

## 标题：基于轻量化 Transformer 的文本检索模型研究

### 摘要：

本文提出了一种基于轻量级 Transformer 的文本检索模型，该模型通过窗口化注意力机制与向量量化技术，使得在消费级显卡上也能完成高效的语义检索任务。实验结果显示，在减少 40% 参数量的情况下，检索精度仅下降 1.7%。

**关键词：** 文本检索、轻量化模型、向量量化、Transformer

## 1. 引言

近年来，语义检索技术发展迅速，但大部分模型仍然需要大量算力。为了让中小企业也能部署本地的智能检索系统，轻量化模型成为趋势。

## 2. 模型结构

模型采用两部分：

- 窗口化多头注意力 W-MHA
- 向量量化嵌入层 VQ-Embedding

公式如下：

$$Q = XW_{Q_r}$$

$$K = XW_{K_r}$$

$$V = XW_V$$

## 3. 实验结果

在 MTEB 中文检索基准中，本模型达到了 **82.1** 的平均得分。

## 4. 结论

该模型适用于本地 RAG 系统中的快速查询场景。