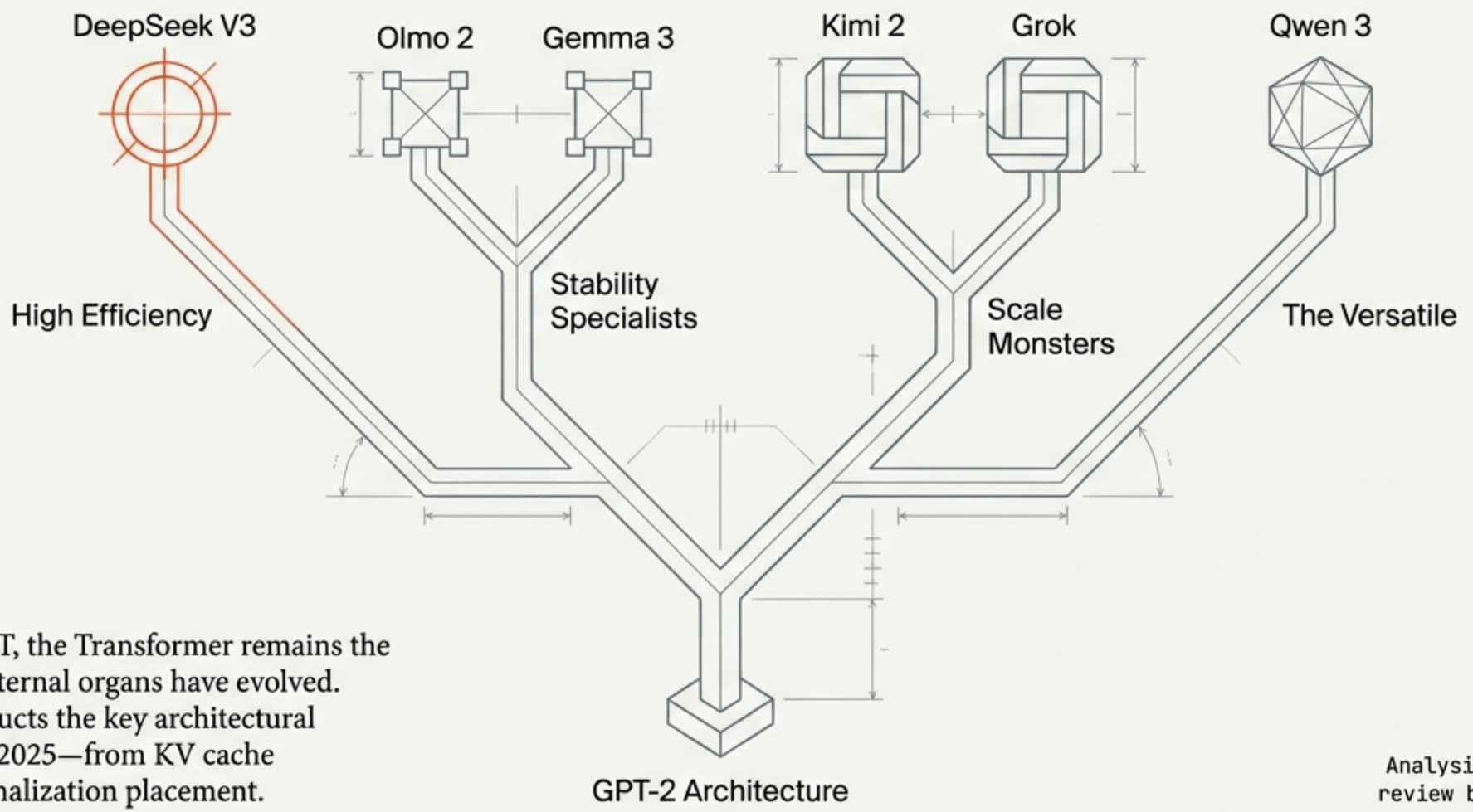
The 2025 LLM Architecture Report

A Comparative Deep Dive: From Scale to Specialization



> Seven years post-GPT, the Transformer remains the backbone, but the internal organs have evolved. This report deconstructs the key architectural innovations of early 2025—from KV cache compression to normalization placement.

Analysis based on technical review by Sebastian Raschka.

2025 State of the Union: The Year of Architectural Efficiency

The Transformer architecture has proven incredibly robust. The core mechanism survives, adapted for massive efficiency gains rather than just raw size.

The KV Bottleneck

Solved via Multi-Head Latent Attention (MLA) and aggressive compression. Moving from memory-heavy caches to compressed latent vectors.

