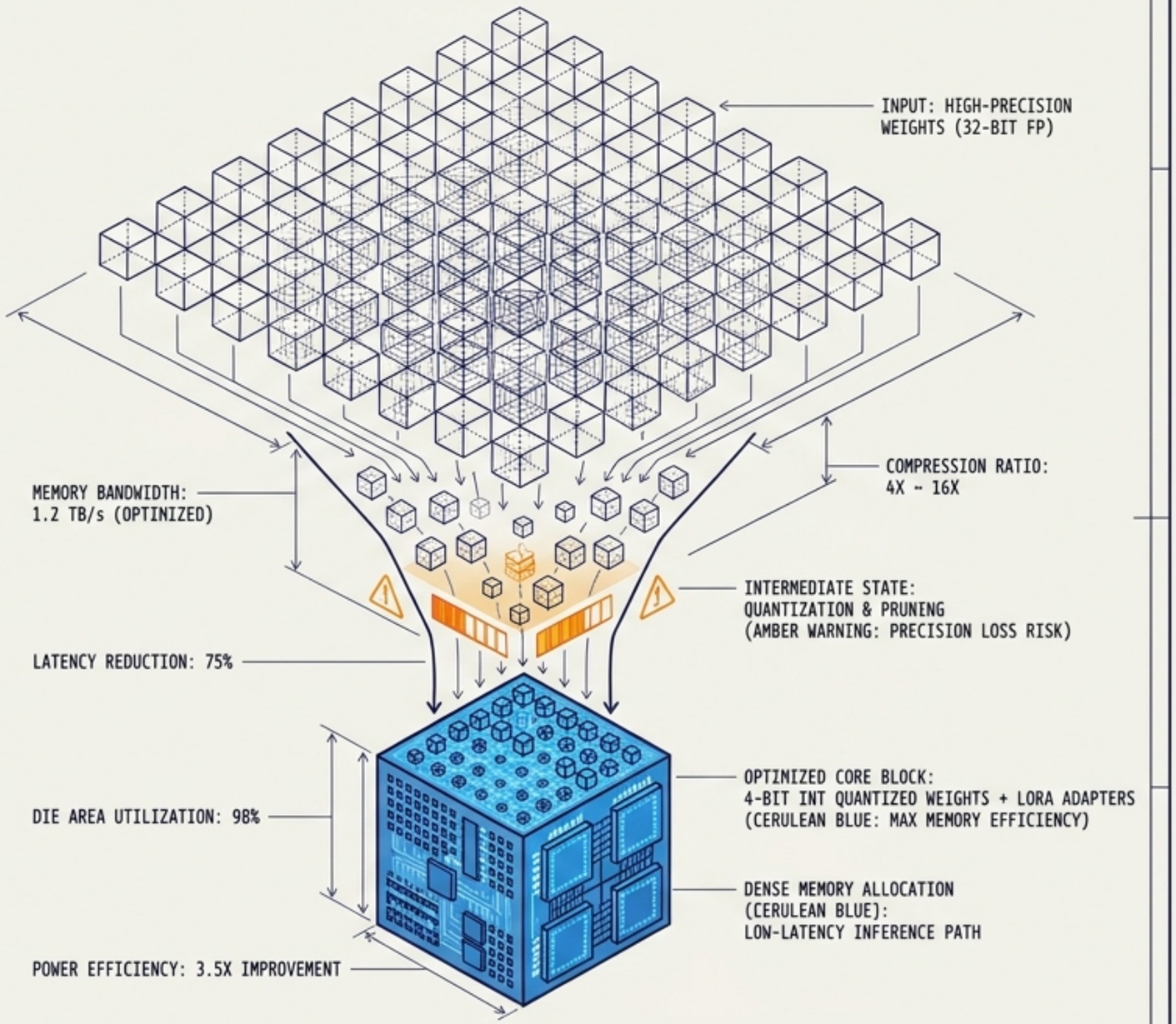
AI Basics — S01E05

Model Compression, Quantization, LoRA & Inference Memory Management

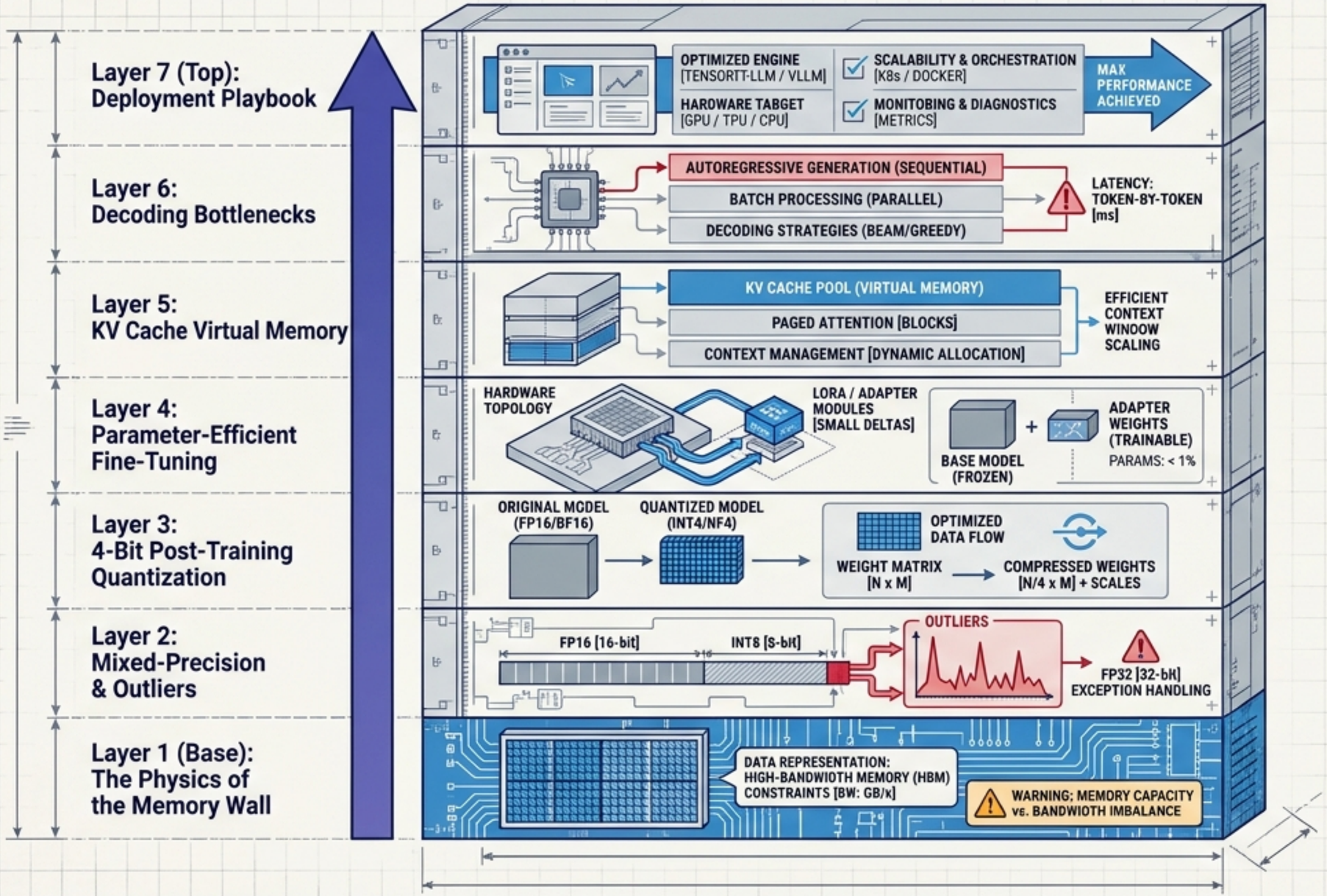
An Architecture Playbook for ML & Infrastructure Engineers



