QBET: Quantum Based Enhanced Transformer for Efficient Language Modeling

Abstract

The large-scale deployment of Generative Artificial Intelligence (AI), predominantly propelled by Large Language Models (LLMs), and simultaneous rapid progress in fault-tolerant quantum computers have motivated research into integrating quantum algorithms with transformer architectures. We introduce QBET (Quantum-Based Enhanced Transformer), a novel architecture that combines trainable position-aware quantum mixing, sparse attention mechanisms, quantum-enhanced attention selectively applied to important tokens, and Mixture-of-Experts feed-forward networks. In experiments conducted on the two datasets Penn Treebank (PTB) and WikiText-2, QBET achieved a perplexity (PPL) score of 93.3 and 212.7 (respectively) in only five and six epochs, respectively, surpassing both classical and Quantum baseline models. This research significantly advances the field of hybrid quantum-AI architectures by demonstrating how quantum properties can enhance the evolution of AI architectures. In conclusion, with ongoing research in quantum-based LLMs and advancements in quantum hardware, QuantumAI systems are rapidly becoming attainable.