`DeepSeek V3`

MoE Renaissance

Shift from 'Few Large Experts' to 'Many Fine-Grained Experts.' Introduction of 'Shared Experts' to handle common syntax and reduce redundancy.

`DeepSeek, Kimi, GLM`

Stability Engineering

Precise Normalization placement (Post-Norm inside residuals, QK residuals, QK Norm) to smooth gradient spikes and prevent training collapse.

`Olmo 2, Gemma 3`

Context Management

Moving beyond brute force. Adoption of Sliding Window Attention (5:1 ratios) and NoPE (No Positional Embeddings) for length generalization.

`Gemma 3, Small LM3`

DeepSeek V3/R1

The Efficiency Catalyst

DeepSeek V3 redefined the baseline for open weights in January 2025. By combining extreme sparsity (activating only ~5.5% of parameters) with MLA, it fits massive intelligence into manageable inference constraints.

This base model powers the R1 reasoning variant.

Spec Card

Total Parameters

671 Billion

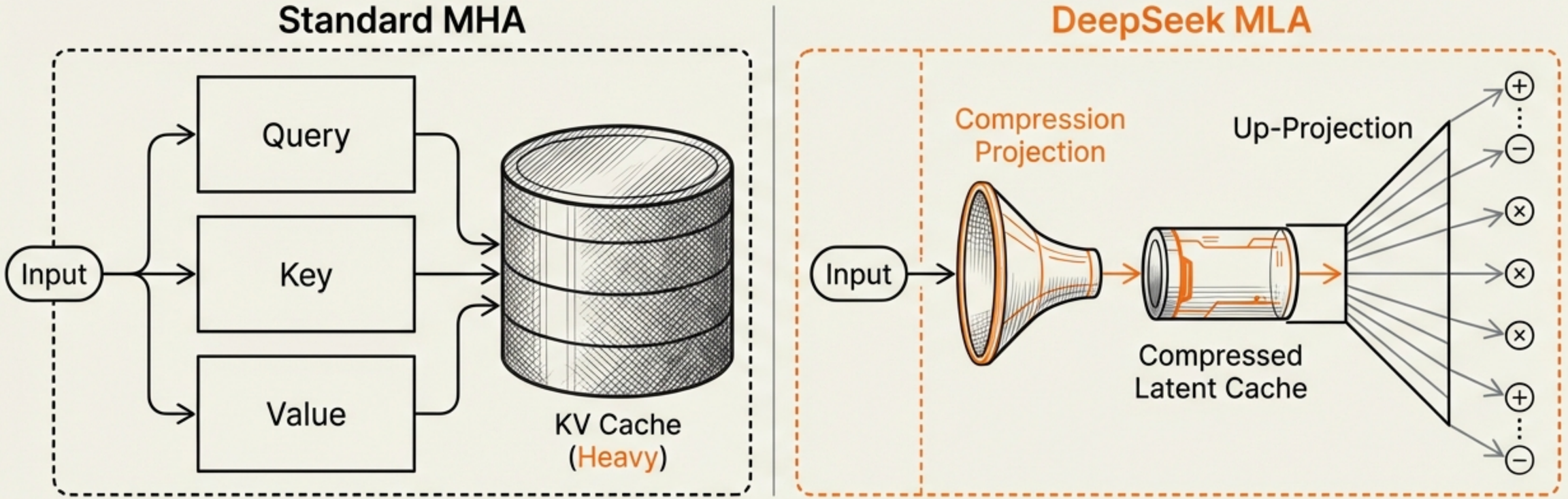
Active Parameters

37 Billion (via MoE)

Multi-Head Latent Attention (MLA)	→ Massive KV Cache Compression
Fine-Grained MoE	→ 256 Total Experts + 1 Shared Expert

Solving the Memory Bottleneck: Multi-Head Latent Attention (MLA)

