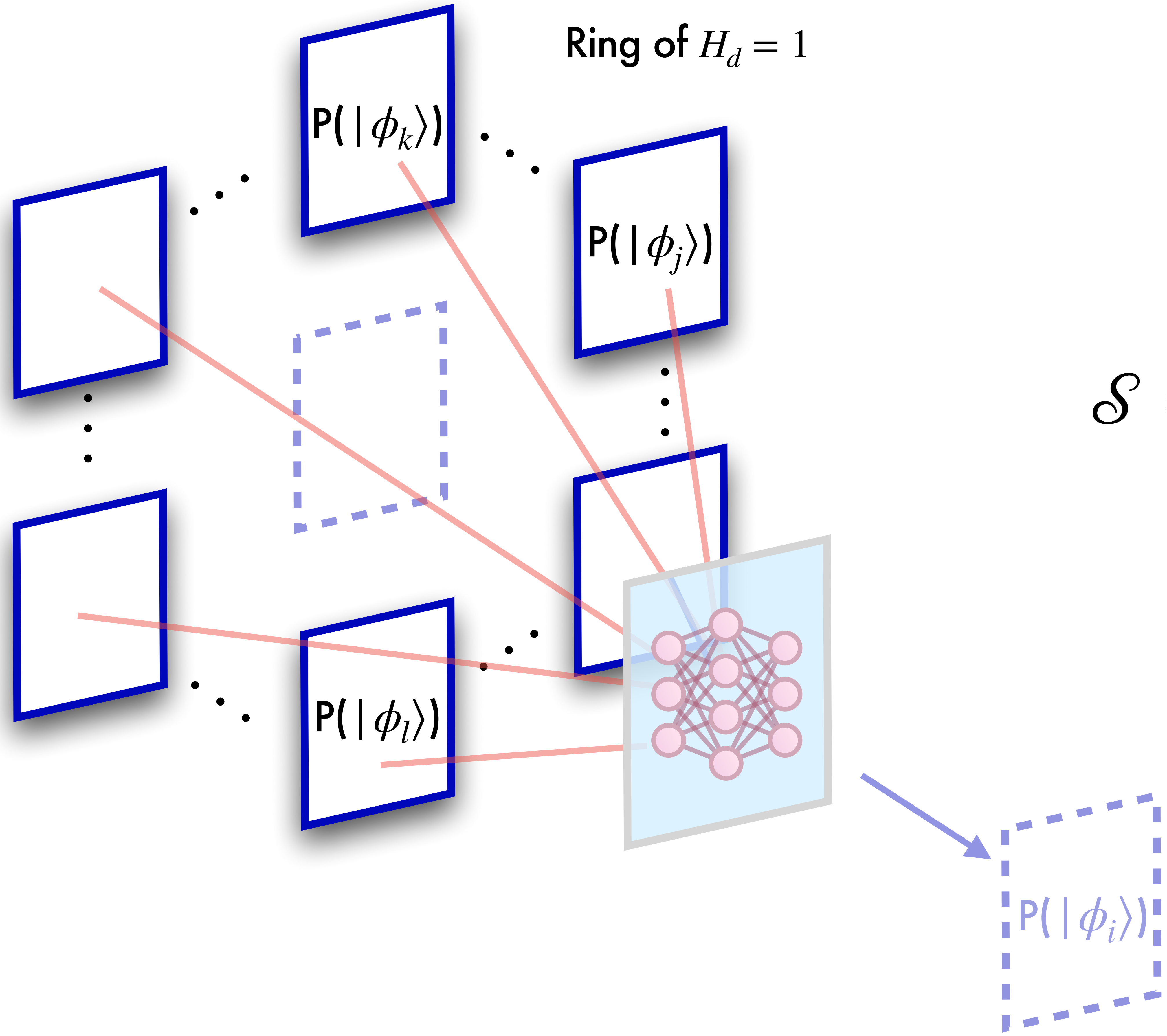


When a basis $|\phi_i\rangle$ is found to be unmeasured, P($|\phi_i\rangle$)=0, a basis subset $\{|\phi_k\rangle, |\phi_j\rangle, \ldots, |\phi_l\rangle\}$ is constructed by collecting the basis that have Hamming distance =1 with $|\phi_i\rangle$.