PROJECT: AI INFRASTRUCTURE | SCHENATIC ID: AI-S01E05-MC-LORA | DATE: OCT 26, 2024 | REVISION: A.1

Gemini Notebook

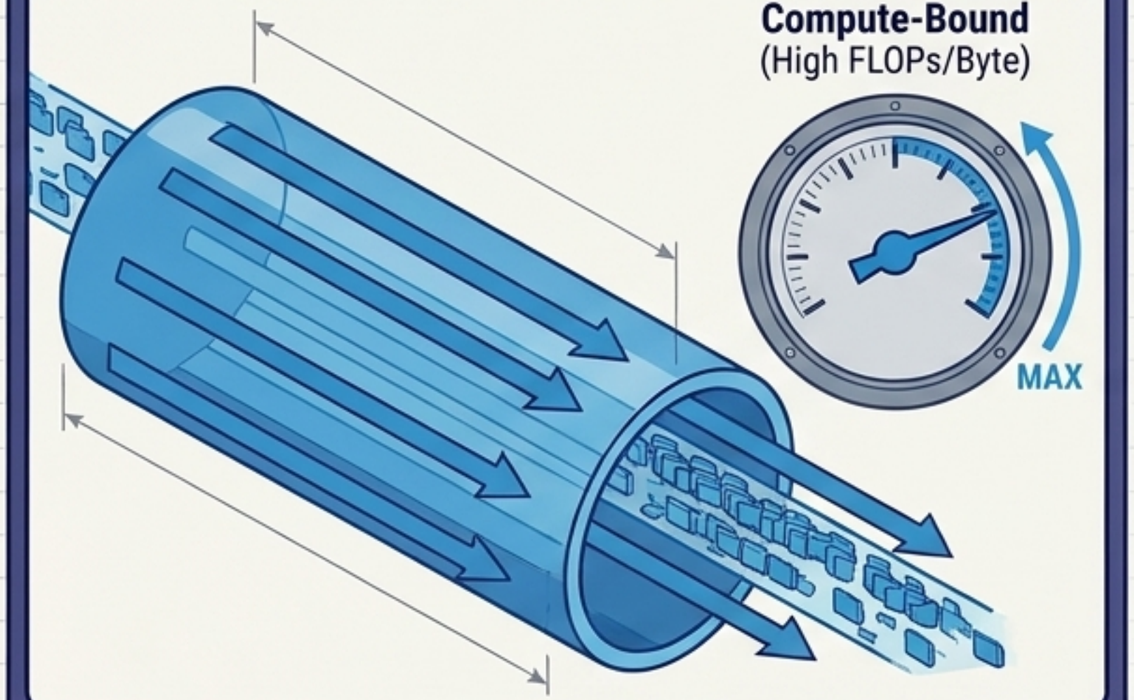
THE INFERENCE OPTIMIZATION STACK



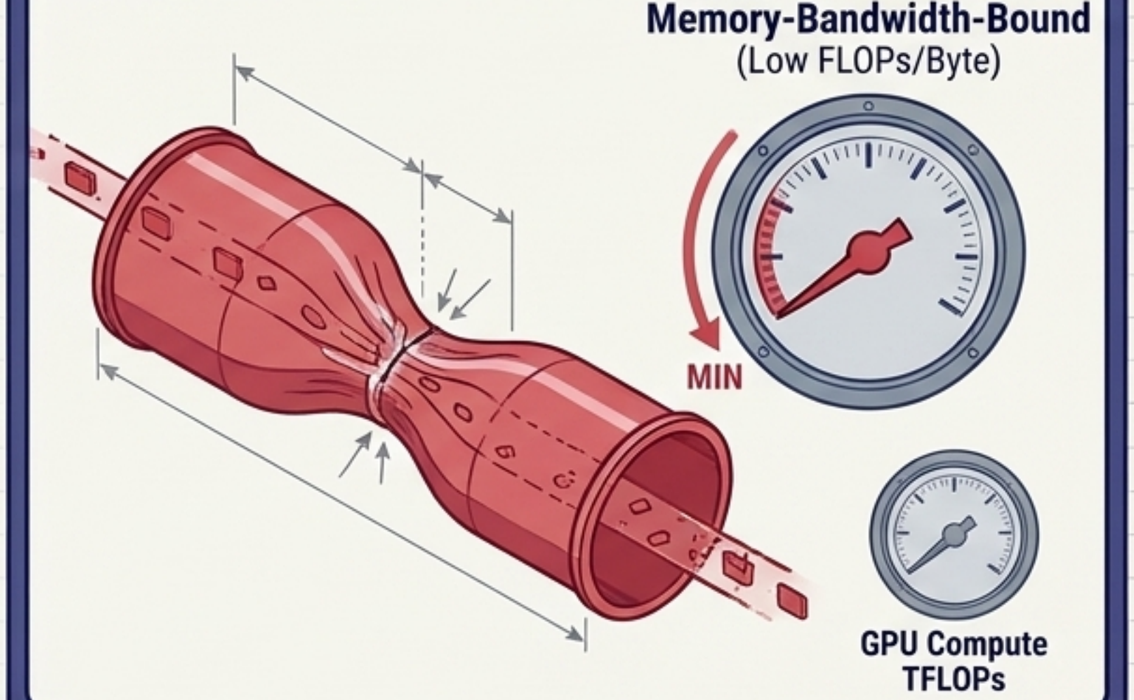
Gemini Notebook

The Physics of Autoregressive Decoding

Prefill Phase



Decode Phase



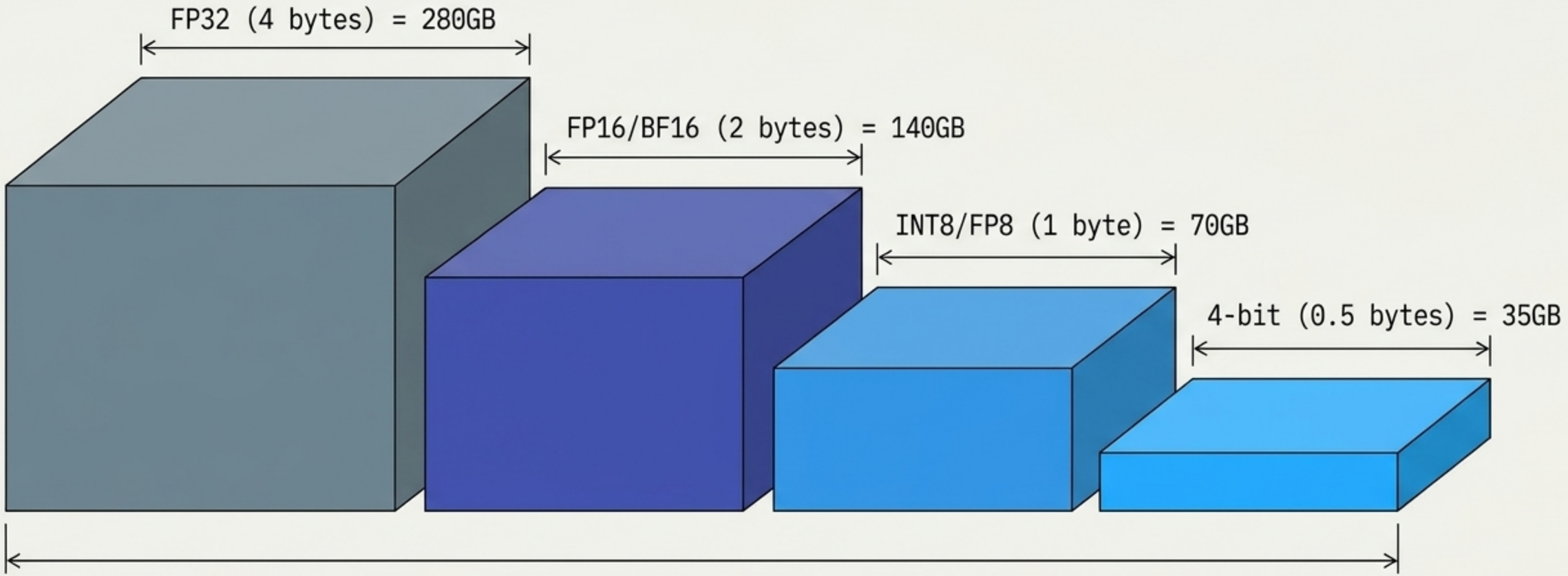
Arithmetic Intensity = FLOPs / Memory Bytes.

Generating exactly 1 token requires reading the entire model weight matrix from