The Exploded View

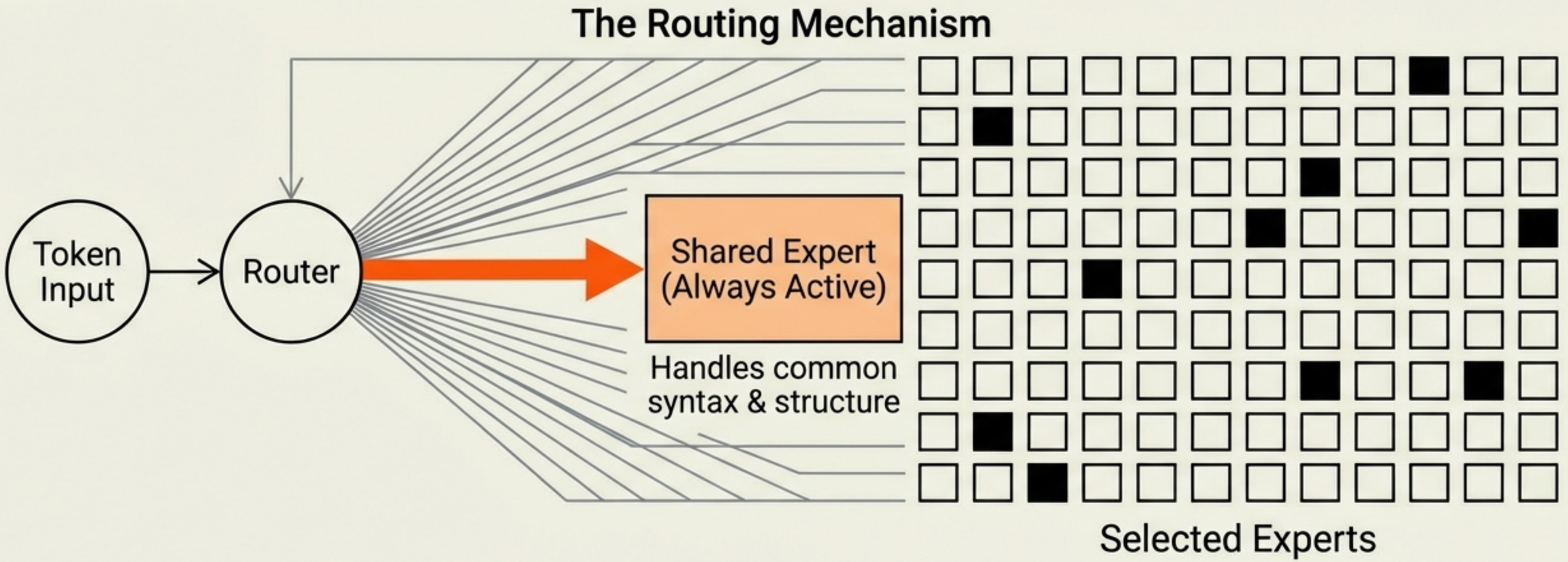


	Method	Params per Token (KV Cache)	Impact
2	Standard MHA	110,000	Baseline Memory High
3	Grouped Query (GQA)	15,000	Memory Low, Accuracy Drop
4	**DeepSeek MLA**	**15,000**	**Memory Low, Accuracy Boost**

MLA trades compute for memory. Storing a compressed latent vector reduces cache size by ~20x without the accuracy penalty of GQA.

The MoE Renaissance: Fine-Grained & Shared Experts

Moving from sparse to fine-grained routing allows a 671B model to run **with the inference cost of a 37B model.**

