Review Notes on Quantum LLMs

Quantum Methods Capable of Enhancing LLMs

Haojun Lin

Academy of Advanced Interdisciplinary Studies
Wuhan University

December 18, 2025

Contents

1 Introduction

2 Quantum-Inspired Techniques in LLMs
- Language Modeling
- Linear Algebra and Tensor Methods

3 Hybrid Quantum-Classical Architecture

4 Towards Quantum-Native LLMs

5 Conclusions

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Introduction

Background

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Definition

A language model architecture that integrates **quantum computing components** with classical LLMs.

Motivation for QLLMs

- Potential quantum speedup effects
- Advantages over classical algorithms **in certain tasks**
  - Quantum system simulation in chemistry;
  - RSA-related computational problems;
  - Language-related applications.

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Quantum-Inspired Techniques in LLMs

Quantum NLP: Model Families Overview

Reference: Varmantchaonala et al.(2024)[3]

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

- **Quantum Bag-of-Words**: quantum-probabilistic representations ignoring word order; efficient but limited.
- **Tensor Product Representations (TPR)**:
  - Positional TPR
  - Contextual TPR
  - Fock-space models: integrate positional and contextual information using harmony operators.
- **Word2Ket / Word2KetXS**

QBoWs and TPRs I

Quantum Bag-of-Words (QBoW) extends classical Bag-of-Words by encoding a document as a quantum state:

$$|d\rangle = \sum_i \delta_i |w_i\rangle, \qquad D = |d\rangle\langle d|.$$

The density matrix D captures term correlations through off-diagonal entries. QBoW is simple and enables quantum measurementbased document classification, but it ignores word order and syntactic structure.

Tensor Product Representations (TPR) aim to encode linguistic structure using tensor products.

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

QBoWs and TPRs II

- **Positional TPR (pTPR)**: bind word and its syntactic position:

$$s_i \otimes n_i.$$

This accurately reflects grammar but leads to rapid dimensional growth.

- **Contextual TPR (cTPR)**: bind a word with its contextual window:

$$|w_{i-1}\rangle \otimes |w_i\rangle \otimes |w_{i+1}\rangle.$$

This reduces structural mixing and controls tensor size, though it still increases dimensionality and cannot fully reflect long-range dependencies.

Word2Ket & Word2KetXS

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

(1) Word2Ket

- Decompose high-dim vector $|\mathbf{v}\rangle$ into tensor product of low-dim vectors:

$$|\mathbf{v}\rangle = \sum_{j=1}^{r} \bigotimes_{i=1}^{n} v'_{ij}$$

---

(2) Word2KetXS: Full Vocabulary Compression

- Extend Word2Ket to compress the entire sparse vocabulary matrix **M**:

$$\mathbf{M} = \sum_{j=1}^{r} \bigotimes_{i=1}^{n} \mathbf{M}'_{ij}$$

Chronological tree of QNLP models

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

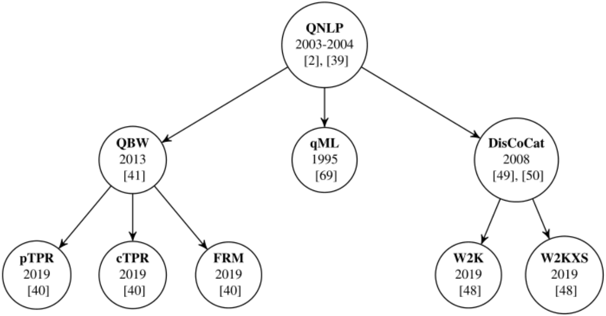


FIGURE 2. Chronological tree of QNLP models. qML stands for quantum machine learning.

