Differential Privacy of Cross-Attention with Provable Guarantee

Yingyu Liang, Zhenmei Shi, Zhao Song, Yufa Zhou

香 港 大 學
THE UNIVERSITY OF HONG KONG

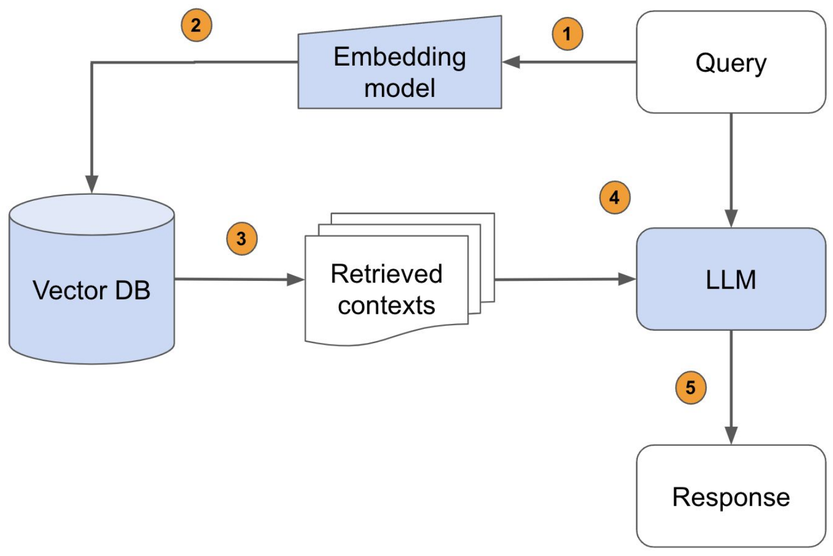
WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Berkeley
UNIVERSITY OF CALIFORNIA