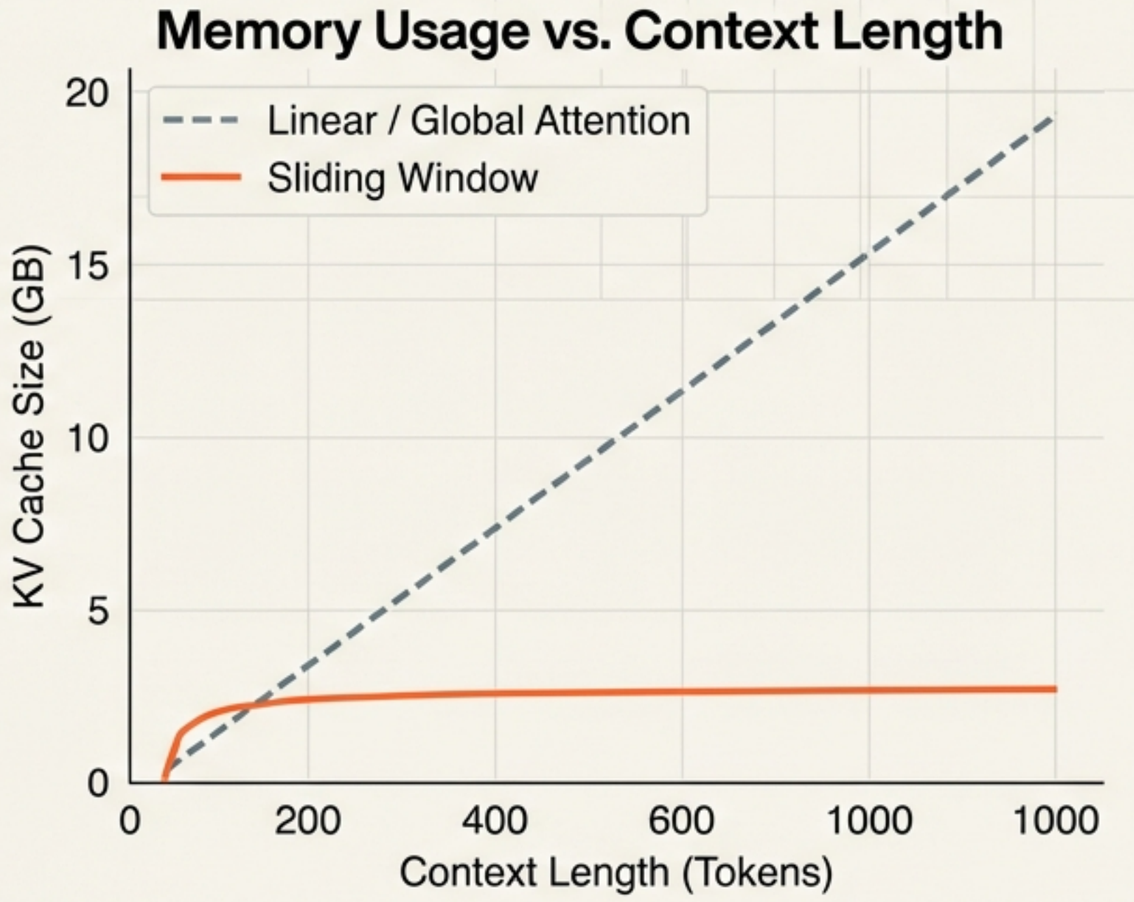


Old School MoE: ~8 Large Experts (2 Active)
2025 Standard (DeepSeek): 256 Small Experts (8 Active + 1 Shared)

Gemma 3: Optimizing Context with Sliding Window Attention

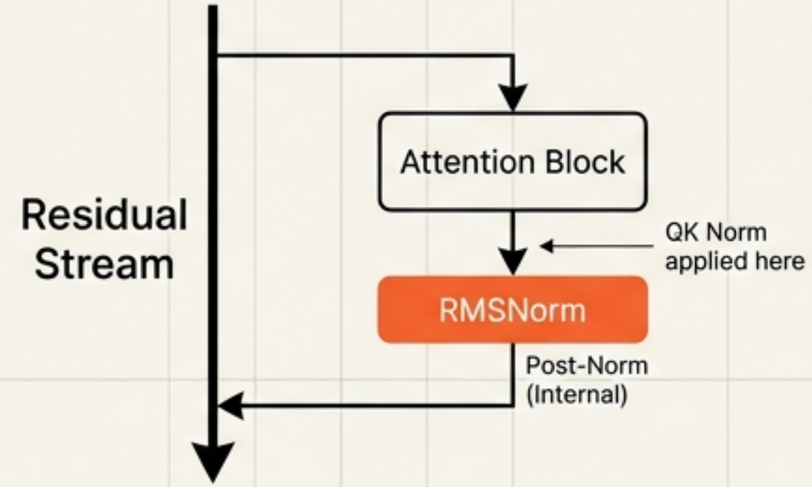
Google’s ablation studies showed that moving from a 1:1 to a 5:1 Local-to-Global ratio had negligible impact on perplexity but massive gains in memory efficiency.

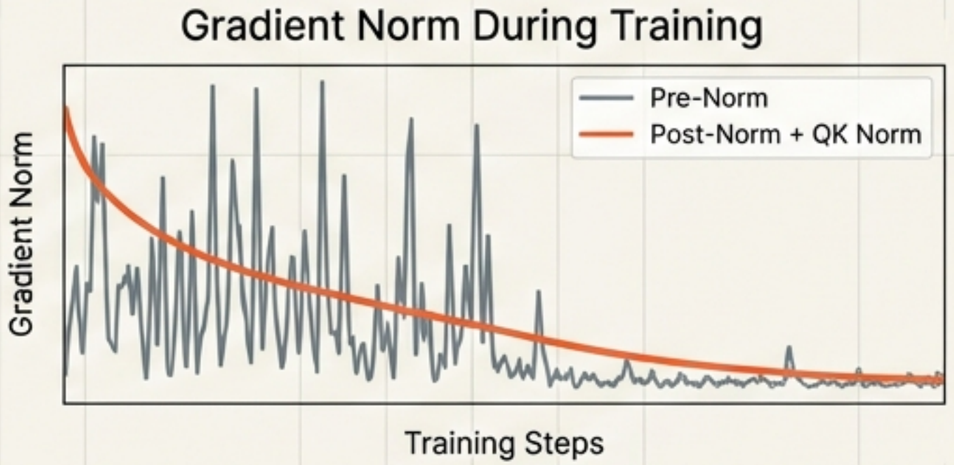
`Ratio: 5 Local Layers : 1 Global Layer.`
`Local Window: 1,024 Tokens.`



Olmo 2 & The Science of Training Stability

Preventing loss spikes through Normalization placement.