The corresponding measurement probabilities of basis in this subset, are then inputted in to a neural network model, where this model gives the estimation of P($|\phi_i\rangle$).

Generative Interpolation (GI)

- Generative Interpolation (GI) of close Hamming distance data



P($|\phi_i\rangle$)=0, $|\phi_i\rangle = |00000\rangle$

Set of Hamming distance = 1

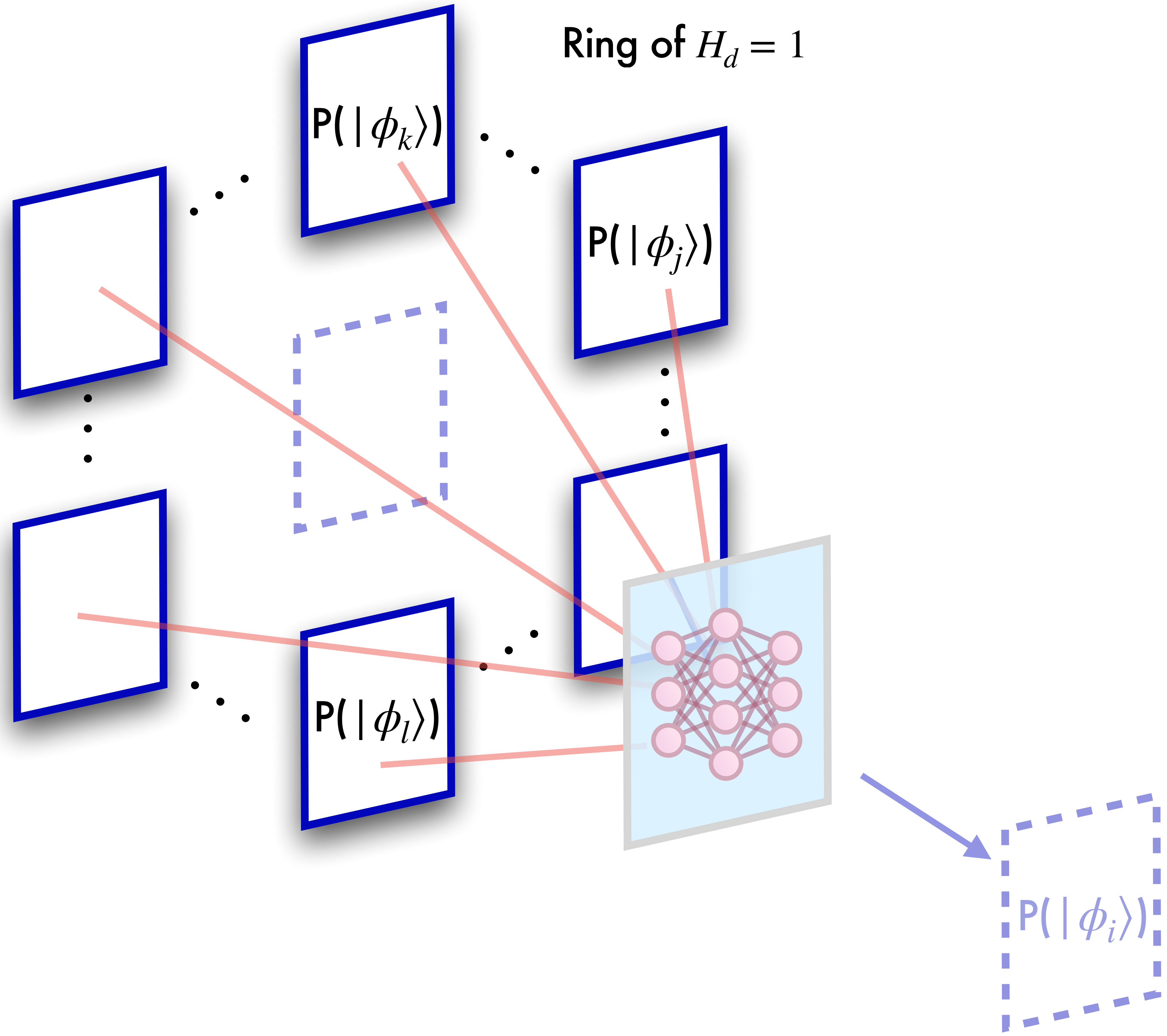
$$\mathcal{S} = \{|10000\rangle, |01000\rangle, |00100\rangle, |00010\rangle, |00001\rangle\}$$