VRAM. This yields an arithmetic intensity of ~1-2 FLOPs/Byte—mathematically capping performance against the memory bandwidth limits of H100s or Apple Silicon.

Gemini Notebook

BYTE MATH: THE VRAM CASCADE

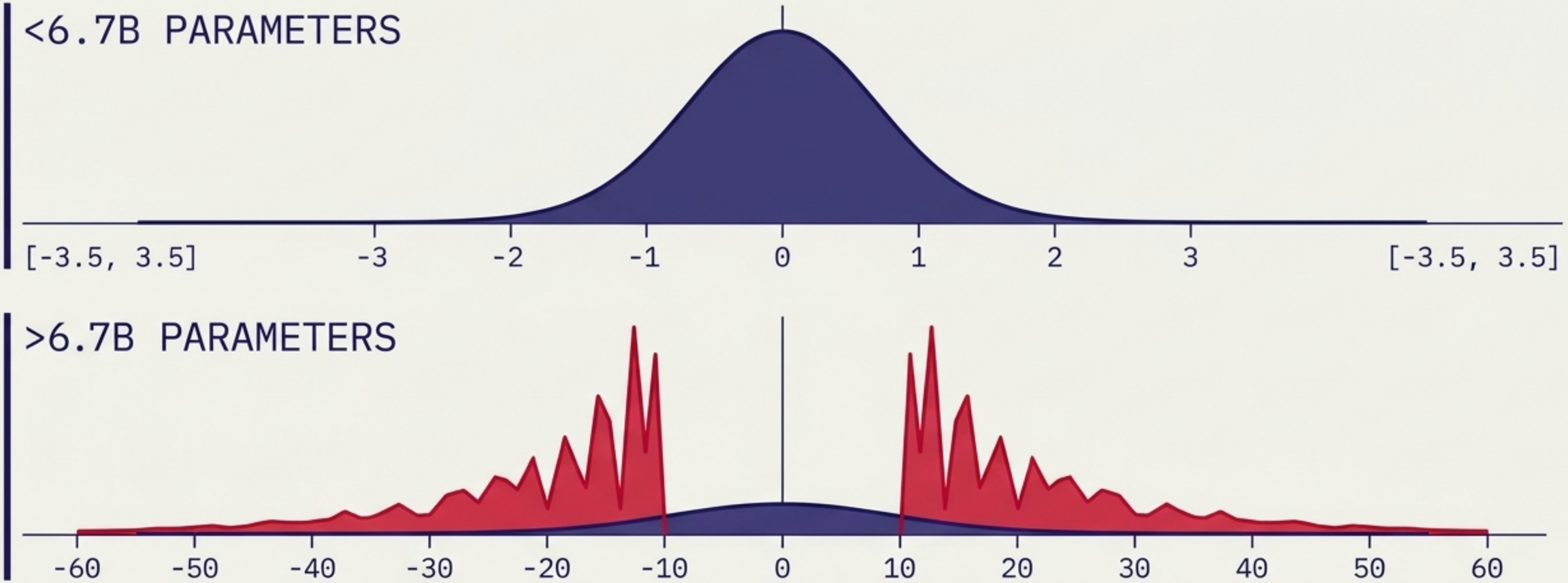
70B parameter model footprint



Moving from FP16 to 4-bit delivers a 4x reduction in total VRAM footprint. Assuming a purely memory-bound constraint, this structurally guarantees a theoretical 4x increase in decode token generation speed.