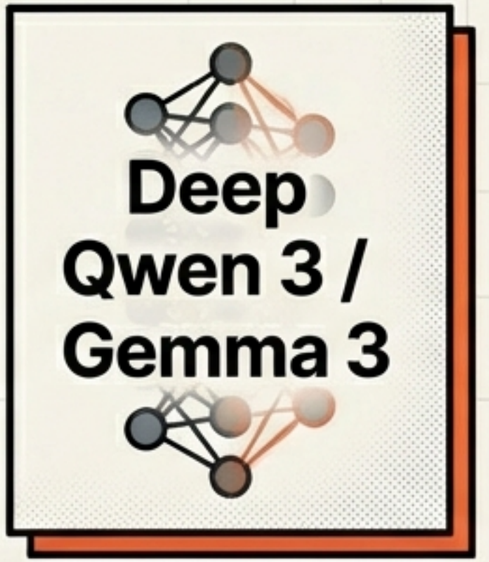




Architecture Shape: Width vs. Depth

40 Layers
40 Attention Heads

Prioritizes Inference Speed.
Fewer layers = Faster sequential processing.



60+ Layers
32 Attention Heads

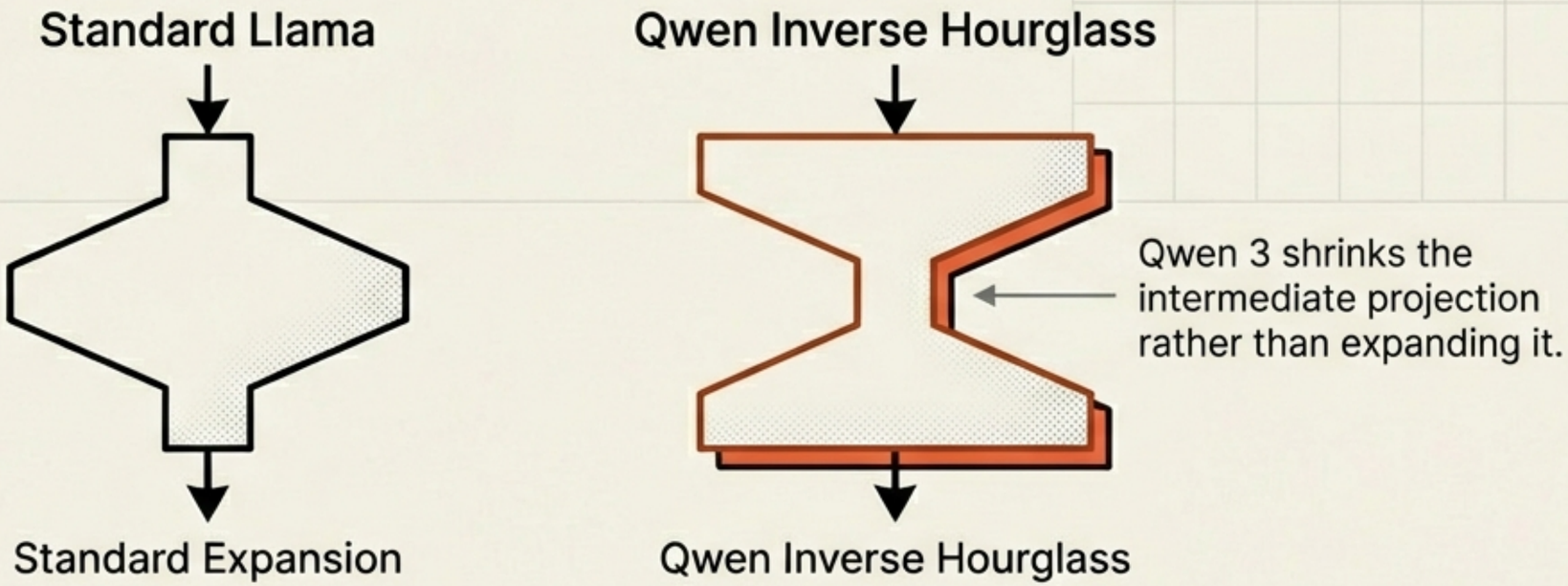
Prioritizes Capacity.
More layers = Better complex reasoning, higher latency.

Trade-off: Mistral aims for Tokens/Sec; Qwen aims for reasoning depth.

