# Quantum-Inspired Sentence Representation: Rethinking Word-Based Density Matrices

Anonymous ACL submission

## Abstract

This paper proposes a novel approach to enhance traditional quantum-inspired models. We introduce a Quantum-Inspired Sentence Representation model (QISR), which transforms word density matrices into representations of entire sentences, improving computational resource efficiency. Compared with traditional quantum-inspired models, the QISR method works at the density matrix layer and has better effects on the overall model as the embedding dimension increases. Even the QPDN model with a word embedding of 768 dimensions only requires 1736MB. This optimization has potential benefits for the overall model architecture, particularly when dealing with large word embedding dimensions. Furthermore, this approach reduces computing resource consumption while maintaining high computational accuracy, highlighting its potential benefits in processing complex language tasks. This research provides a novel approach to sentence representation in quantum-inspired language models and highlights the potential value of improved computational methods in a quantum-inspired context. Our research results are expected to provide modeling support and practical application guidance for future text processing endeavors.

## 1 Introduction

In recent years, traditional quantum-inspired models have primarily focused on the post-hoc interpretability and transparency(Lipton, 2018). Post-hoc interpretability refers to the ability of a model to explain how it worked after it has been executed, while transparency involves self-explanation during the model design phase.

Meanwhile, to meet these needs for clarity and transparency, early quantum language models (QLM)(Sordoni et al., 2013) used density matrices to represent words, capturing word connections in text. Further advancements in this domain, such as the extension of quantum language models to the field of neural networks and the introduction of end-to-end quantum language model (NNQLM)(Zhang et al., 2018a). The NNQLM employs word embeddings for representation and introduces a density matrix computation method for both word and sentence representations. The CNM(Li et al., 2019) model, in the process of converting words to word embeddings, simulates the construction of quantum states through phase embeddings and obtains complex value representations of quantum states using Euler's formula. Additionally, the QINM(Jiang et al., 2020) model enhances interpretability by extensively interacting with queries and documents and using reduced density matrices to model quantum interference between them, making the retrieval process somewhat more in line with human cognition. In addition, there are also tensor networks that serve as a bridge between neural networks and quantum mechanics and have demonstrated good interpretability in processing natural language tasks(Zhang et al., 2018b, 2020a).

Furthermore, apart from the aforementioned models, other significant multimodal models have emerged. The QPM(Tomar et al., 2023) framework includes a complex-valued multimodal representation encoder, a quantum-like fusion network, and a quantum measurement mechanism designed for joint detection of multimodal sarcasm and sentiment. In contrast, the QUIET(Liu et al., 2023) framework is a quantum probability-based multimodal analysis framework specialized in processing text, images, and audio data while considering intermodal correlations to comprehensively analyze sentiment, irony, and emotion across multiple data types. Furthermore, in the field of multimodal analysis, there have emerged methods(Gkoumas et al., 2021b,a; Li et al., 2021; Zhang et al., 2020b; Liu et al., 2021) based on the concepts of quantum entanglement and quantum interference. The development of these methods has further accentuated the importance of quantum-inspired models

in multimodal data analysis, especially in terms of enhancing the interpretability and transparency of the models, thereby providing new perspectives for practical applications.

These quantum-inspired models draw inspiration from quantum mechanics concepts like quantum interference and superposition, aiming to provide explanations for how they work after model execution. They have been applied to various tasks, including information retrieval(Sordoni et al., 2013; Jiang et al., 2020), sentiment analysis (Zhang et al., 2019), and question-answer matching(Li et al., 2019), etc. In previous methods, words in text data are regarded as pure states, and sentences are regarded as mixed states formed by these words. At this time, we inevitably face a challenge: as the word dimension increases, building a density matrix will significantly increases time and computational cost. Therefore, we must rethink how to improve computational efficiency in quantum-inspired frameworks while ensuring interpretability and transparency of model outputs. This challenge formed the motivation for our research.

In this paper, we first theoretically demonstrate the feasibility of representing sentences as mixed states directly in Hilbert space, and propose a new quantum-inspired sentence representation model (QISR) that aims to significantly improve the computational efficiency of existing models. By conducting experiments on different quantum-inspired models, we verify the significant improvements in time and memory efficiency of the QISR model on both CPU and GPU.