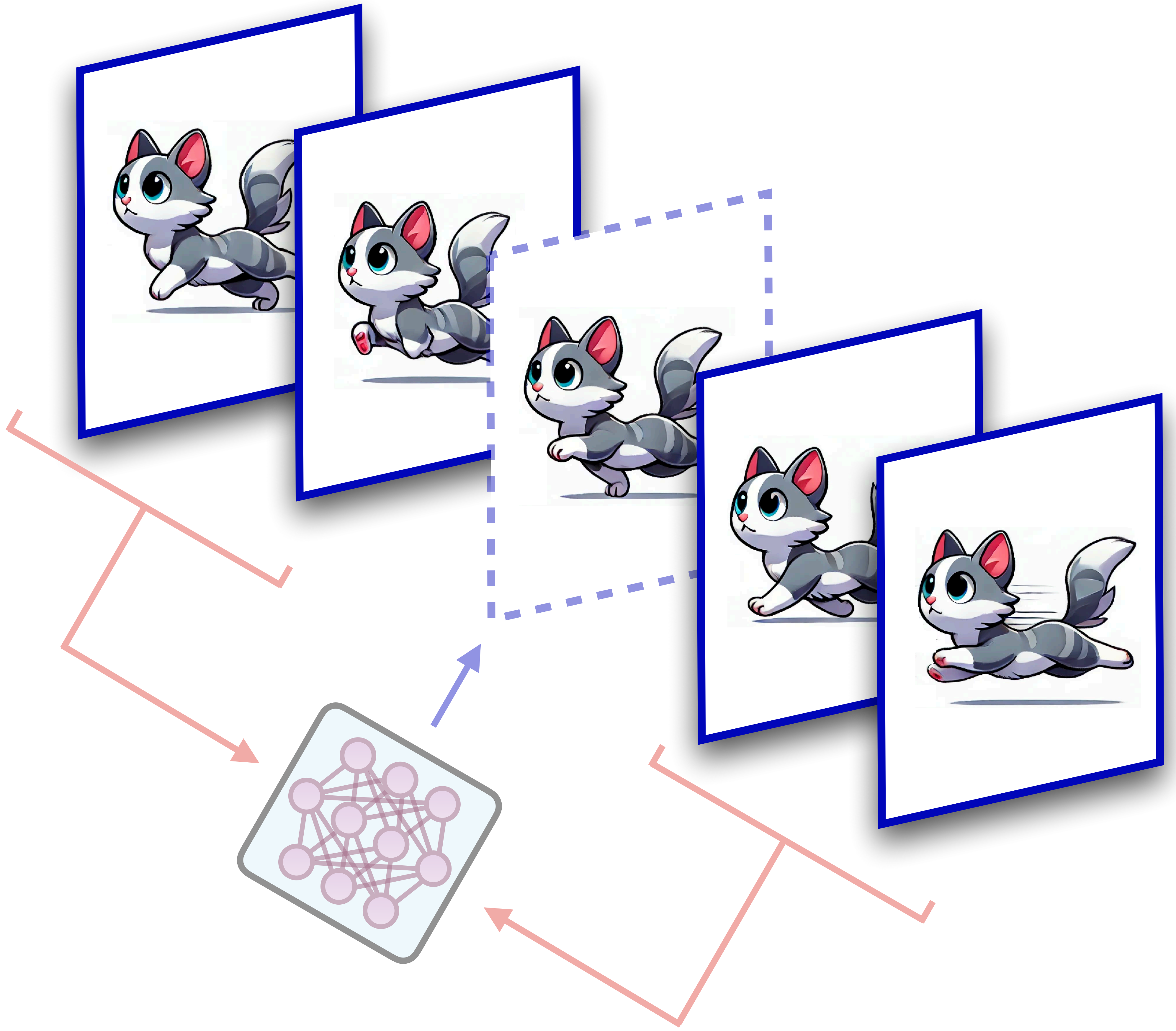
Neural network model F

P($|\phi_i\rangle$)= F(P($\mathcal{S}$))

The parameters of this neural network are learned