Gemini Notebook

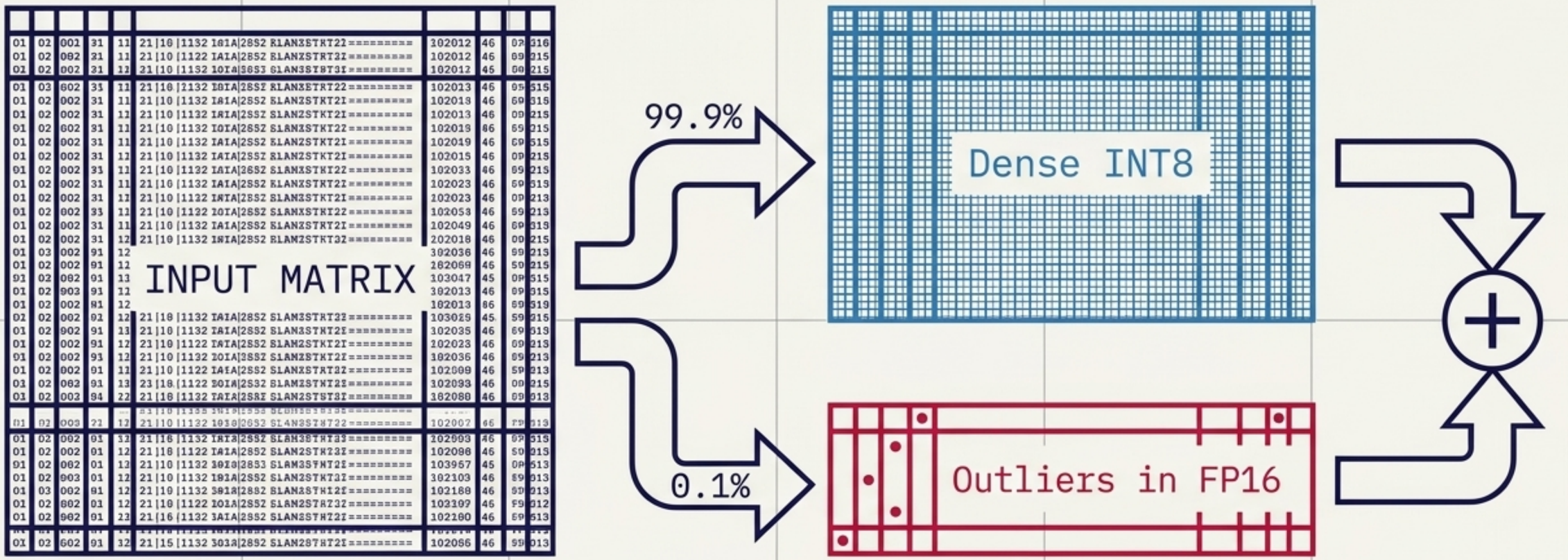
THE 6.7B PARAMETER OUTLIER PHENOMENON



Naive INT8 rounding fails at scale. If an entire tensor is scaled to accommodate a single magnitude-60 outlier, the majority of normal activations are mathematically crushed to zero, wrecking the model's predictive capacity.

Gemini Notebook

LLM.int8() Two-Part Matrix Decomposition

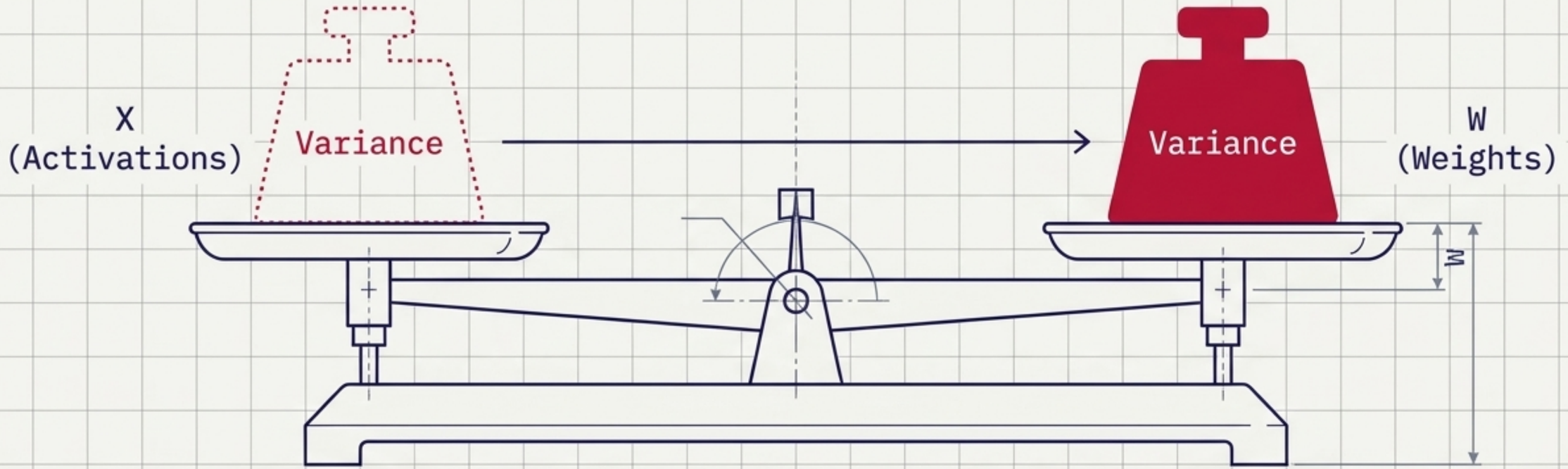


Vector-wise quantization. By isolating just the 0.1% outlier dimensions into 16-bit math, LLM.int8() preserves 99.9% 8-bit efficiency while achieving zero-degradation inference at 175B scale.

Gemini Notebook

SmoothQuant: Migrating Variance Offline

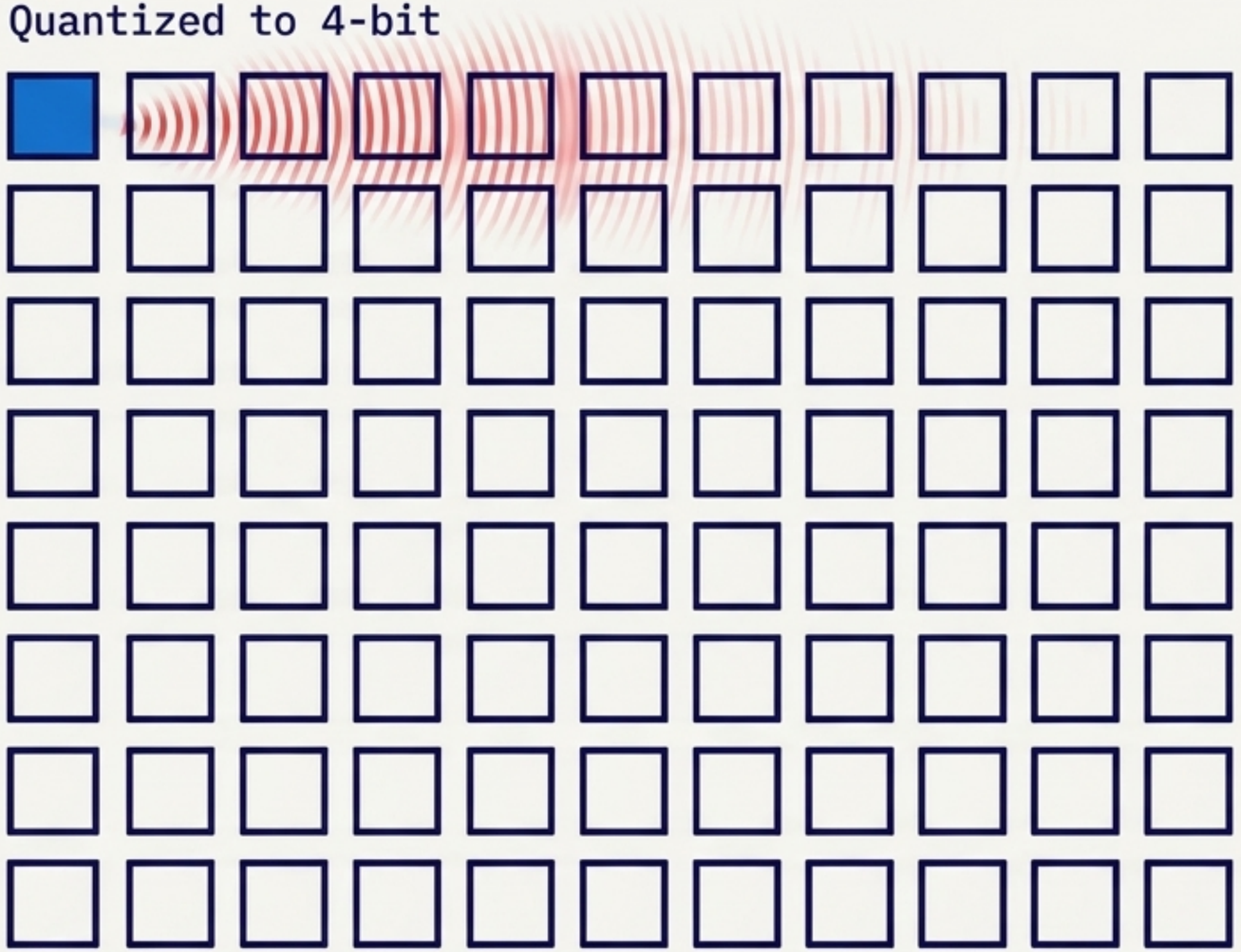
$$Y = (X * \text{diag}(s)^{-1}) * (\text{diag}(s) * W)$$