The key innovations of this research include:

1. Computational Efficiency: This study introduces a sentence-based density matrix approach that changes the order of computation within a layer. This approach reduces floating-point operations (FLOPs) within the density matrix layer by approximately one-half and one-third in real-valued and complex-valued models, respectively. Furthermore, it effectively exploits the parallel computing properties of matrices and hence significantly reduces the computational time of the entire model. Please see table 3 for specific indicators.

2. Memory reduction: In the construction of the density matrix layer, memory consumption is reduced by n times (n representing sentence length) by changing the calculation operations. Overall, as the dimensionality of word embeddings increases, the memory savings become more substantial. This reduction plays a crucial role in alleviating the bottleneck of rising computational costs in traditional quantum-inspired models with large word embedding sizes.

3. Model Adaptability: The QISR model is readily integrable with existing density matrix quantum-inspired models, including monomodal, multimodal, and complex-valued models. This demonstrates its high scalability and practical applicability.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the foundational knowledge related to QISR. Section 3 demonstrates on the advantages of QISR in sentence representation compared to word representation, along with its theoretical explanation. In Section 4, we conducted detailed experimental analysis. Finally, Section 5 concludes this paper and discusses future research directions.

## 2 Background

The quantum-inspired approach is an emerging research direction in the field of natural language processing, drawing on key concepts from quantum mechanics. It enhances the post-hoc interpretability and transparency of models while also offering the potential for improved performance in text processing or multimodal tasks. This section will briefly introduce fundamental theoretical concepts relevant to our research.

### 2.1 Quantum Probability

Quantum probability theory is a generalized probability theory developed by John von Neumann based on linear algebra, with the aim of providing a mathematical foundation for quantum theory. In quantum probability, quantum probability space is defined within the complex Hilbert space $\mathbb{H}$. In this paper, Hilbert space refers to a finite-dimensional inner product space, which is widely used in mathematical analysis and quantum mechanics.

### 2.2 Quantum Superposition

Quantum superposition is a fundamental concept in quantum mechanics that describes the phenomenon

where a quantum system can exist in a superposition of multiple base states under certain conditions. In classical physics, we usually think of objects (such as particles) as having well-defined properties (such as their positions and velocities). However, in quantum physics, the states of particles are inherently uncertain. In quantum mechanics, it is possible to form a superposition state by linearly combining multiple base states. In a two dimensional system, its base states can be represented by $|0\rangle$ and $|1\rangle$. Therefore, the superposition state of a quantum system can be expressed as follows:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

where $|\psi\rangle$ represents the quantum state, and $\alpha$ and $\beta$ are complex amplitudes (complex numbers). $|0\rangle$ and $|1\rangle$ are orthogonal base vectors in a two dimensional Hilbert space, representing the two possible states of a quantum bit.

### 2.3 Quantum Mixed State

Quantum mixed states describe that a quantum system under certain conditions can be composed of a mixture of different pure states, where each pure state is determined by its associated probability weight. The mixed state of a quantum system can be represented as follows:

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i| \tag{2}$$

where $\rho$ represents the quantum mixed state, $p_i$ represents the probability of each pure state $|\psi_i\rangle$, expressed as weights. This method allows us to probabilistically describe multiple possible pure states and their statistical mixtures, contrasting with the deterministic nature of individual pure states $|\psi_i\rangle$.

### 2.4 Measurement

Quantum systems can be in mixed or pure states. A mixed state represents a statistical mixture of multiple quantum states and is described by the density matrix, which is a positive semidefinite matrix and Hermitian matrix. In contrast, a pure state can be described by a state vector even if it is superposed. In order to observe the properties or state information of this system within the state space, it involves a set of operators called measurement operators $\{M_i\}$. When a measurement is performed, the system will collapse to a specific state corresponding to the measurement result with a certain probability:

$$|\psi\rangle \rightarrow \frac{M_i|\psi\rangle}{\sqrt{\langle\psi|M_i^\dagger M_i|\psi\rangle}} \tag{3}$$

where $|\psi\rangle$ represents the initial state of the system, $M_i$ is one of the measurement operators in the set, $M_i^\dagger$ is its adjoint operator, and $p_i = \langle\psi|M_i^\dagger M_i|\psi\rangle$ is the probability of obtaining the measurement result $i$.