Qwen 3: The Versatile All-Rounder

Unique Feed-Forward Design and Hybrid Reasoning

The MLP Hourglass



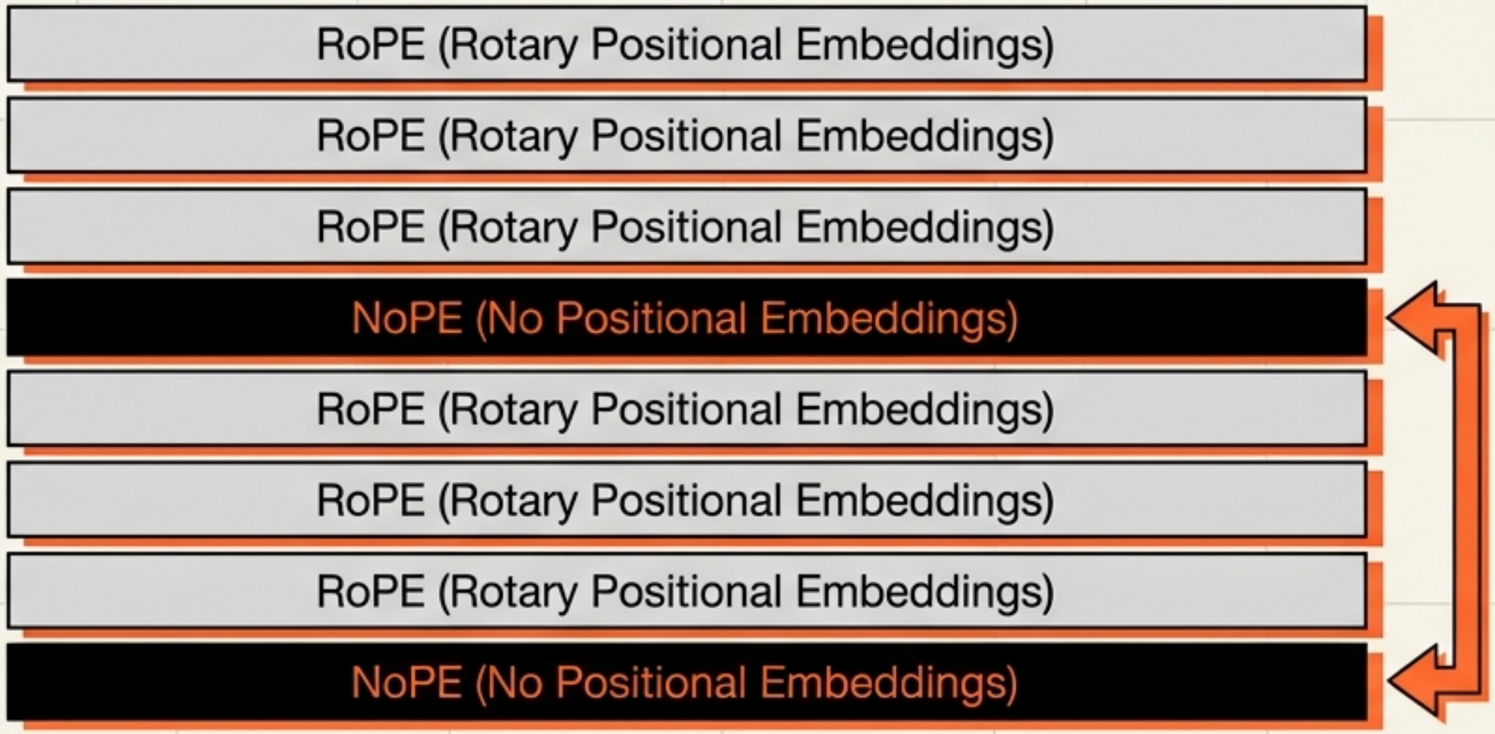
Hybrid Nature: Uses a ‘Think’ token to toggle between Base and Instruct/Reasoning modes within the same architecture. Available in both Dense and MoE variants.

JetBrains Mono: numotion :uses, inks, slates, emotions/ıɒɣn monoɛcat.

Small LM3: The ‘NoPE’ Experiment

Counter-intuitive finding: Explicit position data might overfit the model to specific lengths. Removing embeddings forces intrinsic learning.

