Core Architecture of DeepSeek R1 and V3

Unveiling Mixture-of-Experts and Multi-head Latent Attention

DeepSeek V3 and R1: Architectural Overview

At their core, DeepSeek V3 and R1 leverage a shared Mixture-of-Experts (MoE) foundation for efficient, scalable AI. This architecture allows for specialized processing while maintaining a vast parameter count.

Shared Foundation

- Both models use a **Mixture-of-Experts (MoE)** architecture.

DeepSeek V3

- **671 Billion** parameters total.
- **37 Billion** parameters active per token.
- Optimized for **general-purpose tasks**.
- **64 Experts**, **6 Activated** for a token with **2 shared experts.**

DeepSeek R1

- Built upon **V3-Base**.
- Focused on **complex reasoning** and Chain-of-Thought (CoT) capabilities.

Key Components

- MoE with **sparse activation**.
- **Multi-head Latent Attention (MLA)**.
- **Rotary Positional Embedding (RoPE)**.

Deepseek R1

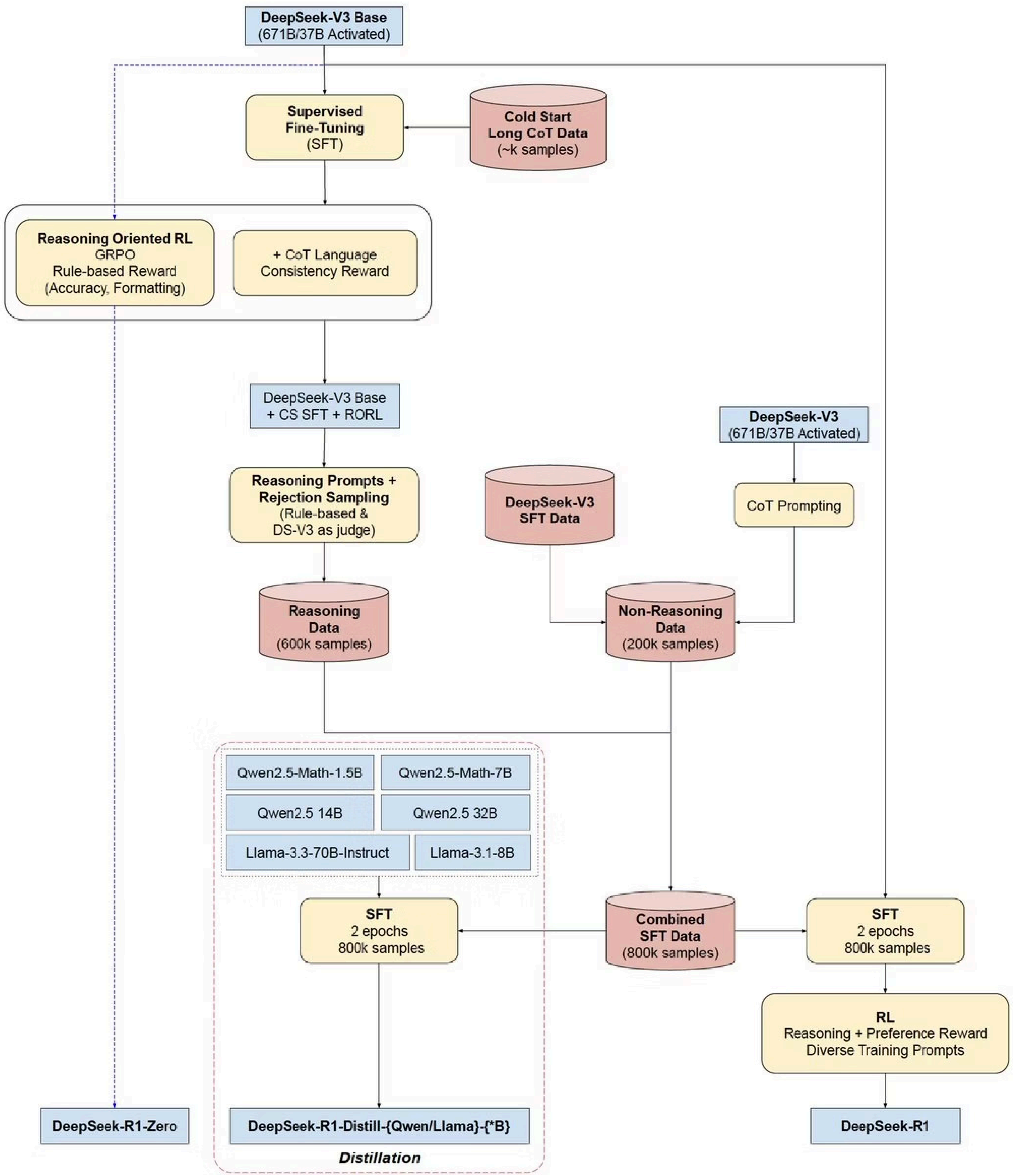

Image Link

Deepseek V3 model

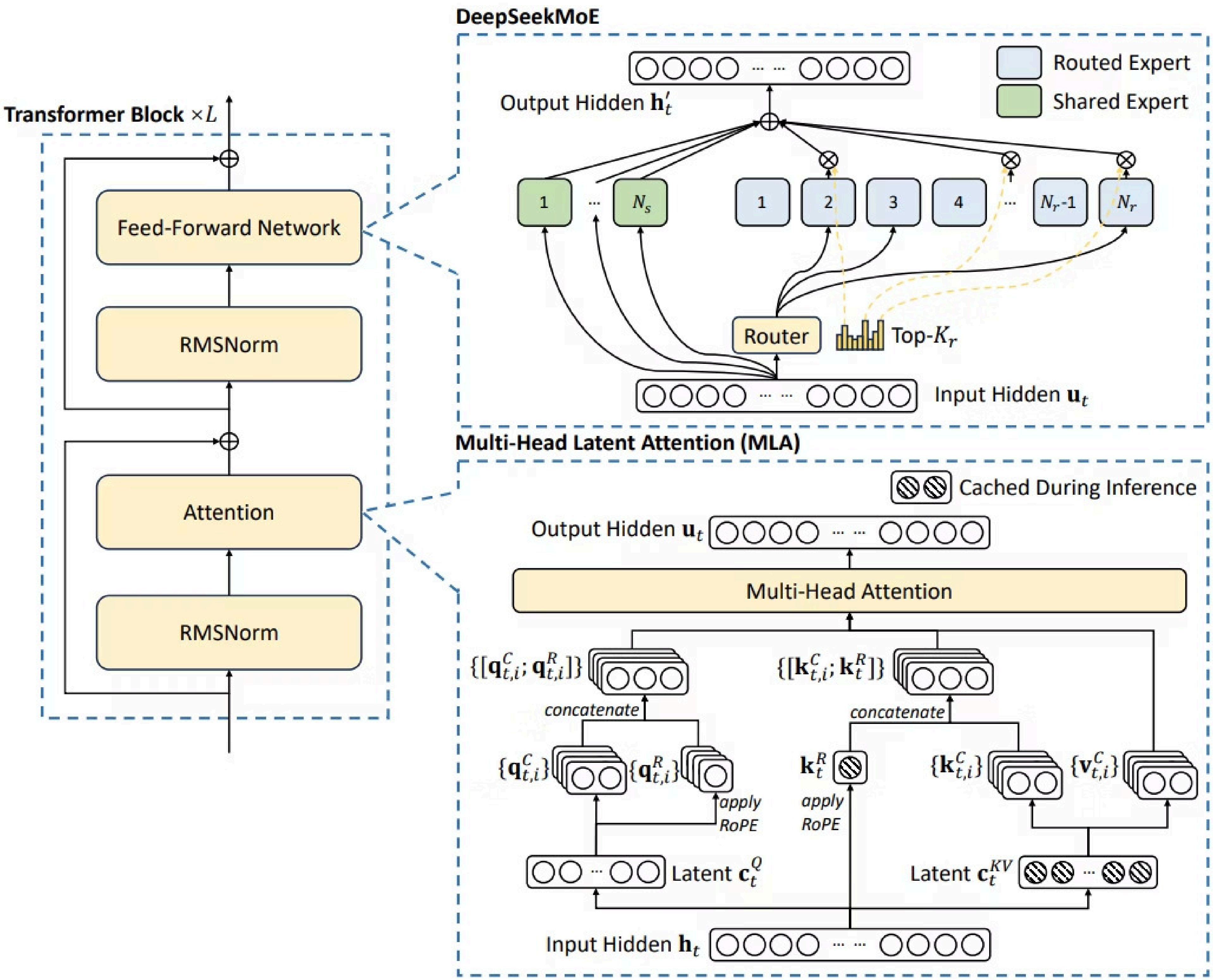

Image Link

Multi-head Latent Attention (MLA)

MLA is a cornerstone of DeepSeek's efficiency, processing attention with compressed representations to conserve memory and compute.

