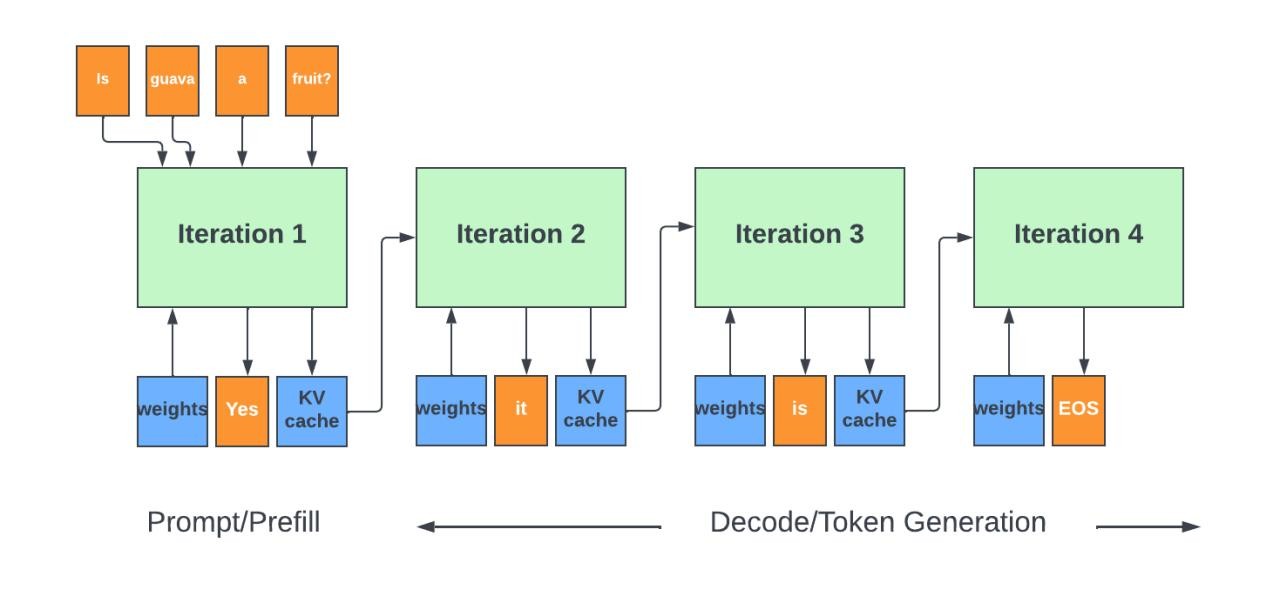
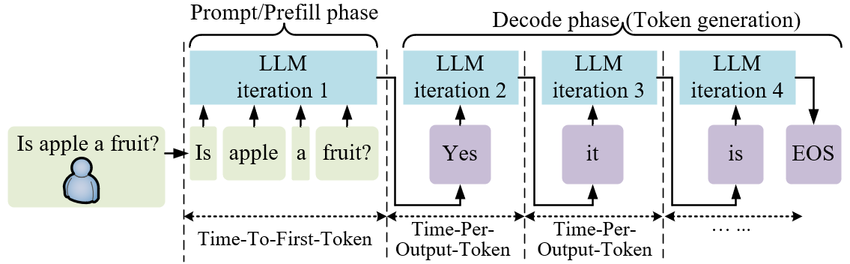
**Inference** - <https://www.linkedin.com/pulse/llm-inference-hwsw-optimizations-sharada-yeluri-wfdyc/>





**Embeddings**: <https://www.youtube.com/watch?v=8kJStTRuMcs>

By default, the length of the embedding vector will be 1536 (output dimensions) for text-embedding-3-small or 3072 for text-embedding-3-large

**Vector database**: <https://youtu.be/dN0lsF2cvm4?si=UtvkXY5g-oW_487V&t=43>