## 3 Sentence Representation

### 3.1 Word-Based Density Matrices Representation

Inspired by quantum theory, existing quantum-inspired models that construct density matrices word-based embeddings all utilize the concept of quantum mixed states as shown in Figure 1-a, which is the most direct process for constructing density matrices based on the quantum mixed state, specifically as depicted in Formula 2. Even models like CNM and QPDN(Wang et al., 2019), as shown in Figure 1-b, begin by obtaining quantum states through amplitude-phase relationships and Euler's formula before constructing the density matrix. These models can be broadly viewed as transforming word embeddings into the form of density matrices.

### 3.2 Sentence-Based Density Matrices Representation

Compared to traditional Convolutional Neural Networks (CNN) (LeCun et al., 1989) and Long Short-Term Memory networks (LSTM) (Shi et al., 2015) and their derivative models, our approach offers greater directness and clarity in terms of interpretability and transparency. However, this process from words to sentences using density matrices requires a large amount of computing resources, resulting in high computational costs. Therefore, we explored whether efficiency could be improved by directly constructing sentence-based density matrices. Our QISR model demonstrates the potential of this method, as illustrated in Figures 1-c and 1-d, utilizing matrix properties in quantum-inspired models to accelerate processing speed and reduce computational resource usage. In subsequent sections, we will mathematically demonstrate the theoretical equivalence of the sentence-based and word-based density matrix approaches.

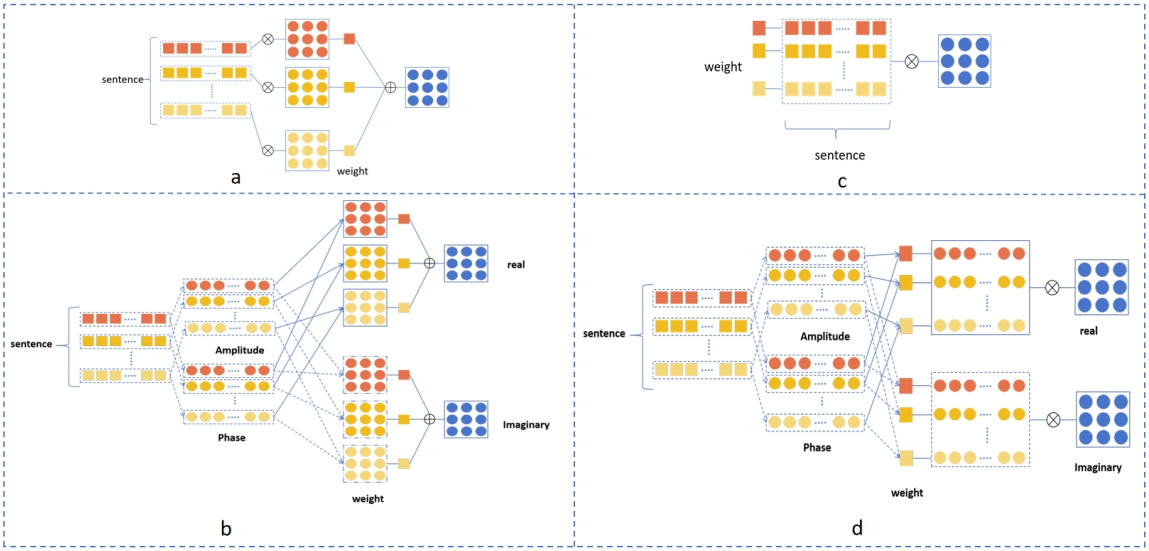


Figure 1: In the figure, $\otimes$ represents the tensor product operation, and $\oplus$ represents the summation operation. Figure 1-a shows the construction of a classical quantum-inspired model based on the density matrix, primarily using word-based embeddings to encode inputs, as shown in Formula 2. Figure. 1-b shows the construction of the density matrix in complex-valued quantum-inspired models, inspired by Euler's formula. Figure 1-c shows the process of the QISR model using sentence embeddings to build a density matrix in the classical quantum-inspired model. Figure 1-d shows the use of QISR in the complex-valued quantum-inspired model to construct the density matrix.