Example of QLM

Reference: Basile et al.(2017)[4]

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

For a sequence $w = (w_1, \ldots, w_n)$, encode as $w \mapsto |w\rangle$.
Let $\Pi_w := |w\rangle\langle w|$ and define conditional steps:

$$\text{Init:} \quad P(w_1; \rho_0, U) = \mathrm{Tr}(\rho_0 \Pi_{w_1}),$$

$$\text{Projection:} \quad \rho_1' = \frac{\Pi_{w_1} \rho_0 \Pi_{w_1}}{\mathrm{Tr}(\Pi_{w_1} \rho_0 \Pi_{w_1})},$$

$$\text{Evolution:} \quad \rho_1 = U \rho_1' U^\dagger.$$

Termination:

$$P(w \mid \rho_0, U) = P(w_1; \rho_0, U) \prod_{i=2}^{n} P(w_i \mid w_{1,\cdots,i-1}; \rho_0, U).$$

Memory via entanglement

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

- Perplexity (to be minimised):

$$\Gamma(\rho_0, U) = \exp\left(-\frac{1}{C}\sum_{w\in S}\log P(w \mid \rho_0, U)\right).$$

Ancillary system

$$\mathcal{H}_2 = \mathcal{H}_{\mathrm{anc}} \otimes \mathcal{H}.$$

Projectors lift to $\Pi_w^{(2)} = I_D \otimes \Pi_w$.

Learning the unitary operator

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Quantum embedding via low-dimensional vectors. With idea following word2vec and GloVe, each word w is mapped through a fixed embedding:

$$w \mapsto (\alpha_1(w), \alpha_2(w), \ldots, \alpha_p(w)).$$

Learning a small set of base unitaries. The model learns only p base unitaries:

$$\{U_1, U_2, \ldots, U_p\}, \qquad U_i \in \mathbb{C}^{DN \times DN}.$$

Strengths and Limitations

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Strengths

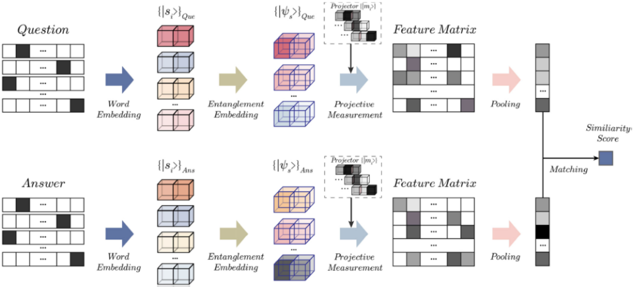
- Interference & entanglement $\rightarrow$ sequence-level correlations
- Compact trace formula for $P(w)$
- Fast inference once unitaries are precomputed

Limitations

- Scalability: DN grows with vocabulary size
- Current model works only for small vocabularies
- Training is computationally intensive (unitarity constraints + many parameters)

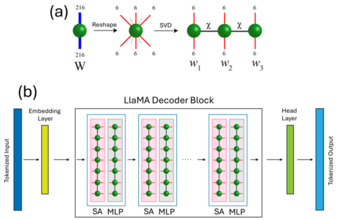
QLM-EE

Reference: Chen et al.(2021)[5]



CompactifAI

Reference: Tomut et al.(2024)[6], Vaswani et al.(2017)[7]



Weight Matrices in Transformer Layers

Self-Attention Linear Projections

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V.$$

Attention Output: $O = \mathrm{Attention}(Q, K, V)\, W_O.$
Feed-Forward Network

$$H = \sigma(XW_1 + b_1), \quad Y = HW_2 + b_2.$$

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Conceptual parallel (MPS)

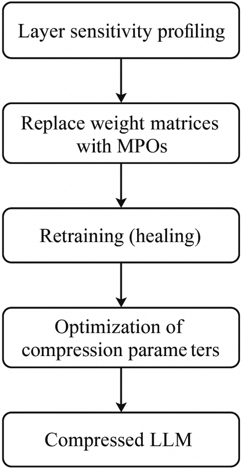
$$|\psi\rangle = \sum_{i_1, i_2, \ldots, i_N} A^{[1]}_{i_1} A^{[2]}_{i_2} \cdots A^{[N]}_{i_N} \, |i_1 i_2 \cdots i_N\rangle.$$

Matrix Product Operator, MPO

$$W_{i_1 \cdots i_L,\, j_1 \cdots j_L} = \sum_{\alpha_1, \ldots, \alpha_{L-1}} G^{(1)}_{i_1 j_1,\, \alpha_1} \, G^{(2)}_{i_2 j_2,\, \alpha_1 \alpha_2} \cdots G^{(L)}_{i_L j_L,\, \alpha_{L-1}}.$$

- The bond dimension χ: the compression ratio.