Activations are dynamic and exceptionally hard to quantize. Weights are static. SmoothQuant mathematically pushes the outlier variance into the weight tensor offline, enabling hardware-efficient W8A8 (Weight-8, Activation-8) execution without mixed-precision overhead.

Gemini Notebook

GPTQ & Optimal Brain Quantization

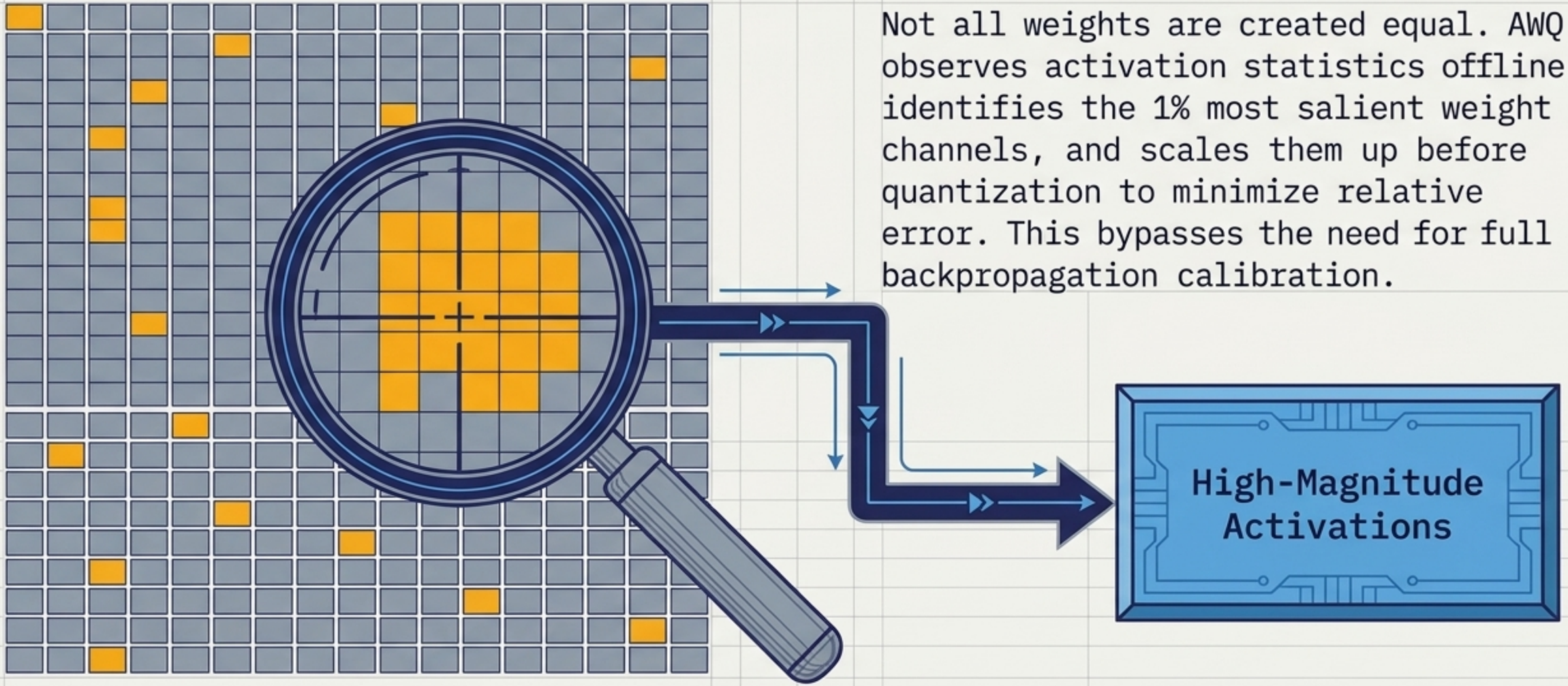


GPTQ doesn't just clip and round. Utilizing approximate second-order Hessian data, it greedily compensates for the quantization error of early weights by actively adjusting the remaining weights in the row.

Result: Near-lossless 3-4 bit compression in GPU-hours.

Gemini Notebook

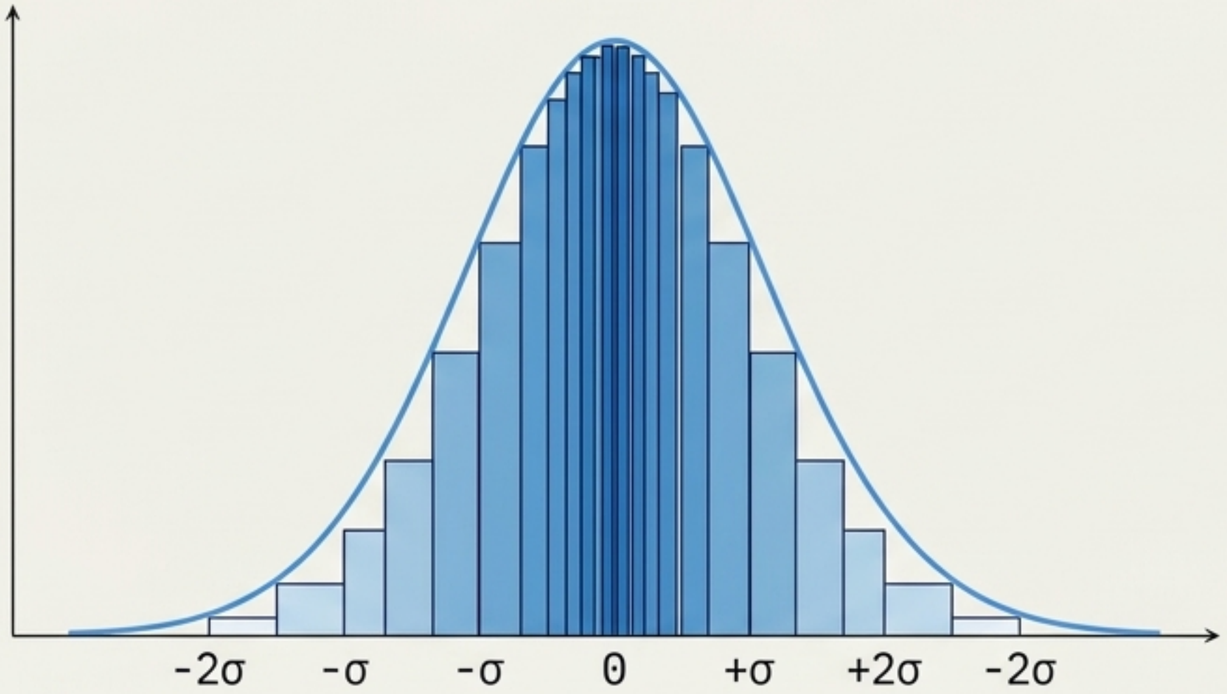
AWQ: Activation-Aware Weight Quantization