along with the Quantum Parameter Adaptation process.

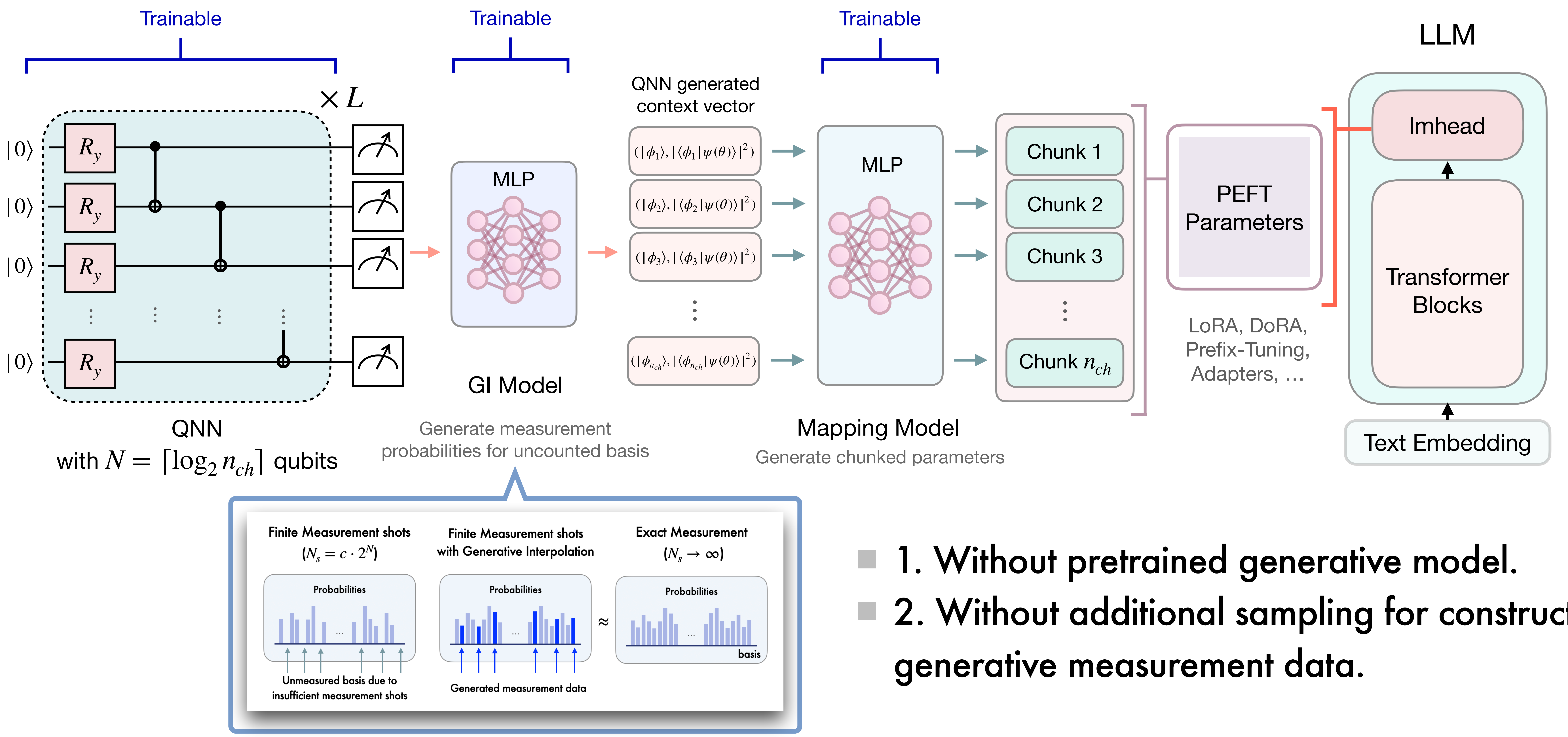
- Generative Interpolation (GI) of close Hamming distance data





