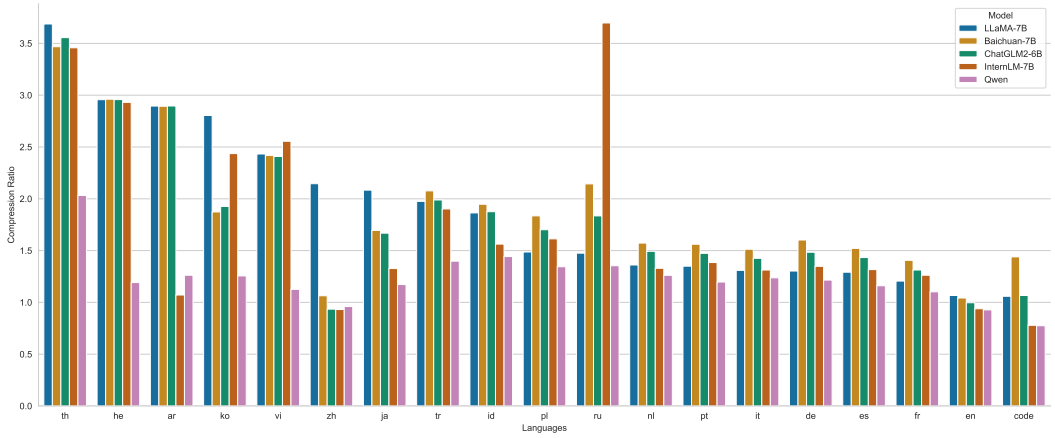

Figure 3: **Encoding compression rates of different models.** We randomly selected 1 million document corpora of each language to test and compare the encoding compression rates of different models (with XLM-R (Conneau et al., 2019), which supports 100 languages, as the base value 1, not shown in the figure). As can be seen, while ensuring the efficient decoding of Chinese, English, and code, QWEN also achieves a high compression rate for many other languages (such as th, he, ar, ko, vi, ja, tr, id, pl, ru, nl, pt, it, de, es, fr, etc.), equipping the model with strong scalability as well as high training and inference efficiency in these languages.

2.2 TOKENIZATION

The design of vocabulary significantly impacts the training efficiency and the downstream task performance. In this study, we utilize byte pair encoding (BPE) as our tokenization method, following GPT-3.5 and GPT-4. We start with the open-source fast BPE tokenizer, tiktoken (Jain, 2022), and select the vocabulary cl100k base as our starting point. To enhance the performance of our model on multilingual downstream tasks, particularly in Chinese, we augment the vocabulary with commonly used Chinese characters and words, as well as those in other languages. Also, following Touvron et al. (2023a;b), we have split numbers into single digits. The final vocabulary size is approximately 152K.

The performance of the QWEN tokenizer in terms of compression is depicted in Figure 3. In this comparison, we have evaluated QWEN against several other tokenizers, including XLM-R (Conneau et al., 2019), LLaMA Touvron et al. (2023a), Baichuan Inc. (2023a), and InternLM InternLM Team (2023). Our findings reveal that QWEN achieves higher compression efficiency than its competitors in most languages. This implies that the cost of serving can be significantly reduced since a smaller number of tokens from QWEN can convey more information than its competitors. Furthermore, we have conducted preliminary experiments to ensure that scaling the vocabulary size of QWEN does not negatively impact the downstream performance of the pretrained model. Despite the increase in vocabulary size, our experiments have shown that QWEN maintains its performance levels in downstream evaluation.

2.3 MODEL

2.3.1 ARCHITECTURE

QWEN is designed using a modified version of the Transformer architecture. Specifically, we have adopted the recent open-source approach of training large language models, LLaMA (Touvron et al., 2023a), which is widely regarded as the top open-source LLM. Our modifications to the architecture include:

- **Embedding and output projection**. Based on preliminary experimental findings, we have opted for the untied embedding approach instead of tying the weights of input embedding and output projection. This decision was made in order to achieve better performance with the price of memory costs.