### 3.3 Justification and Comparative Analysis

#### 3.3.1 Algorithm Equivalence Proof

In this section, we will conduct a theoretical analysis to compare classical quantum-inspired architectures, such as those shown in Figure 1-a. We will theoretically evaluate the computational costs of transforming word-based density matrices into sentence-based density matrices.

Consider a sentence consisting of $n$ words, denoted as $\{w_1, w_2, \ldots, w_n\}$. For each word $w_i$, its corresponding embedding is a vector in a $d$-dimensional space, represented as $e_i = \{e_{i1}, e_{i2}, \ldots, e_{id}\}$, where each $e_{ij}$ represents the $j^{th}$ component of the embedding vector for the $i^{th}$ word and corresponding weight coefficients $p$, where $\sum_{i=1}^{n} p_i = 1$.

First, we start with the word-based density matrix, we need to perform a density matrix operation on this sentence, i.e., Equation 2. This initially involves the outer product operation of word embeddings.

$$|w_i\rangle\langle w_i| = \begin{bmatrix} e_{i1} \cdot e_{i1} & \ldots & e_{i1} \cdot e_{id} \\ e_{i2} \cdot e_{i1} & \ldots & e_{i2} \cdot e_{id} \\ \vdots & \ddots & \vdots \\ e_{id} \cdot e_{i1} & \ldots & e_{id} \cdot e_{id} \end{bmatrix} \quad (4)$$

where $w_i$ refers to the $i$-th ($1 \leq i \leq n$) word in the sentence.

Next, we proceed with the remaining operations in Formula 2, multiplying the current $|w_i\rangle\langle w_i|$ by a coefficient $p_i$, and finally summing them up.

$$\sum_{i=1}^{n} p_i |w_i\rangle\langle w_i| = \begin{bmatrix} \psi_{11} & \psi_{12} & \ldots & \psi_{1d} \\ \psi_{21} & \psi_{22} & \ldots & \psi_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{d1} & \psi_{d2} & \ldots & \psi_{dd} \end{bmatrix} \quad (5)$$

where $\psi_{ij}$ ($1 \leq i, j \leq d$) represents a value in the density matrix of a sentence composed of $n$ words, given by $p_1 e_{1i} e_{1j} + p_2 e_{2i} e_{2j} + \ldots + p_n e_{ni} e_{nj}$.

As we commence the construction of the sentence density matrix, it is important to note that our approach differs slightly from the word-based construction method. We begin by performing the multiplication of word embeddings with their respective weight coefficients, denoted as

$$word\_emb = q_i |w_i\rangle \quad (6)$$

where $q_i$ represents the weight coefficients of individual words prior to constructing sentence embeddings, and $q_i = \sqrt{p_i}$ with the constraint $\sum_{i=1}^{n} q_i^2 = 1$.

Next, we convert the embedding representation of each word into matrix form, where the embedding vector of each word becomes a row in the

matrix. This process creates a sentence embedding of dimension (n * d), denoted as

$$s\_e = \begin{bmatrix} q_1 e_{11} & q_1 e_{12} & \cdots & q_1 e_{1d} \\ q_2 e_{21} & q_2 e_{22} & \cdots & q_2 e_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_n e_{n1} & q_n e_{n2} & \cdots & q_n e_{nd} \end{bmatrix} \quad (7)$$

Next, we will perform a matmul product operation on sentence embedding, denoted as

$$s\_e^\top \cdot s\_e = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \cdots & \Psi_{1d} \\ \Psi_{21} & \Psi_{22} & \cdots & \Psi_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{d1} & \Psi_{d2} & \cdots & \Psi_{dd} \end{bmatrix} \quad (8)$$

where $\Psi_{ij}$ is defined as $q_1^2 e_{1i} e_{1j} + q_2^2 e_{2i} e_{2j} + \ldots + q_n^2 e_{ni} e_{nj}$.

