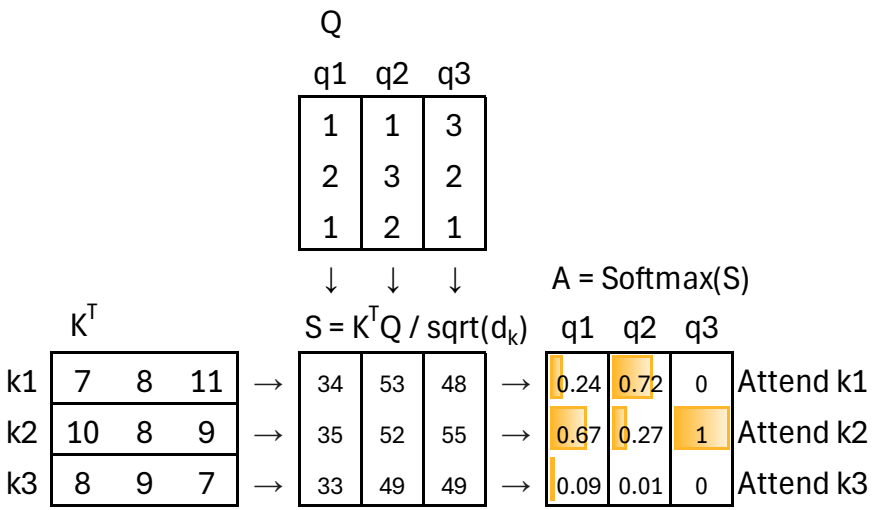
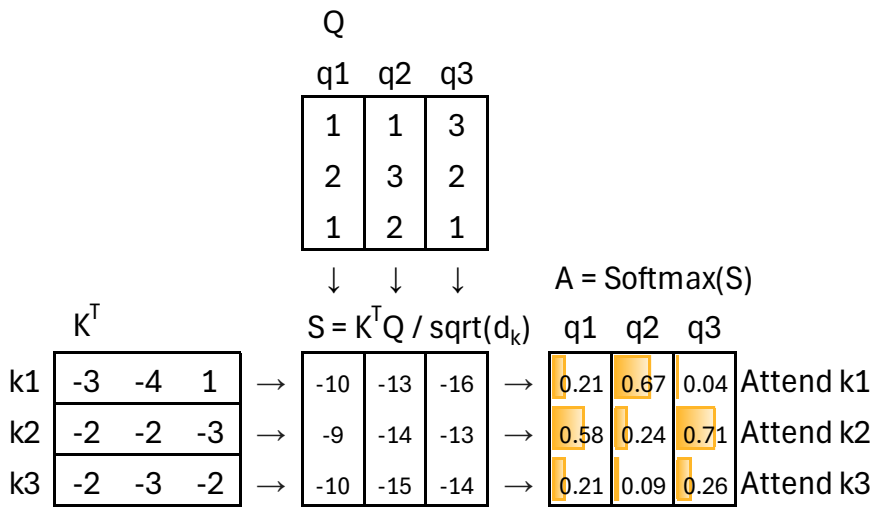
Ignore: Attend to None

Traditional QKV Attention

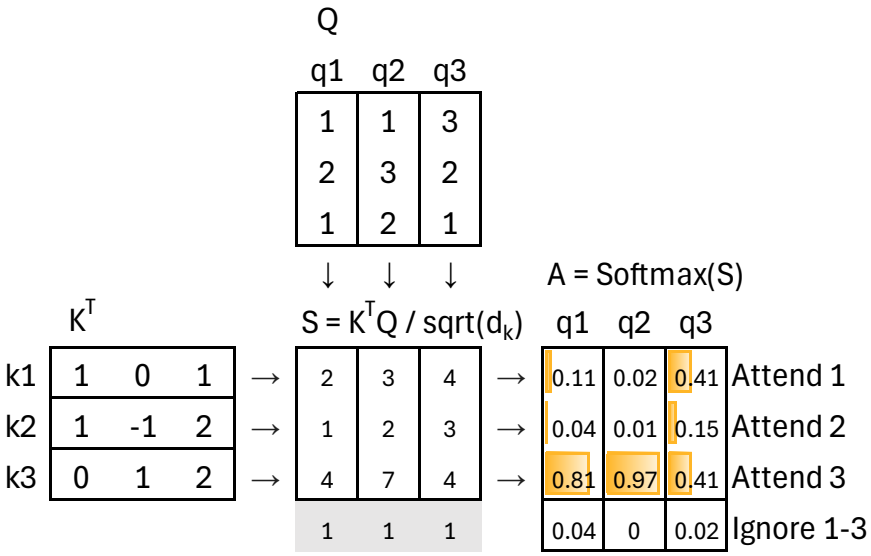
Traditional QKV attention can't ignore tokens—softmax forces the model to assign some attention, even when all similarity scores are low or irrelevant.