Efficient Attention

Query: It says that what this token wants next.
Key: It says that what the current tokens can offer.
Value: It says that what the current tokens are offering(token-values).

- Compresses t
- he input from **2048D to 128D** using linear layers.
- Splits into **16 heads**, each handling **128D content** and **64D positional data**.
- Concatenates content and positional embeddings before the attention mechanism.

Benefits

- Significantly **reduces memory footprint**.
- Lowers **computational cost**.
- Maintains **rich contextual awareness** through compressed representations.

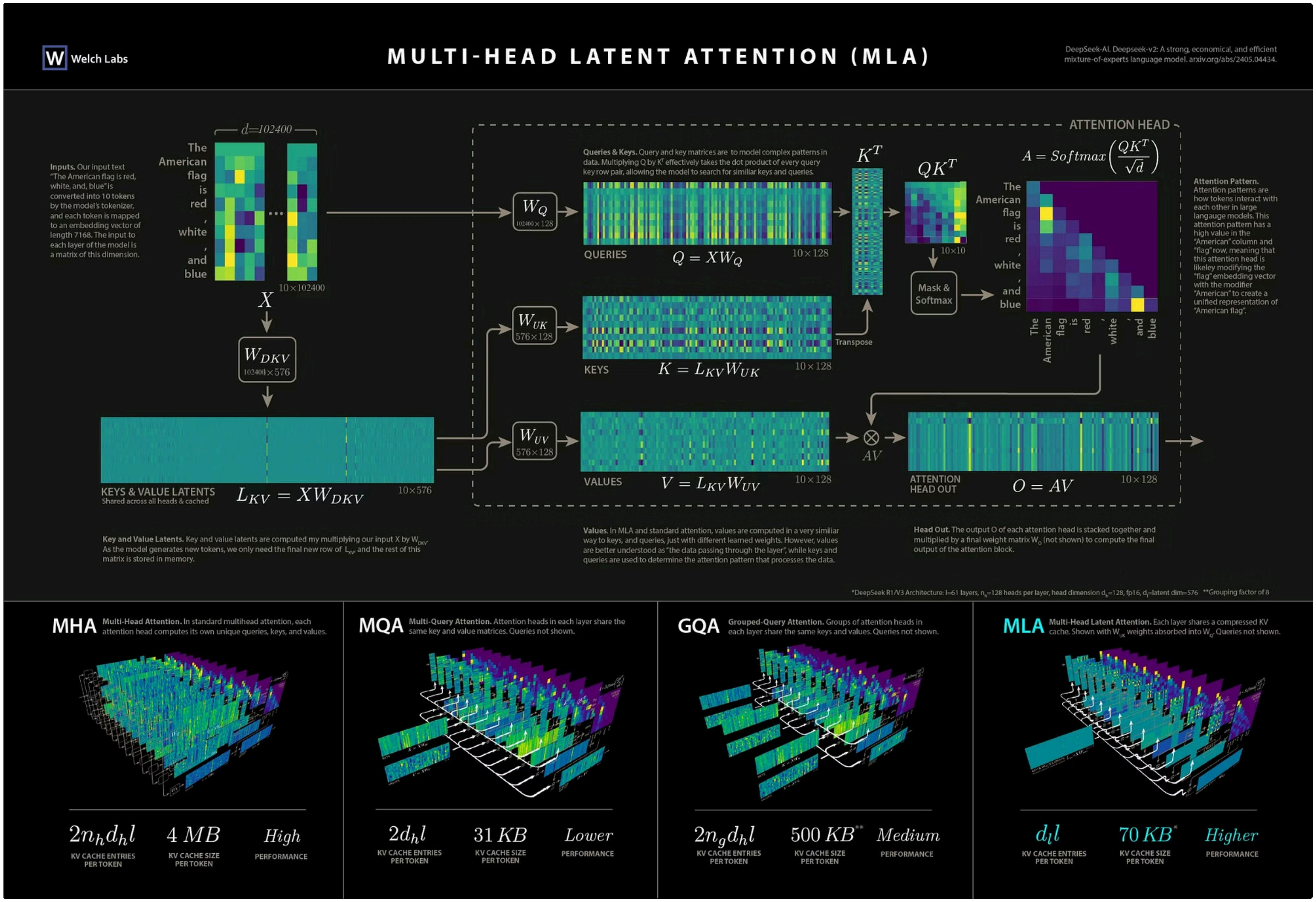

Image Link

$$\boxed{\mathbf{c}_t^{KV}} = W^{DKV}\mathbf{h}_t,$$

$$[\mathbf{k}_{t,1}^C; \mathbf{k}_{t,2}^C; \ldots; \mathbf{k}_{t,n_h}^C] = \mathbf{k}_t^C = W^{UK}\mathbf{c}_t^{KV},$$