From Equation 6, it can be derived that $\Psi_{ij}$ is also defined as $p_1 a_i a_j + p_2 b_i b_j + \ldots + p_n n_i n_j$. Therefore, we arrive at a conclusion that the results obtained from the density matrices constructed from words and from sentences are consistent.

#### 3.3.2 Memory Cost Comparison

In this section, we focus on comparative analysis of the differences in memory costs between the two methods. For the word-based density matrix construction process, Equation 5 shows that under the assumption that the density operator representation of each word requires computation, the minimum storage space for effectively expressing word-based density matrix information is $n \cdot d \cdot d$. Conversely, in the case of QISR construction, as shown in Equations 7 and 8, the minimum storage requirement is between $d^2$ and $n \cdot d$. In actual scenarios where $n$ is usually less than $d$, the actual storage space required converges to $d^2$.

This represents that during the model training process, when constructing the density matrix using the QISR approach, the reduction in memory overhead can be up to a factor of "sentence length". The reduction in memory overhead is especially important when word embedding dimensions are large, as it reduces the need for memory resources. This not only enhances computational efficiency but also enables the model to better handle various text processing tasks. This optimization holds significant importance for improving performance and reducing costs.

#### 3.3.3 Computational Cost Comparison

In this section, we conduct a detailed analysis of the computational costs involved in constructing real-valued and complex-valued density matrices. The key metrics for this analysis are FLOPs and parallelization efficiency. It is important to note that the computational processes for both real-valued and complex-valued density matrix construction are similar.

First, we compare different models in terms of FLOPs. In traditional quantum-inspired models, the real-valued model requires $2n \cdot d \cdot d$ FLOPs (Figure 1-a), while the complex-valued model consumes $6n \cdot d \cdot d$ FLOPs for complex embedding operations (Figure 1-b).

Then, we assess the QISR-based models. Differing from traditional approaches, each word is multiplied by weight coefficients before constructing the density matrix for the whole sentence. In the QISR framework, the real-valued version needs $n \cdot d + n \cdot d \cdot d$ FLOPs (Figure 1-c), and the complex-valued version requires $4n \cdot d \cdot d + 2n \cdot d$ FLOPs (Figure 1-d).

Overall, in the real-valued model, the FLOPs for constructing the density matrix layer in the QISR method are approximately $1/2 - 1/(2d)$ of the traditional method, especially when $d$ is large, approximating to half of the traditional method. In the complex-valued model, the FLOPs for the QISR method are roughly $2/3 + 1/(3d)$ of the traditional method, approximating to two-thirds for large values of $d$.

Finally, we conducted a thorough analysis of the parallelization performance of both real-valued and complex-valued models on GPUs. To ensure comparability of FLOPs between the two approaches, the impact of weight coefficients was removed. The experiment utilized a range of parameters from Table 1, focusing on evaluating how the dimensions of word embeddings affect parallelization efficiency. During the 1000 iterations test for both models, we meticulously recorded the computational time required for each iteration. As shown in Table 2, our study compares the time consumption of traditional (non-parallel) methods with that of QISR (parallel) methods in processing word embeddings across various dimensions. The research results indicate that the parallel processing method using QISR significantly reduces the calculation time compared to traditional methods, particularly when processing high-dimensional data. This observation suggests

| Setting | Value |
|---|---|
| lr | 0.005 |
| epoch | 50 |
| batch size | 64 |
| measurements | 20 |
| length | 64 |
| slide | 16 |
| seed | 0 |
| run times | Take the maximum of 6 |
| cpu | i5-10505 |
| gpu | V100 16G |

Table 1: Model hyperparameters and device model.

| Dim | Real | Complex |
|---|---|---|
| 100 | 1.998 | 8.542 |
| 100-QISR | 0.035 | 0.141 |
| 200 | 6.636 | 36.054 |
| 200-QISR | 0.035 | 0.261 |
| 500 | 24.778 | 198.201 |
| 500-QISR | 0.036 | 2.698 |

Table 2: Computational time efficiency (in milliseconds (ms)) of complex versus real quantum-inspired models with and without QISR optimization is compared. It is evident that the parallelization performance benchmarks for models utilizing the QISR method significantly outperform those of the conventional approach.

that quantum-inspired models utilizing QISR can process high-dimensional word embedding tasks more quickly, thus improving overall efficiency.