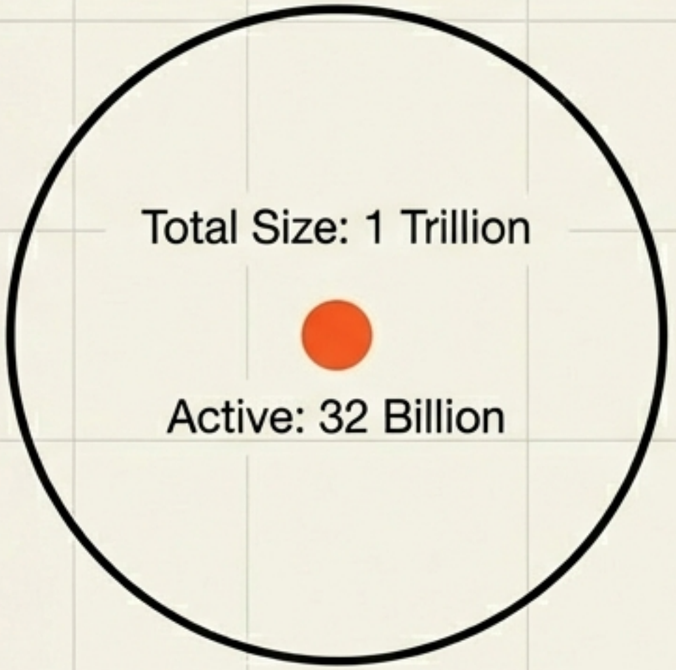
Layer Strategy



Scaling Up: Kimi 2 (1 Trillion Parameters)

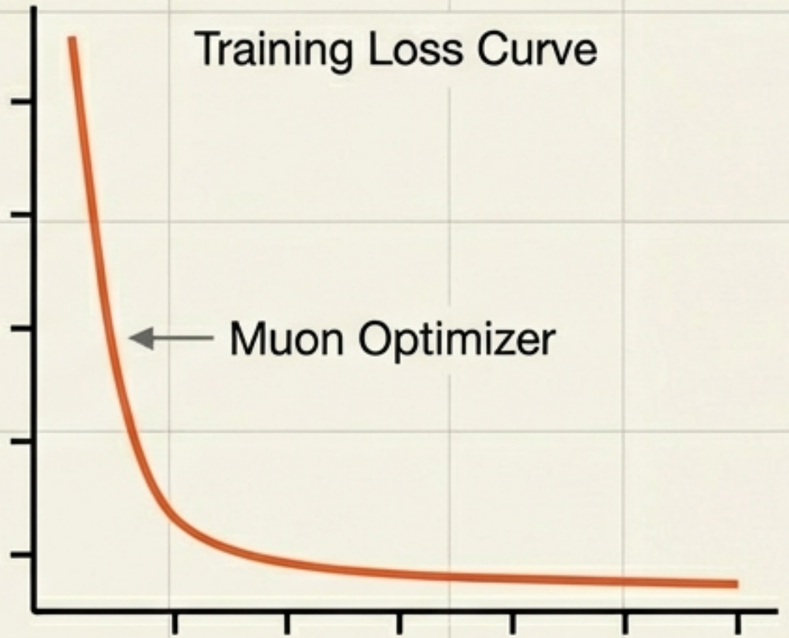
Comparison of model scaling, active parameters, and optimization techniques.

The Efficiency Paradox



Larger total size than DeepSeek (671B), but FEWER active parameters per token.

Muon Optimizer Impact



Uses momentum update on orthogonal non-linearities for steeper, faster learning.

First 3 blocks are Dense to prevent expert collapse.

Technical Editorial

The Incumbents: GPT-OSS & Legacy Design

The first open weights from OpenAI since GPT-2 (likely GPT-4o-mini).

The Linear Layer Inspection

Modern Standard (Llama/Mistral)

$$Y = X * W$$

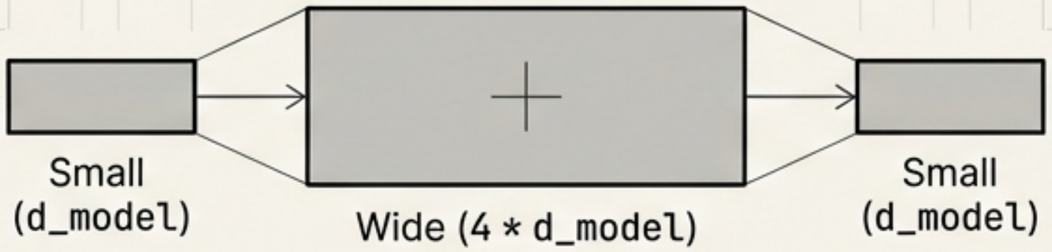
GPT-OSS (Legacy)

$$Y = X * W + [\text{Bias Vector}]$$

Bias Vector: A legacy holdover from GPT-2. Adds compute with negligible performance benefit.