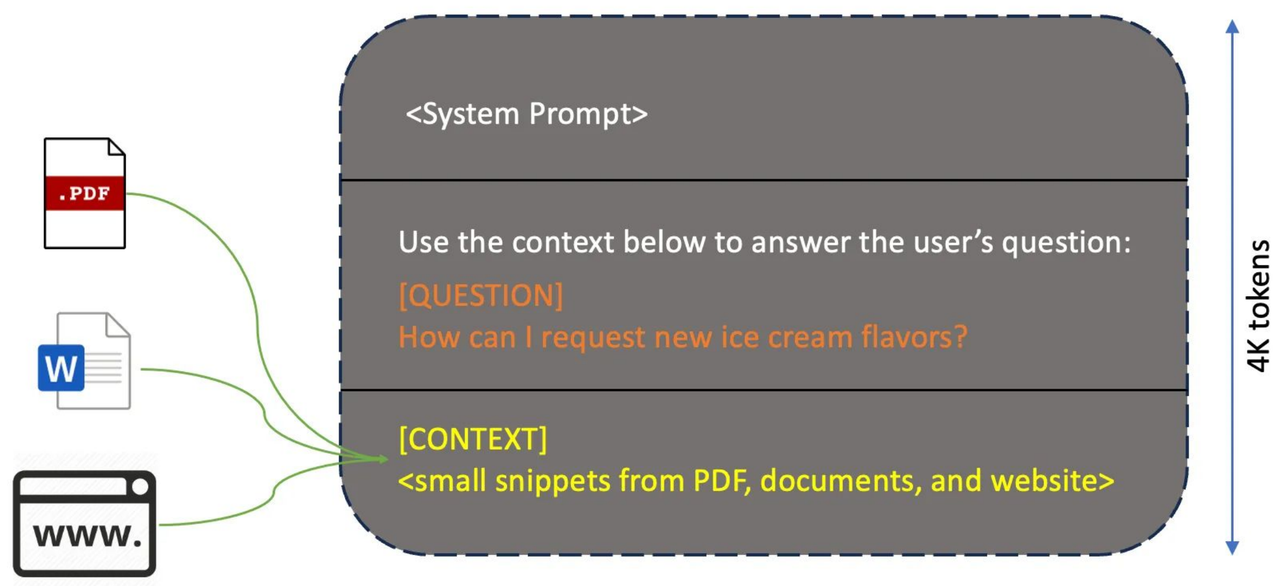
Penn
University of Pennsylvania

SCAN ME

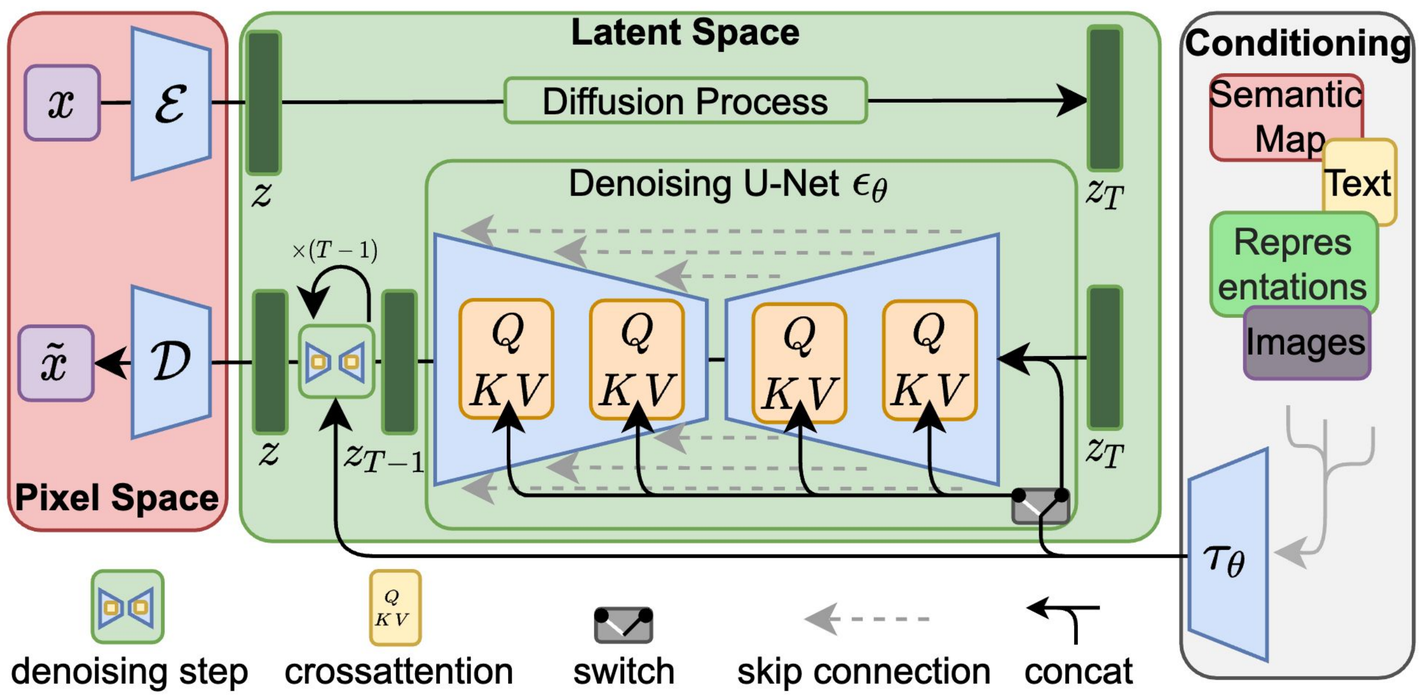
Background



Retrieval-Augmented Generation (RAG)
Source from “Building RAG-based LLM Applications for Production” in Anyscale

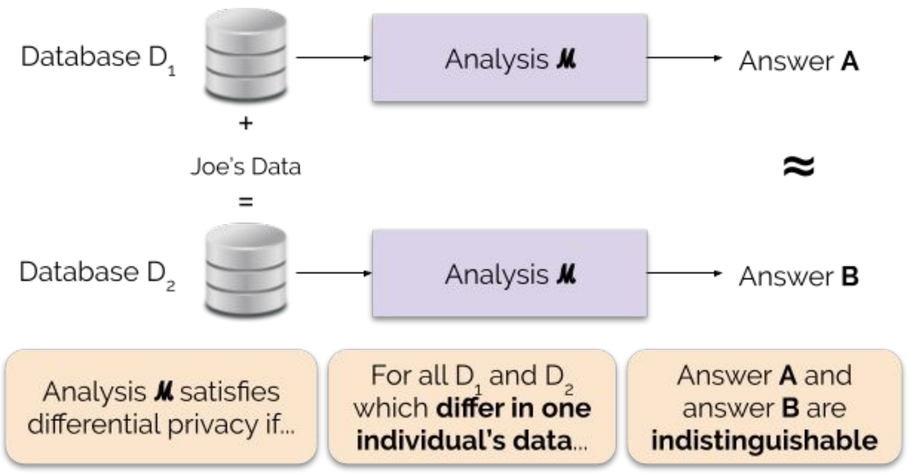


System Prompt
Source from “The LLM wants to talk” in Medium



Guided Diffusion
Source from “High-Resolution Image Synthesis with Latent Diffusion Models”

Motivation



Source from “Differential privacy” in Wikipedia

Differential Privacy (DP)

Definition 1

Suppose $\epsilon > 0$, $\delta \geq 0$, a randomized algorithm A is (ϵ, δ)-DP, if for all $S \subseteq \text{Range}(A)$ and for all neighboring datasets X, X′:

$$\Pr[A(X) \in S] \leq \exp(\epsilon) \Pr[A(X') \in S] + \delta.$$

When $\delta = 0$, the algorithm is said to have pure differential privacy.

How can we use differential privacy to protect the security of cross-attention in Large Generative Models?

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.