## 4 Experiments

### 4.1 Experimental Design and Evaluation Metrics

This study aims to optimize density matrix-based quantum-inspired models to reduce computational and memory overhead. The experiment was conducted in two stages. In the first stage, by applying the QISR method with various word embedding dimensionalities, we use the running time of CPU and GPU and the memory consumption of GPU as indicators to evaluate the processing time and memory overhead. Subsequent, in the second stage, we first analyze the differences brought about by the QISR method, and then compare the accuracy performance of the model before and after applying QISR on the real task data set to test whether the QISR optimization may have a negative impact on the model performance. In conclusion, our experiments not only test whether QISR optimization can significantly reduce the time and space complexity of the model, but also whether QISR optimization can at least not reduce the performance of the model.

### 4.2 Datasets

The experiments of this study mainly focus on text classification tasks. Text classification was chosen because it is relatively simple and can clearly demonstrate how to improve computational efficiency without sacrificing performance metrics. The classification dataset used in the experiment is as follows.

**The Stanford Sentiment Treebank**(Socher et al., 2013). SST is released by Stanford University and is mainly used for sentiment classification of movie reviews. The dataset is divided into two parts: SST-2 (Binary Classification) and SST-5 (Five-Level Classification). SST-2 contains 11,855 movie reviews, divided into 8,544 training samples, 1,101 development samples, and 2,210 test samples; SST-5 contains 6,920 training samples, 872 development samples, and 1,821 test samples.

**Movie Reviews**(Pang and Lee, 2005). MR is a dataset designed for sentiment analysis experiments, comprising annotated movie review documents with overall sentiment orientation and subjective states, along with the sentiment orientation and subjective states of individual sentences.

**The Corpus of Linguistic Acceptability**(Warstadt et al., 2018). CoLA is a dataset comprises 10,657 sentences from 23 linguistics publications, annotated for acceptability by original authors. The public release has 9,594 sentences in training and development sets, excluding 1,063 for a held-out test set.

It's worth noting that while quantum-inspired models are suitable for various tasks, the focus of this experiment is on demonstrating their application in text classification, which may not be advantageous for some models like CNM.

### 4.3 Baselines and Parameter Scale

We used well-known quantum-inspired models based on density matrices, including NNQLM-I, NNQLM-II, QPDN, and CNM. Our experimental parameters, as shown in Table 1, were initialized with 50-dimensional, 200-dimensional, and 300-dimensional GloVe vectors. Subsequently, in the QISR model, we employed 768-dimensional word embeddings to simulate the computational cost of

| Model | | Ori | | | Ori+QISR | | |
|---|---|---|---|---|---|---|---|
| | | c-t(ms) | g(MB) | g-t(ms) | c-t(ms) | g(MB) | g-t(ms) |
| NNQLM-I | 50 | 11.12 | 194 | 0.10 | **0.35** | **154** | **0.10** |
| | 200 | 205 | 798 | 5.30 | **4.00** | **172** | **0.10** |
| | 300 | 381 | 1602 | 11.86 | **7.34** | **194** | **0.11** |
| | 768 | - | - | - | **34.88** | **338** | **0.23** |
| NNQLM-II | 50 | 23.13 | 402 | 0.30 | **14.03** | **390** | **0.24** |
| | 200 | 335 | 1006 | 17.19 | **199** | **772** | **9.03** |
| | 300 | 816 | 1808 | 69.40 | **510** | **1280** | **29.53** |
| | 768 | - | - | - | **3506** | **12456** | **387** |
| QPDN | 50 | 76.82 | 428 | 4.65 | **5.28** | **192** | **1.04** |
| | 200 | 1306 | 4612 | 52.22 | **38.93** | **330** | **1.05** |
| | 300 | 3158 | 10236 | 130 | **76.82** | **468** | **1.25** |
| | 768 | - | - | - | **419** | **1736** | **12.92** |
| CNM | 50 | 731 | 2606 | 30.91 | **58.87** | **406** | **2.35** |
| | 200 | 1152 | OOM | - | **714** | **3652** | **21.78** |
| | 300 | 2861 | OOM | - | **1804** | **8104** | **51.32** |
| | 768 | - | - | - | **9121** | **OOM** | - |

Table 3: "ori" represents the classification form of the original model. "ori-QISR" represents the classification model of ori using QISR. “c-t” and “g-t” represent the time required to conduct the experiment on CPU and GPU respectively, while “g” represents the memory consumption of GPU. The "-" indicates that no experimental statistics are performed because the memory is OOM.

using BERT(Devlin et al., 2018) as word embeddings.