Traditional 4x Expansion

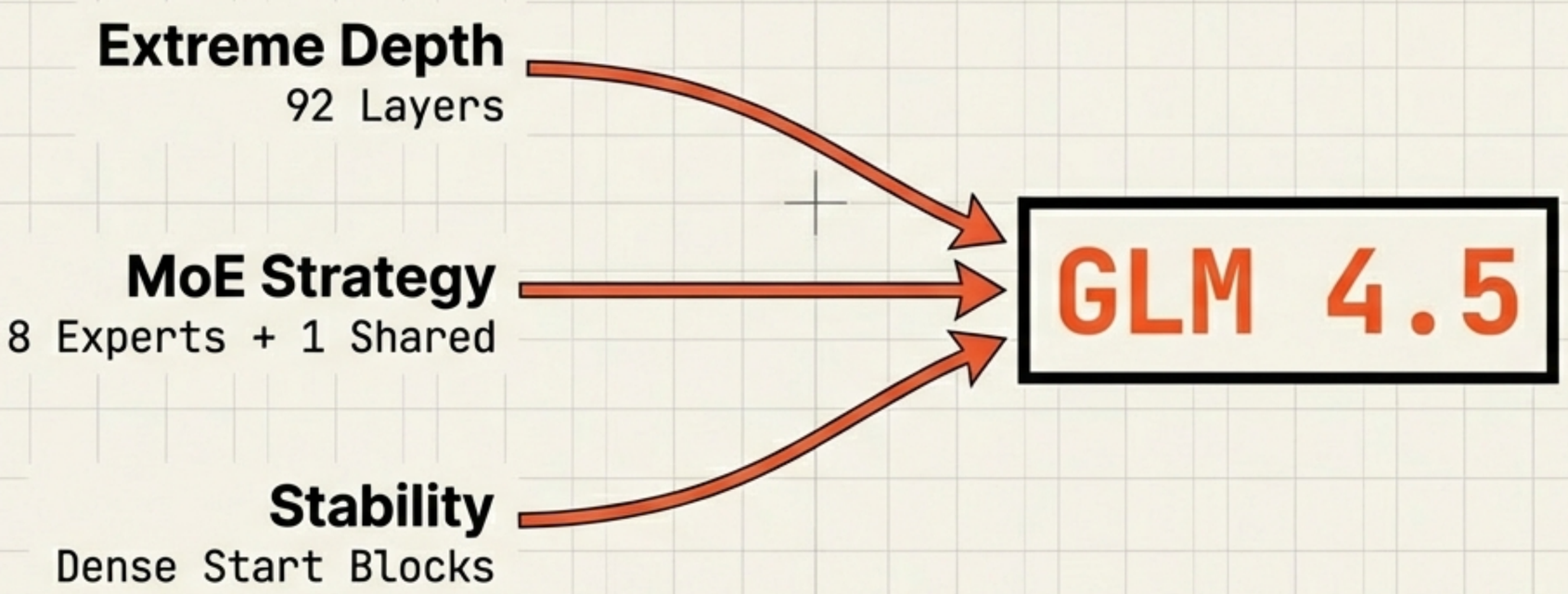
"Inverse Hourglass" Feed-Forward



Technical Editorial

High Performance Convergence: GLM 4.5

GLM 4.5 represents the synthesis of 2025's best practices, combining extreme depth with the Shared Expert MoE strategy.



A hybrid design ensuring training stability before switching to highly efficient MoE layers.

Technical Editorial

The 2025 Architecture Cheat Sheet

Model	Attention	Norm Strategy	MoE / Structure	Unique Feature
DeepSeek V3	MLA (Compressed)	RMSNorm	MoE (Shared + Fine)	**Active Params < 6%**
Gemma 3	Sliding Window (5:1)	Post+Pre+QK	Dense (Deep)	**Context Efficiency**
Olmo 2	Standard MHA	Post-Norm (Internal)	Dense	Training Stability
Qwen 3	GQA	RMSNorm	Hourglass MLP	**Hybrid Reasoning**
Kimi 2	GQA	RMSNorm	MoE (Dense Start)	**Muon Optimizer**
Mistral Small	GQA	RMSNorm	Wide / Shallow	Inference Speed

Technical Editorial

Conclusion: The Robustness of the Transformer

Despite distinct approaches—DeepSeek's compression, Google's windowing, and Kimi's scaling—the underlying Transformer architecture remains surprisingly uniform. It survives diverse engineering paths to converge on high performance.

Base Architecture (2025)

The Next Frontier

The focus shifts from how the model is BUILT to how it THINKS. Verifiable reasoning chains (DeepSeek R1 / Qwen Instruct).