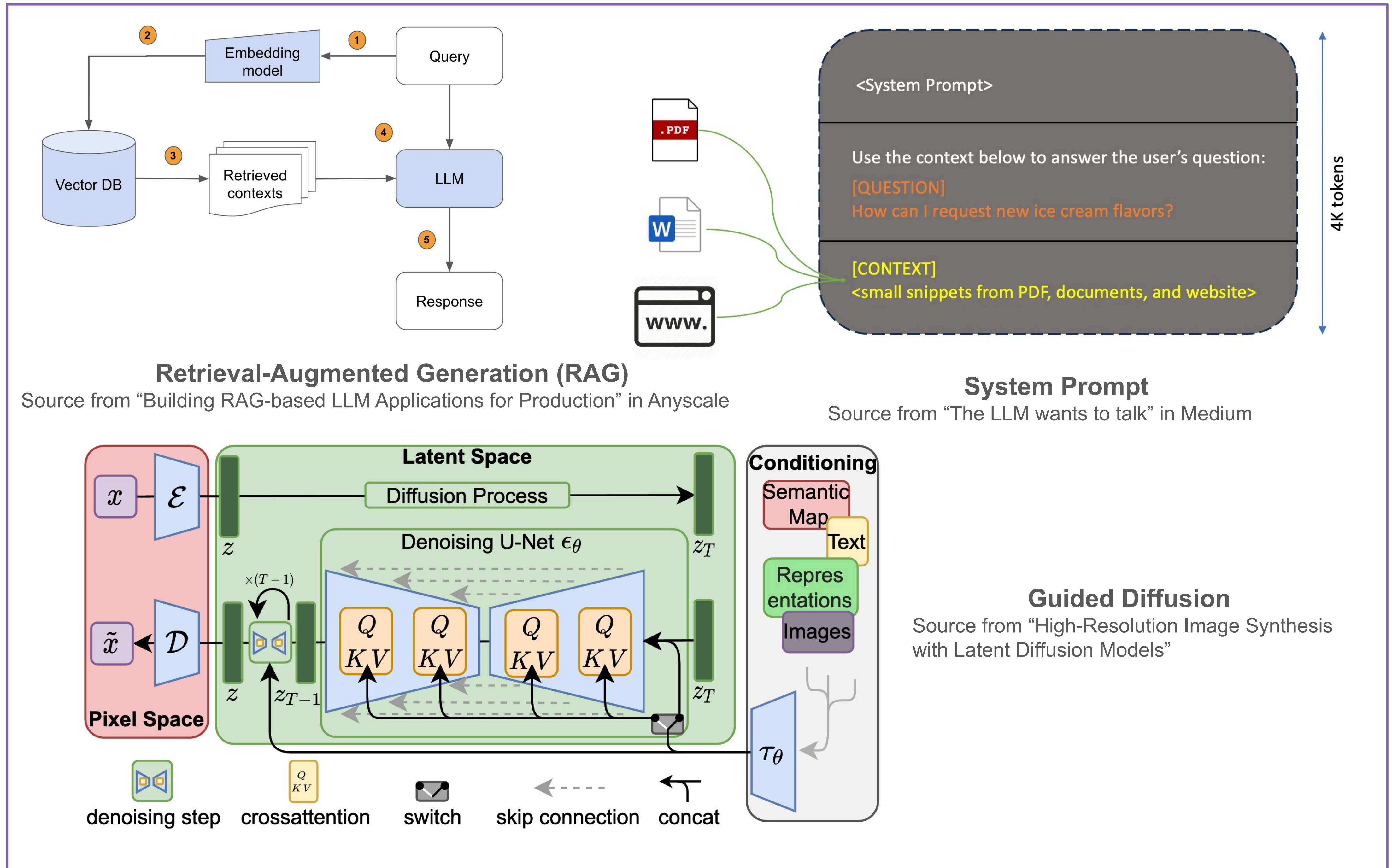


# Differential Privacy of Cross-Attention with Provable Guarantee

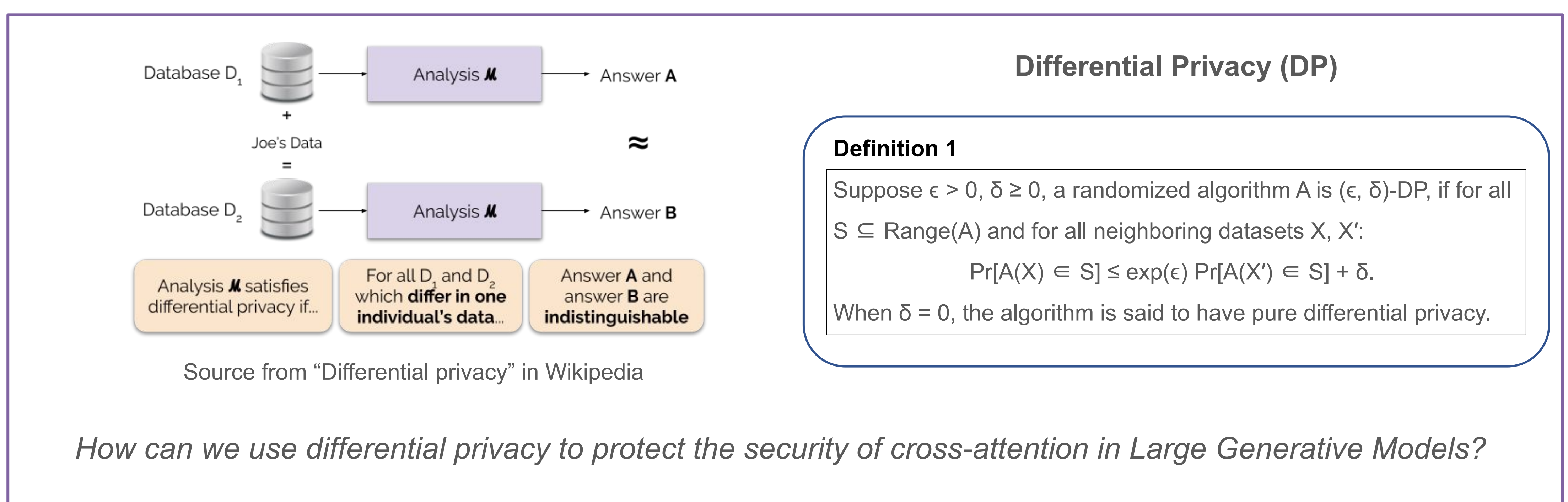
Yingyu Liang, Zhenmei Shi, Zhao Song, Yufa Zhou



## Background



## Motivation



## Main Results

### Theorem 1 (DP Cross-Attention)

Consider the cross-attention mechanism where queries, keys, and values are defined as in the standard framework. There exists an algorithm that uses polynomial kernel approximations to efficiently compute cross-attention by treating it as a weighted distance problem between query and key embeddings, weighted by the value embeddings. With high probability, the algorithm ensures differential privacy for the cross-attention process, is robust to repeated adaptive queries, and achieves small relative and additive errors. Furthermore, the additive error decreases as the number of input tokens grows, eventually diminishing to zero.