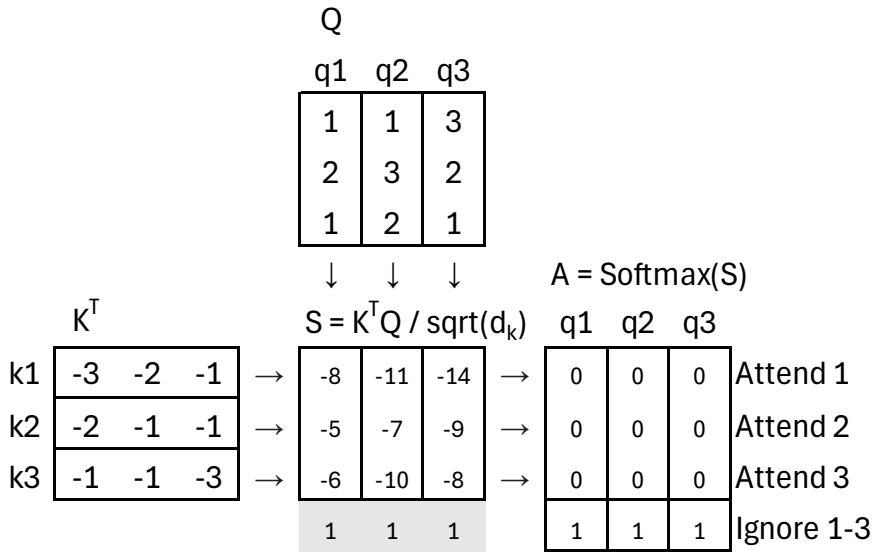




Off-by-One Softmax

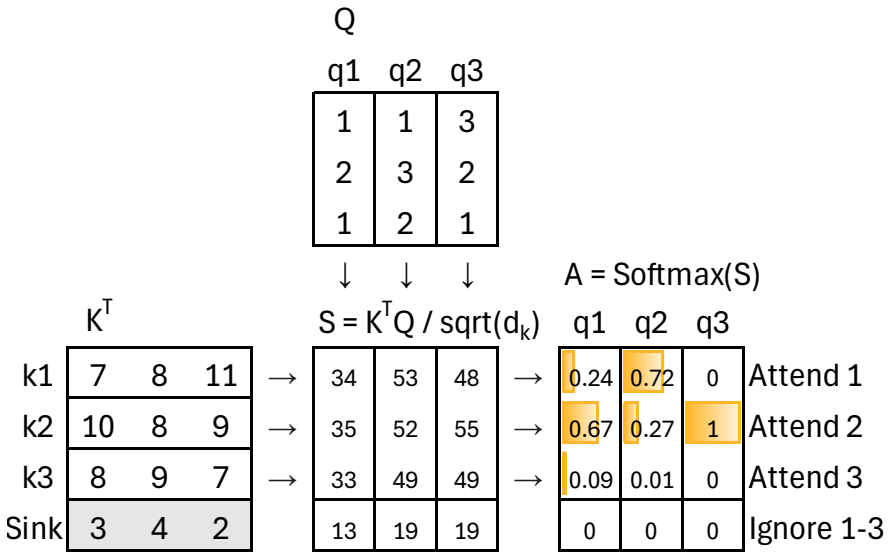
Off-by-one attention adds a fixed bias to the softmax, so tokens with similarity scores significantly smaller than the bias receive near-zero attention—letting the model ignore weak matches. Similarity scores higher than the bias work the same way as in standard attention.

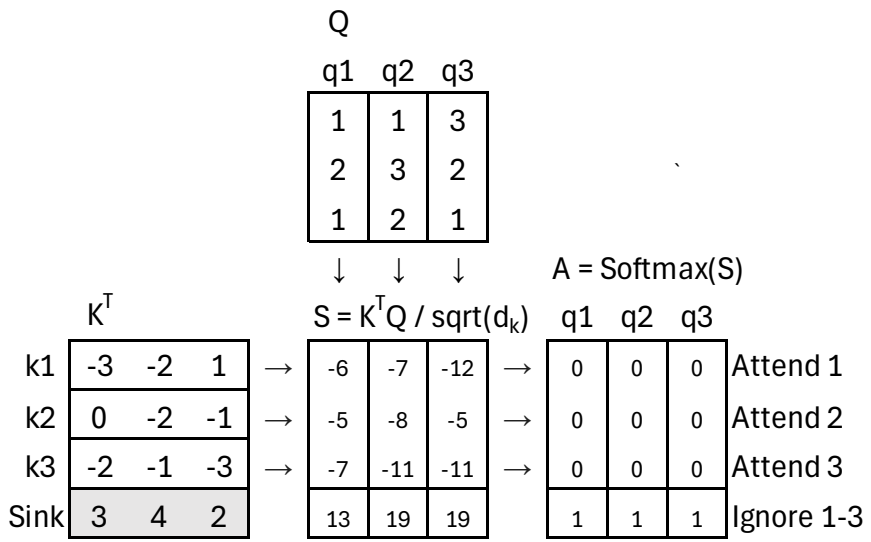




Sink Token

Sink token attention introduces a special sink that is always present. When a query has no strong matches among the other keys, attention shifts toward the sink, allowing the model to effectively ignore all the keys. Keys with high similarity still receive attention as usual, just like in standard attention.





GPT-OSS: Learnable Softmax Bias

This method adds a learnable bias to the softmax, allowing thattention head to dynamically control how easily it ignores low-similarity keys. When similarity scores are significantly smaller than the learned bias, the resulting attention weights become near zero—enabling the model to suppress weak matches. Higher similarity scores still behave like in standard attention.