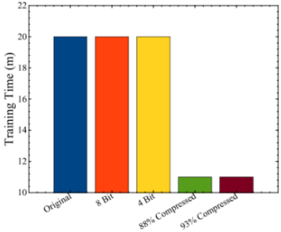
CompactifAI

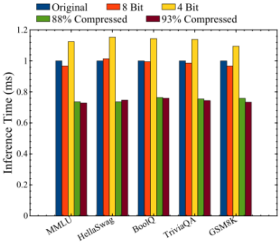


Experiment results

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

- 93% memory reduction; 70% parameter reduction
- approximately 25% faster inference; around 50% faster training
- a 2 - 3 % drop in accuracy





Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Hybrid Quantum-Classical Architecture

QLLM via Tensor Network Disentanglers

Reference: Aizpurua et al.(2024)[8]

Introduction

Quantum-Inspired Techniques in LLMs

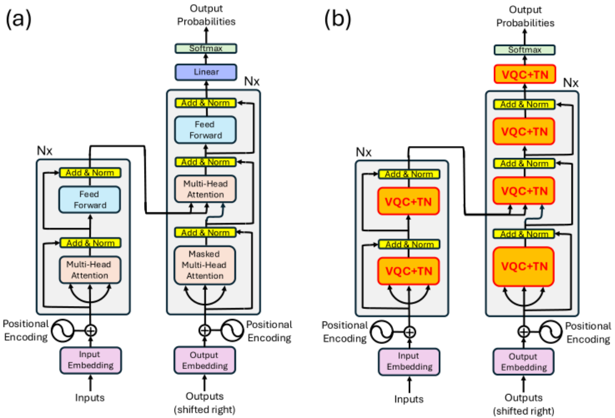
Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

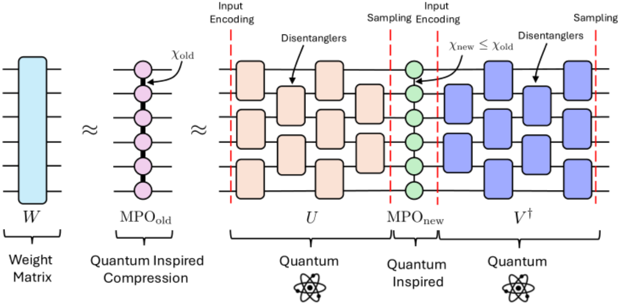
Towards Quantum-Native LLMs

Conclusions



Quantum Disentanglers

- $M = U_{\text{left}}^{\dagger} M_{\text{new}} U_{\text{right}}$.



Key Experimental Observations

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Observations

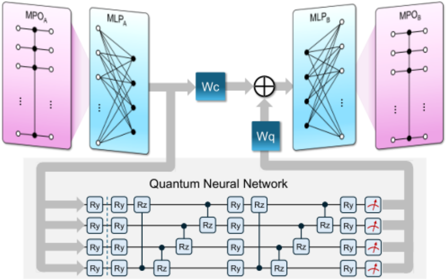
- Required bond dimension: **significantly reduced** $\Longrightarrow$ Quantum circuit effectively captures higher-order correlations.
- For the **same bond dimension**, the quantum-assisted MPO yields a **substantially lower reconstruction error**.
- Performance: remains very close to the original model.

Quantum enhanced fine tuning

Reference: Kong et al.(2025)[9]

Quantum Tensor Hybrid Adaptation,QTHA

- A parameter-efficient fine-tuning (PEFT) method;
- Dynamically adjusts feature weights through parameter tuning and outputs a combination of features from MPO and QNN.