### 4.4 Experiment Results

#### 4.4.1 Performance Evaluation

Table 3 shows the experimental results under four different word embedding dimensions. In particular, to simulate BERT, we conduct experiments using 768-dimensional word embeddings alone in the QISR method. According to the analysis in Section 3, the QISR method can reduce memory overhead (n times) in the density matrix construction stage. However, in addition to the density matrix layer, several other modules are included in the quantum inspired model. Therefore, the effect of this memory overhead reduction will vary in different models. For example, in the NNQLM-I model, when the word vector dimension is 200, the QISR method can be used to reduce parameters by 4 times. When the dimension is increased to 300, the reduction can be up to 8 times. However, since the NNQLM-II model uses convolution in its module, its QISR acceleration effect is not as obvious as the NNQLM-I model.

Overall, the QISR method shows excellent performance in quantum-inspired models. Not only does it effectively reduce the computational cost, but it also enables the QISR model to increase the dimension of word embedding, thereby surpassing the limitations of the traditional bag-of-word model and laying the foundation for future integration with pre-trained models. As shown in Table 3, even using 768-dimensional word embedding, our QISR still only requires about 1736MB of memory on QPDN, which can maximize the potential of quantum-inspired models on limited resources.

Since some models cannot run properly when the word embedding is 300 dimensions, we only conduct experiments on 50-dimensional and 200-dimensional word embeddings in the following experiments.

| Data Seed | MAE | MSE |
|---|---|---|
| 10 | 1.3956e-06 | 3.6820e-12 |
| 100 | 1.4382e-06 | 3.9215e-12 |
| 200 | 1.6851e-06 | 5.4577e-12 |

Table 4: In this table, our input parameters (n, d) are (64, 50). "Data Seed" represents the random numbers generated using different seed values.

#### 4.4.2 Accuracy Loss Analysis

To compare the accuracy loss, this study initially explores the differences caused by two methods. Specifically, we selected the density matrix layers shown in Figures 1-b and 1-d and conducted comparative experiments using different random seeds.

| Model | Dim | Task Acc. | | | Task Mcc |
|---|---|---|---|---|---|
| | | SST2 | SST5 | MR | CoLA |
| NNQLM-I | 50 | 71.29 | 35.65 | 66.53 | 4.53 |
| | 200 | 73.97 | 37.19 | 67.53 | 0 |
| NNQLM-I+QISR | 50 | 72.93 | 36.29 | 65.73 | 4.8 |
| | 200 | 73.86 | 37.71 | 68.63 | 0 |
| NNQLM-II | 50 | 72.69 | 35.97 | 74.13 | 9.51 |
| | 200 | 71.31 | 36.28 | 72.59 | 8.17 |
| NNQLM-II+QISR | 50 | 72.16 | 37.54 | 74.74 | 10.59 |
| | 200 | 71.49 | 36.25 | 73.09 | 9.80 |
| QPDN | 50 | 83.96 | 43.36 | 81.99 | 14.08 |
| | 200 | 83.80 | 43.97 | 81.49 | 15.72 |
| QPDN+QISR | 50 | 84.24 | 43.57 | 82.09 | 14.72 |
| | 200 | 83.96 | 44.66 | 81.79 | 16.37 |
| CNM | 50 | 78.14 | 38.77 | 76.13 | 5.97 |
| | 200 | 77.75 | 37.78 | 74.74 | 0.89 |
| CNM+QISR | 50 | 78.13 | 39.23 | 76.19 | 8.16 |
| | 200 | 77.38 | 37.96 | 75.03 | 0.92 |
| TextCNN | 50 | 78.86 | 40.90 | 79.39 | 16.37 |
| | 200 | 82.53 | 43.57 | 82.59 | 21.16 |

Table 5: Accuracy comparison of QISR and non-QISR models using four different models and different dimensions. And added TextCNN(Kim, 2014) under the same hyperparameters.