Generative Interpolation (GI)

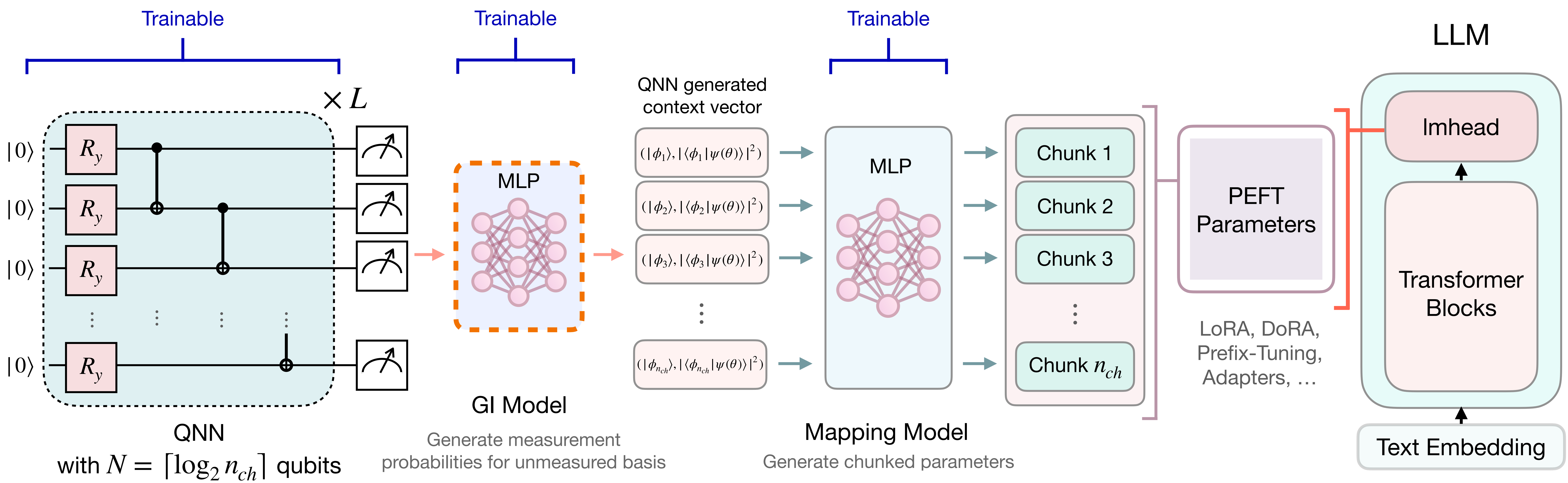
- Generative Interpolation (GI) enhanced Quantum Parameter Adaptation



- 1. Without pretrained generative model.
- 2. Without additional sampling for constructing generative measurement data.

ICLR International Conference On Learning Representations
國立臺灣大學 National Taiwan University

Generative Interpolation (GI) enhanced Quantum Parameter Adaptation



(This work)

Original QPA

Add random interpolation ?

Generative approach

- Result of Generative Interpolation (GI) for few measurement shots