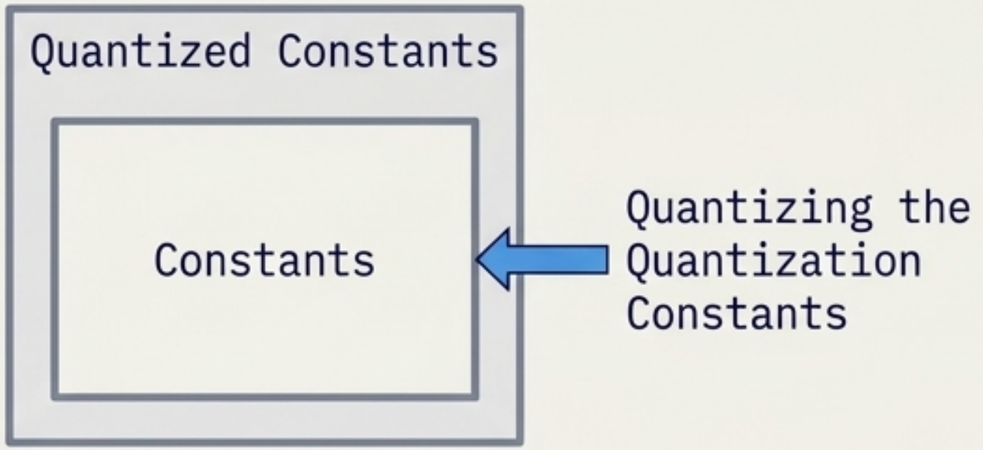


Not all weights are created equal. AWQ observes activation statistics offline, identifies the 1% most salient weight channels, and scales them up before quantization to minimize relative error. This bypasses the need for full backpropagation calibration.

Gemini Notebook

The Information-Theoretic Peak: NormalFloat4 (NF4)

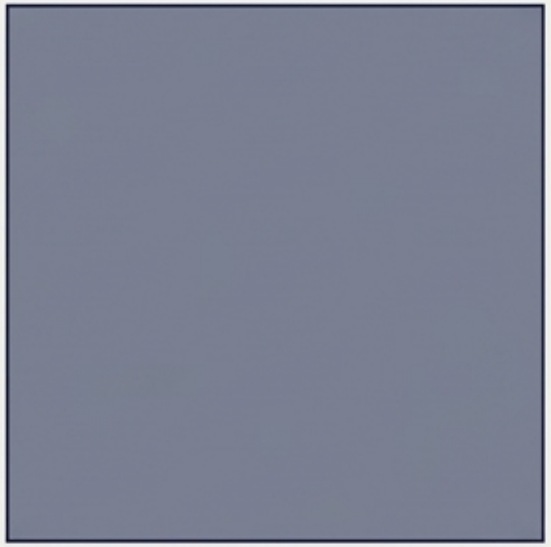




1	**NF4 Bins:** 16 optimally spacd bins designed strictly for the normal distribution of pre-trained weights.
2	**Double Quantization:** Quantizing the scales themselves saves an additional 0.4 bits/parameter (shrinking a 65B model by an extra ~3GB).

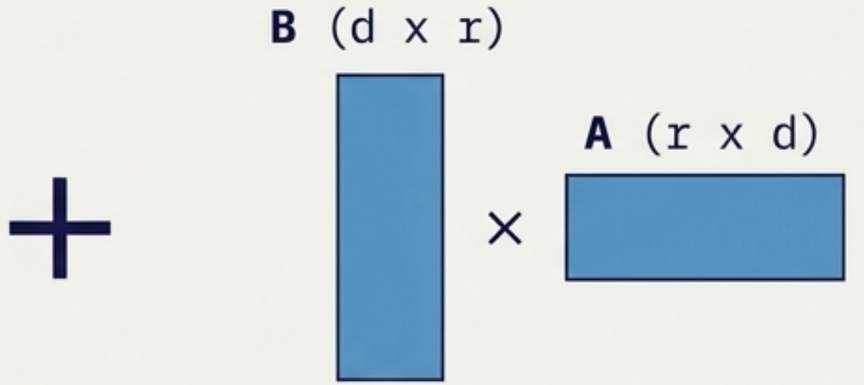
Gemini Notebook

The Intrinsic Rank Hypothesis (LoRA)



W (d x d)