In Section 3.3, we theoretically demonstrated the equivalence of these two methods. However, during practical computation, we observed that the summation operation in Formula 5 leads to more significant precision loss compared to Formula 7. Specifically, in the experiments, it was noted that the mantissa part of the model was set to zero, resulting in a decrease in precision, with the mean absolute error (MAE) reaching the level of $10^{-6}$, as detailed in Table 4. Our proposed QISR method exhibited higher precision in constructing mixed states, thoroughly demonstrating the overall effectiveness of our model.

Subsequently, we investigated whether applying the QISR method would adversely affect the overall model, as detailed in Table 5. When processing different tasks, the model performance showed minor variations. Specifically, for the relatively simple task SST-2, the model accuracy seemed unaffected by the use of different methods. However, for the more complex SST-5 and CoLA tasks, the QISR method performed slightly better than the non-QISR method, indicating a certain degree of performance enhancement. This result confirms the performance improvement of the QISR method in handling challenging tasks and its advantages in evaluation metrics.

## 5 Discussions

In this study, we propose a quantum-inspired sentence representation model (QISR) that shows significant effectiveness in terms of processing time and memory overhead. Provides an efficient solution to the limitations of traditional quantum-inspired models in dealing with high dimensions. Furthermore, this study not only demonstrates the application potential of quantum-inspired in the field of natural language processing, but also provides new possibilities for efficient language processing in resource-constrained environments. Despite these achievements, results as shown in Table 5 show that current quantum-inspired models, including our proposed QISR, exhibit certain limitations in specific tasks.

Therefore, the first goal of future work is to address these identified weaknesses and improve the overall performance of the model in various dimensions. We expect this study to inspire more future research and applications of quantum-inspired models in the field of natural language processing. The next work will focus on further improving the accuracy and optimization performance of the QISR model to enhance its adaptability and interpretability in various natural language processing tasks.

## References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimitrios Gkoumas, Qiuchi Li, Yijun Yu, and Dawei Song. 2021a. An entanglement-driven fusion neural network for video sentiment analysis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1736–1742. International Joint Conferences on Artificial Intelligence Organization.

Dimitris Gkoumas, Qiuchi Li, Shahram Dehdashti, Massimo Melucci, Yijun Yu, and Dawei Song. 2021b. Quantum cognitively motivated decision fusion for video sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 827–835.

Yongyu Jiang, Peng Zhang, Hui Gao, and Dawei Song. 2020. A quantum interference inspired neural matching model for ad-hoc retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–28.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. 2021. Quantum-inspired neural network for conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13270–13278.

Qiuchi Li, Benyou Wang, and Massimo Melucci. 2019. Cnm: An interpretable complex-valued network for matching. *arXiv preprint arXiv:1904.05298*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. 2021. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 871–880.

Yaochen Liu, Yazhou Zhang, and Dawei Song. 2023. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. *IEEE Transactions on Affective Computing*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. 2013. Modeling term dependencies with quantum language models for IR. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 653–662.

Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. 2023. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3926–3933.

Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. 2019. Semantic hilbert space for text representation learning. In *The World Wide Web Conference*, pages 3293–3299.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint 1805.12471*.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. 2018a. End-to-end quantum-like language models with application to question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. 2018b. A quantum many-body wave function inspired language modeling approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1303–1312.

Peng Zhang, Jing Zhang, Xindian Ma, Siwei Rao, Guangjian Tian, and Jun Wang. 2020a. TextTN: Probabilistic encoding of language on tensor network.

Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. 2019. Quantum-inspired interactive networks for conversational sentiment analysis.

Yazhou Zhang, Dawei Song, Xiang Li, Peng Zhang, Panpan Wang, Lu Rong, Guangliang Yu, and Bo Wang. 2020b. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, 62:14–31.