State Revolution in QNN

$$|\Psi(x)\rangle = \bigotimes_{i=1}^{r} R_Z(x_i)\,|0\rangle,$$

$$|\phi(x)\rangle = U(\theta)|\psi(x)\rangle = \prod_{l=1}^{L}\left(\bigotimes_{i=1}^{q} RY(\theta_{l,i}) \cdot \mathrm{CZ}(\theta_{l,i,j})\right),$$

Expectation values

$$y_i = \langle\phi(x)|Z_i|\phi(x)\rangle \in [-1,1].$$

Main Contributions

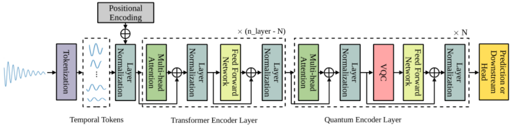
Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

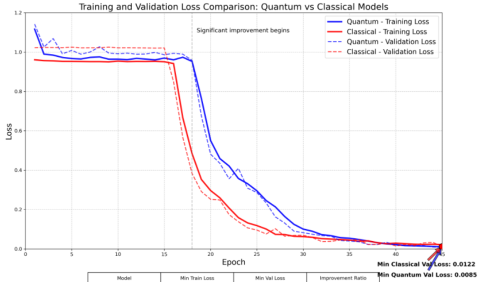
- First implementation of quantum computing inference for LLM on **quantum hardware**;
- Significant reduction in trainable parameters (**76**% compared to LoRA);
- Enhancing the performance of LLM fine-tuning.

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Quantum Enhanced Self Attention

Reference: Chen et al.(2025)[10]





Model	Min Train Loss	Min Val Loss	Improvement Ratio
Quantum	0.0106	0.0085	105.73
Classical	0.0219	0.0122	43.94

Key Observations:
1. Both models show similar training patterns but different convergence rates
2. The quantum model achieves lower final training and validation loss
3. The classical model shows more validation loss fluctuation after epoch 30

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Towards Quantum-Native LLMs

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

DisCoCat Framework

- **Complex Types**: Constructed via left cancellation n^l and right cancellation n^r. For example:
  - Adjectives: with type $n \cdot n^l$.
  - Transitive verbs: with type $n^r \cdot s \cdot n^l$.
- **Syntactic Parsing**

$$\text{Cats}(n) \qquad \text{eat}(n^r \cdot s \cdot n^l) \qquad \text{fish}(n)$$

$$\mathrm{Meaning}(\mathit{sentence}) = \epsilon_N \otimes 1_S \otimes \epsilon_N(\overrightarrow{\mathrm{Cats}} \otimes \overrightarrow{\mathrm{eat}} \otimes \overrightarrow{\mathrm{fish}})$$

Syntactic structure $\Longrightarrow$ Tensor network structure $\Longrightarrow$ Quantum circuit structure

Quantum-Native Self Attention I

Reference: Shi et al.(2025)[12]

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

$$x_i \longmapsto |x_i\rangle \xrightarrow{U_{\text{att}}} |\mathrm{Att}(x_i)\rangle$$

1. **Quantum Embedding**

$$x_i \in \mathbb{R}^d \;\mapsto\; |x_i\rangle \in \mathbb{C}^{2^n}$$

2. **Quantum Key–Query Inner Product**

$$q_i k_j^\top$$

SWAP-test / Hadamard-test / parameterized circuits:

$$\mathrm{sim}(i,j) = \langle q_i | k_j \rangle$$

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

3. **Quantum Attention Weighting** Parameterized controlled rotation:

$$\mathrm{softmax}(s_{ij}) \approx f_\theta(s_{ij})$$

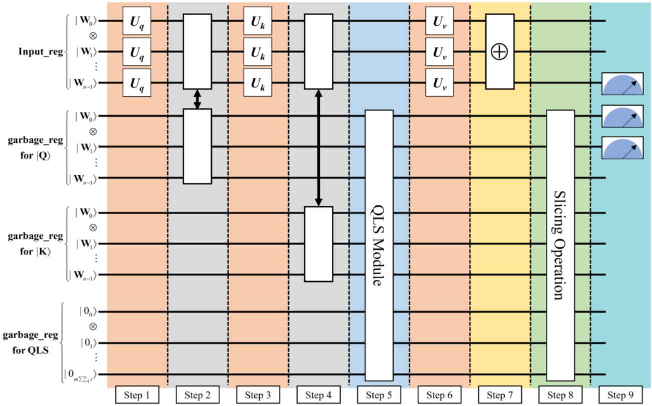
Final attention state:

$$|\mathrm{Att}(x_i)\rangle \propto \sum_j f_\theta(s_{ij}) \, |v_j\rangle$$

Quantum-Native Self Attention III

Reference: Shi et al.(2025)[12]