Frozen Base Matrix

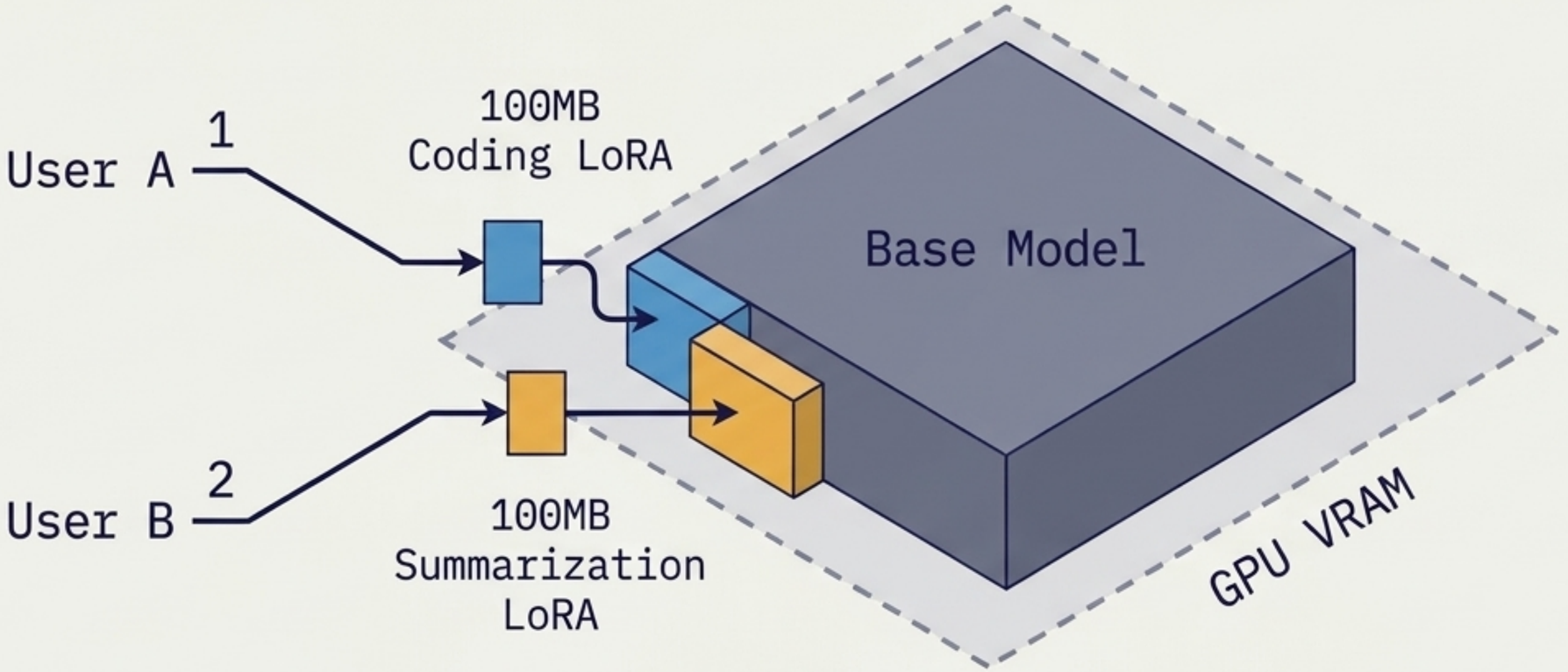


r << d

1	**Full fine-tuning overwrites W**	2	**Inference**
	LoRA freezes **W** and trains $\Delta \mathbf{W} = \mathbf{B} \times \mathbf{A}$. This eliminates full-matrix Adam optimizer states, shrinking trainable parameters by 10,000x.		Note: At runtime, **B** × **A** is merged permanently back into **W** for zero-latency inference overhead.

Gemini Notebook

Multi-Tenant Serving & Dynamic Swapping



Dynamic LoRA routing allows a single expensive GPU cluster to simultaneously serve hundreds of highly specialized, fine-tuned models with near-zero VRAM duplication.

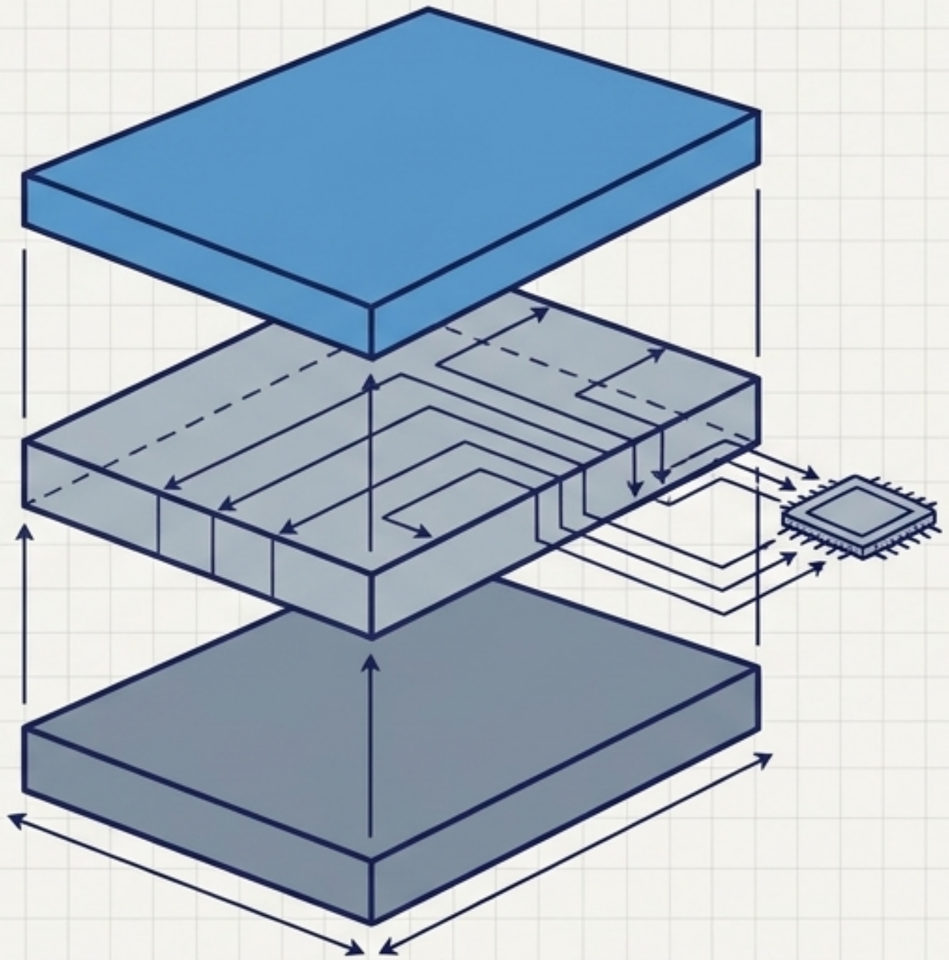
Gemini Notebook

QLoRA: The Ultimate Tuning Stack

Compute: Trainable FP16 LoRA Adapters

Memory Mgmt: Double Quantization + Paged Optimizers (managing memory spikes to CPU)