$$\boxed{\mathbf{k}_t^R} = \mathrm{RoPE}(W^{KR}\mathbf{h}_t),$$

$$\mathbf{k}_{t,i} = [\mathbf{k}_{t,i}^C; \mathbf{k}_t^R],$$

$$[\mathbf{v}_{t,1}^C; \mathbf{v}_{t,2}^C; \ldots; \mathbf{v}_{t,n_h}^C] = \mathbf{v}_t^C = W^{UV}\mathbf{c}_t^{KV},$$

$$\mathbf{c}_t^Q = W^{DQ}\mathbf{h}_t,$$

$$[\mathbf{q}_{t,1}^C; \mathbf{q}_{t,2}^C; \ldots; \mathbf{q}_{t,n_h}^C] = \mathbf{q}_t^C = W^{UQ}\mathbf{c}_t^Q,$$

$$[\mathbf{q}_{t,1}^R; \mathbf{q}_{t,2}^R; \ldots; \mathbf{q}_{t,n_h}^R] = \mathbf{q}_t^R = \mathrm{RoPE}(W^{QR}\mathbf{c}_t^Q),$$

$$\mathbf{q}_{t,i} = [\mathbf{q}_{t,i}^C; \mathbf{q}_{t,i}^R],$$

$$\mathbf{o}_{t,i} = \sum_{j=1}^{t} \mathrm{Softmax}_j\left(\frac{\mathbf{q}_{t,i}^T \mathbf{k}_{j,i}}{\sqrt{d_h + d_h^R}}\right)\mathbf{v}_{j,i}^C,$$

$$\mathbf{u}_t = W^O[\mathbf{o}_{t,1}; \mathbf{o}_{t,2}; \ldots; \mathbf{o}_{t,n_h}],$$

Rotary Positional Embedding (RoPE)

RoPE is critical for encoding token positions in DeepSeek models, enabling effective generalization to longer sequences while preserving essential vector properties.

Purpose

- Encodes the **absolute position** of tokens.
- Enhances the **relative positional information** for attention mechanisms.

Mechanism

- Splits 64-dimensional vectors into **32 pairs**.
- Applies **rotation to each pair** based on position and frequency.
- The transformation is integrated directly into the **query and key vectors** within MLA.

$$\begin{bmatrix} x'_i \\ x'_{i+1} \end{bmatrix} = \begin{bmatrix} \cos(m\theta_i) & -\sin(m\theta_i) \\ \sin(m\theta_i) & \cos(m\theta_i) \end{bmatrix} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix}$$

$$x'_i = x_i \cos(m\theta_i) - x_{i+1} \sin(m\theta_i)$$
$$x'_{i+1} = x_i \sin(m\theta_i) + x_{i+1} \cos(m\theta_i)$$

Where **m** is the token position and **θ_i** is a predefined frequency for each dimension pair.

Benefits

- **Extends context window** without performance degradation.
- Preserves **vector magnitude** and long-term dependencies.

Training and Data Strategies

DeepSeek R1 and V3 utilize advanced training methodologies and vast, diverse datasets to achieve their impressive capabilities, ensuring robust performance across various tasks.

1

Curated Datasets

Training data includes a mix of text, code, and mathematical content, carefully filtered for quality and diversity to prevent bias.

2

Distributed Training

Leverages large-scale GPU clusters and optimized distributed training frameworks to handle the immense model size and data volume.

3

Fine-tuning & Alignment

Further refined through supervised fine-tuning and reinforcement learning with human feedback (RLHF) to align with human preferences and instructions.

Conclusion and Q&A

> The synergy of MoE, MLA, and RoPE empowers DeepSeek V3 and R1 to achieve remarkable performance with unprecedented efficiency, setting a new standard for large language models.

1. **MoE for Efficiency**

   Sparse activation provides computational gains without sacrificing model capacity.

2. **MLA for Context**

   Compressed attention maintains broad contextual awareness with reduced resource demands.

3. **RoPE for Scale**

   Positional embeddings enable robust performance over very long sequences.

Questions & Discussion

We invite your questions on the technical aspects and implications of DeepSeek R1 and V3's core architecture.