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions



Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Conclusions

Future Outlook

Reference: Pan et al.(2025)[13]

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Based on AQCF Framework (Bridging Classical and Quantum Computing)

- Gradual transition: Classical LLM $\rightarrow$ Hybrid Q-Enhanced $\rightarrow$ Quantum-Native
- Principles: progressive, scalable, hardware-feasible
- Focus: shallow quantum modules; adaptive circuits; noise-aware design

Potential Strategic Directions

- Scalable quantum text processing for NISQ
- Quantum-enhanced attention
- Quantum memory and semantic retrieval

References I

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield.
Quantum-like contextuality in large language models.
Proceedings of the Royal Society A, 481(2319):20240399, 2025.

Michael A. Nielsen and Isaac L. Chuang.
Quantum Computation and Quantum Information: 10th Anniversary Edition.
Cambridge University Press, New York, NY, USA, 2010.

Charles M Varmantchaonala, Jean Louis KE Fendji, Julius Schöning, and Marcellin Atemkeng.
Quantum natural language processing: A comprehensive survey.
IEEE Access, 12:99578–99598, 2024.

Ivano Basile and Fabio Tamburini.
Towards quantum language models.
In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1840–1849, 2017.

Yiwei Chen, Yu Pan, and Daoyi Dong.
Quantum language model with entanglement embedding for question answering.
IEEE Transactions on Cybernetics, 53(6):3467–3478, 2021.

References II

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Andrei Tomut, Saeed S Jahromi, Abhijoy Sarkar, Uygar Kurt, Sukhbinder Singh, Faysal Ishtiaq, Cesar Muñoz, Prabdeep Singh Bajaj, Ali Elborady, Gianni del Bimbo, et al.
Compactifai: extreme compression of large language models using quantum-inspired tensor networks.
arXiv preprint arXiv:2401.14109, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
Advances in neural information processing systems, 30, 2017.

Borja Aizpurua, Saeed S. Jahromi, Sukhbinder Singh, and Roman Orus.
Quantum large language models via tensor network disentanglers, 2024.

Xiaofei Kong, Lei Li, Zhaoyun Chen, Cheng Xue, Xiaofan Xu, Huanyu Liu, Yuchun Wu, Yuan Fang, Han Fang, Kejiang Chen, et al.
Quantum-enhanced llm efficient fine tuning.
arXiv preprint arXiv:2503.12790, 2025.

References III

Introduction

Quantum-Inspired Techniques in LLMs

Language Modeling

Linear Algebra and Tensor Methods

Hybrid Quantum-Classical Architecture

Towards Quantum-Native LLMs

Conclusions

Chi-Sheng Chen and En-Jui Kuo.
Quantum adaptive self-attention for quantum transformer models.
arXiv preprint arXiv:2504.05336, 2025.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark.
Mathematical foundations for a compositional distributional model of meaning.
arXiv preprint arXiv:1003.4394, 2010.

Jinjing Shi, Ren-Xin Zhao, Wenxuan Wang, Shichao Zhang, and Xuelong Li.
Qsan: A near-term achievable quantum self-attention network.
IEEE Transactions on Neural Networks and Learning Systems, 2024.

Yi Pan, Hanqi Jiang, Junhao Chen, Yiwei Li, Huaqin Zhao, Lin Zhao, Yohannes Abate, Yingfeng Wang, and Tianming Liu.
Bridging classical and quantum computing for next-generation language models.
arXiv preprint arXiv:2508.07026, 2025.

Introduction
Quantum-Inspired Techniques in LLMs
Language Modeling
Linear Algebra and Tensor Methods
Hybrid Quantum-Classical Architecture
Towards Quantum-Native LLMs
Conclusions

Thank You!