

1 A Key Dimensionality Analysis with RoPE

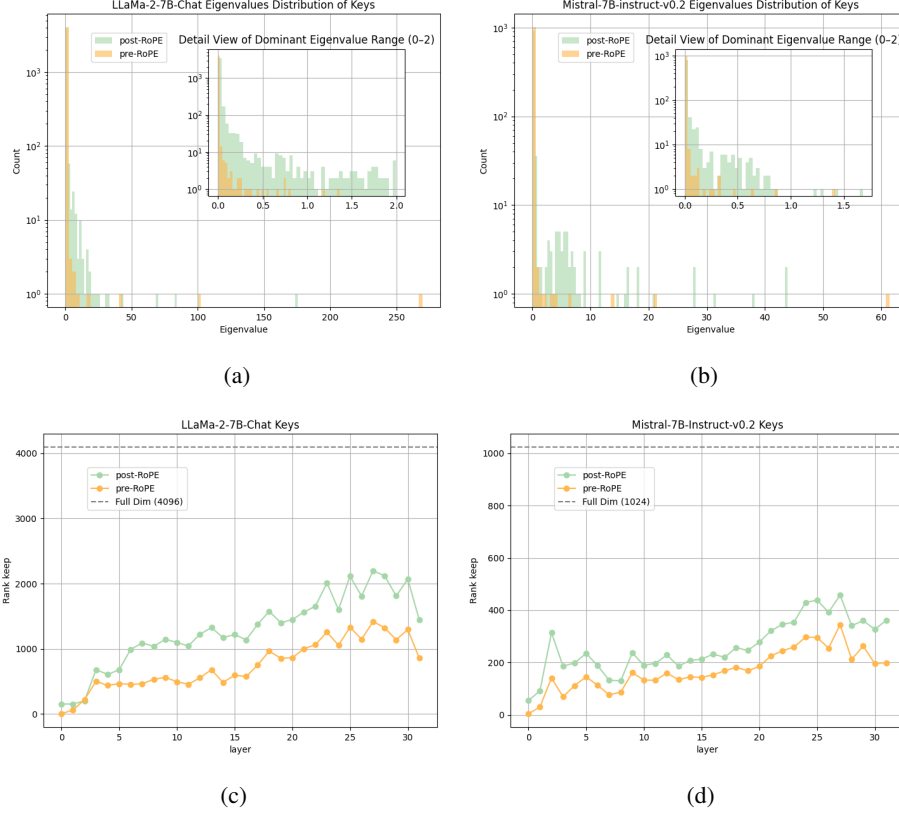


Figure 1: (a)–(b): Eigenvalue distributions of key covariance matrices in LLaMA-2-7B-Chat and Mistral-7B-Instruct-v0.2 before and after applying Rotary Position Embedding (RoPE). (c)–(d): Number of principal components required to retain 90% of the total variance across transformer layers, indicating changes in effective rank after RoPE.

In this section, we conduct a numerical analysis to quantify how rotary position embedding (RoPE) alters the principal-component structure of the key states. We perform principal component analysis (PCA) on the key tensors before and after applying RoPE, denoted as *pre-RoPE* and *post-RoPE*, respectively. Concretely, we first compute the covariance matrix of the key states and then carry out an eigenvalue decomposition. The magnitude of each eigenvalue reflects the contribution of its associated eigenvector: the presence of many small eigenvalues indicates that the corresponding representation is effectively low rank. To relate rank to the spectrum more systematically, we adopt the metrix introduced in Loki[1]:

$$\text{Rank}_l(v) = \min \left\{ d \in \mathbb{Z}_+ : \sum_{i=1}^d \lambda_l^{(i)} \geq \frac{v}{100} \right\},$$

where $\lambda_l^{(i)}$ denotes the i -th largest eigenvalue of the covariance matrix for the keys in layer l . This definition specifies the smallest number of principal components needed to explain at least $v\%$ of the total variance, enabling a direct comparison of the intrinsic dimensionality before and after RoPE.

Figure 1(a)–(b) illustrate the eigen-value spectra of the first layer ($l = 0$) for Llama-2-7B-Chat and Mistral-7B-Instruct-v0.2 under the pre-RoPE and post-RoPE conditions. The pre-RoPE spectra exhibit a markedly larger number of small eigenvalues, whereas the post-RoPE spectra are shifted upward, corroborating our earlier finding that RoPE increases the overall variance. Figure 1(c)–(d) present the layer-wise $\text{Rank}_l(90)$ values, i.e. the minimal dimensionality needed to retain 90% of the variance. For both models, the post-RoPE condition consistently requires a higher rank, reflecting the broader spectra observed in panels (a)–(b). In addition, the required rank varies substantially across layers, indicating that a layer-adaptive rank selection scheme could further enhance compression efficiency.

22 **References**

- 23 [1] Prajwal Singhanian, Siddharth Singh, Shwai He, Soheil Feizi, and Abhinav Bhatele. Loki: Low-
24 rank keys for efficient sparse attention. *arXiv preprint arXiv:2406.02542*, 2024.