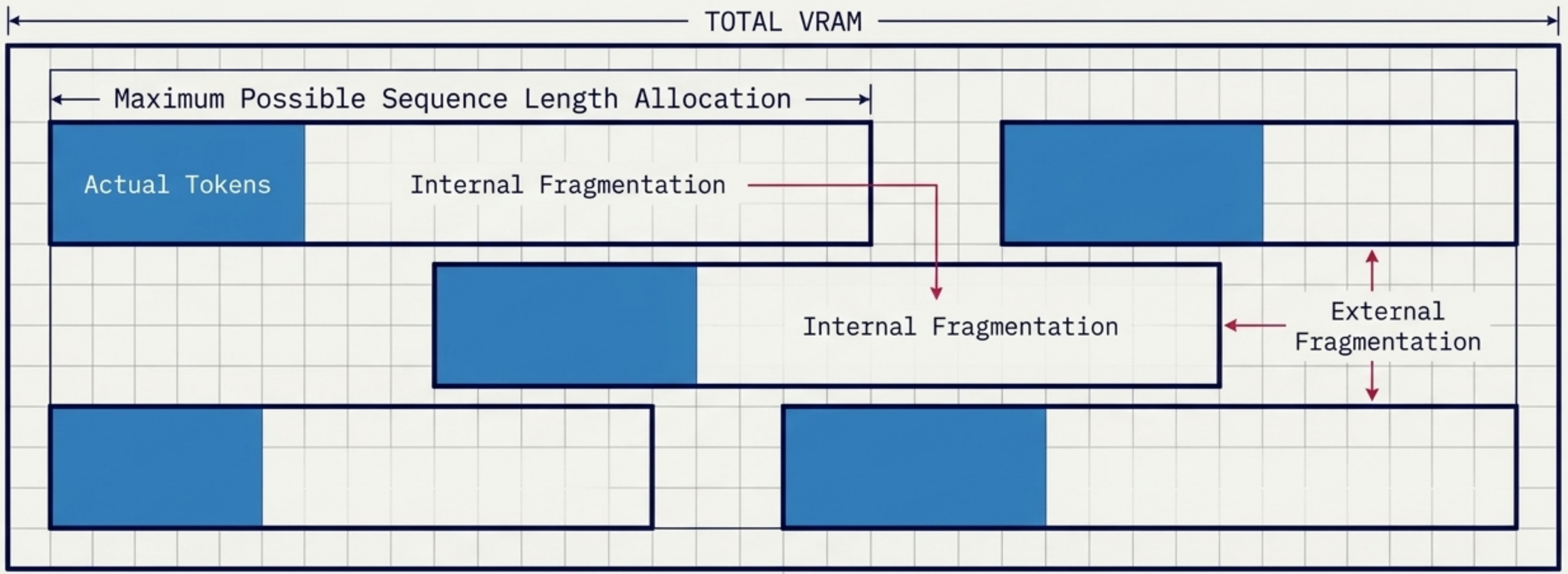
Storage: Frozen NF4 Base Model (4-bit)



By combining 4-bit PTQ with low-rank adapters, QLoRA enables a 65B parameter model to be rigorously fine-tuned on a single 48GB consumer GPU with zero performance degradation.

Gemini Notebook

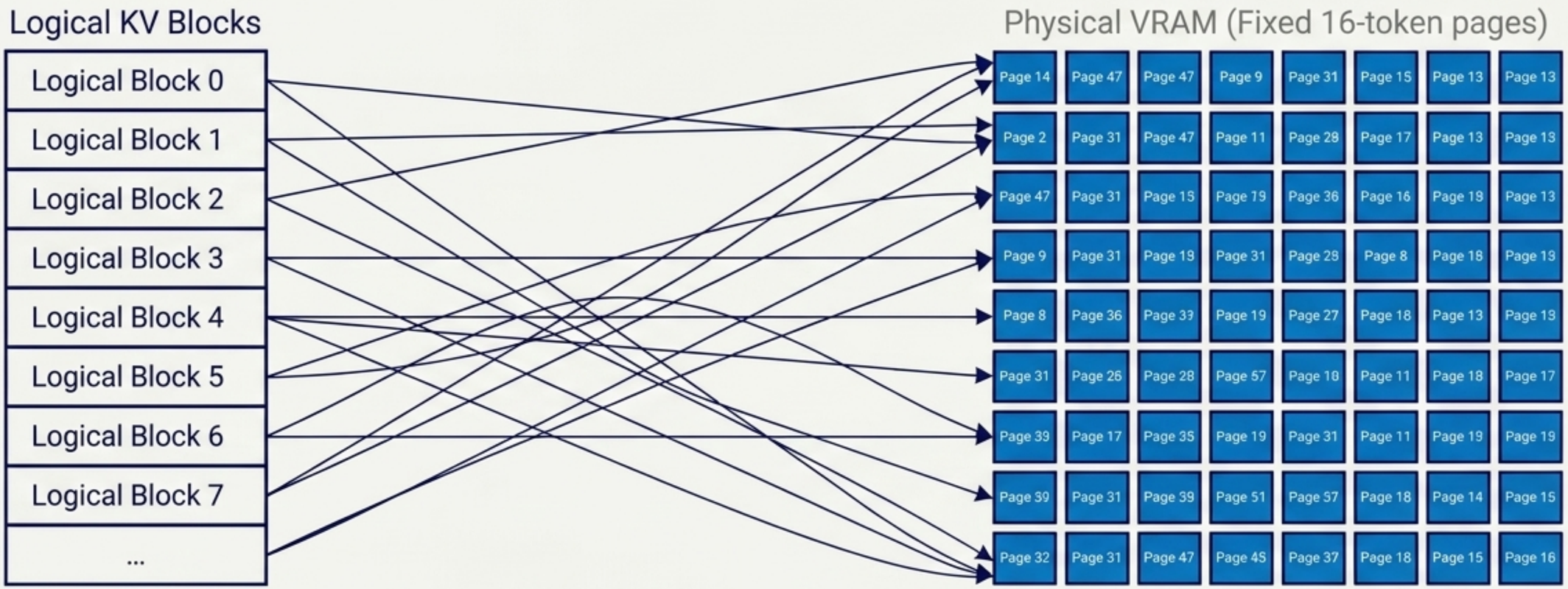
The KV Cache Memory Leak



In standard serving engines, KV cache sizing is unpredictable. Mandating contiguous memory allocation results in up to 60% of total VRAM being completely wasted, severely limiting batch sizes.

Gemini Notebook

vLLM & PagedAttention

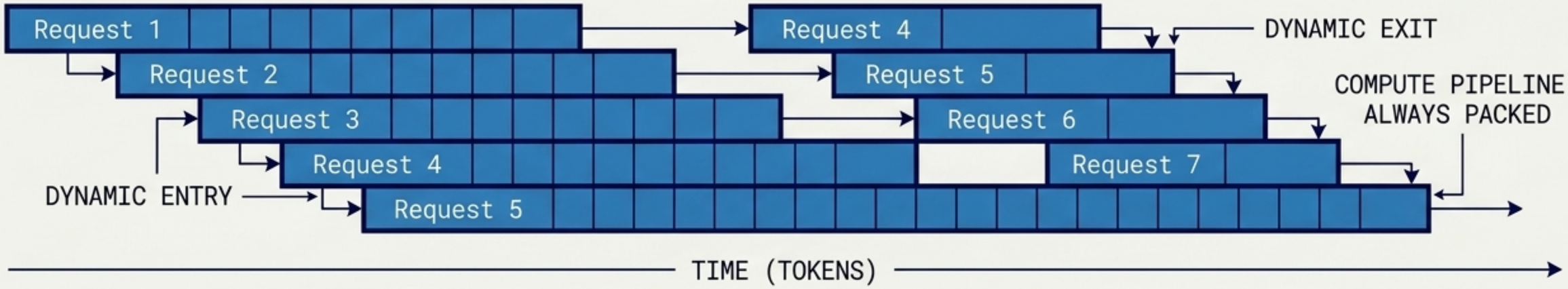


- Inspired by OS virtual memory, PagedAttention breaks the KV cache into fixed-size blocks mapped to non-contiguous physical VRAM.
- **Result**: Near-zero memory waste. Allows batch sizes to scale 2x–4x higher on identical hardware.

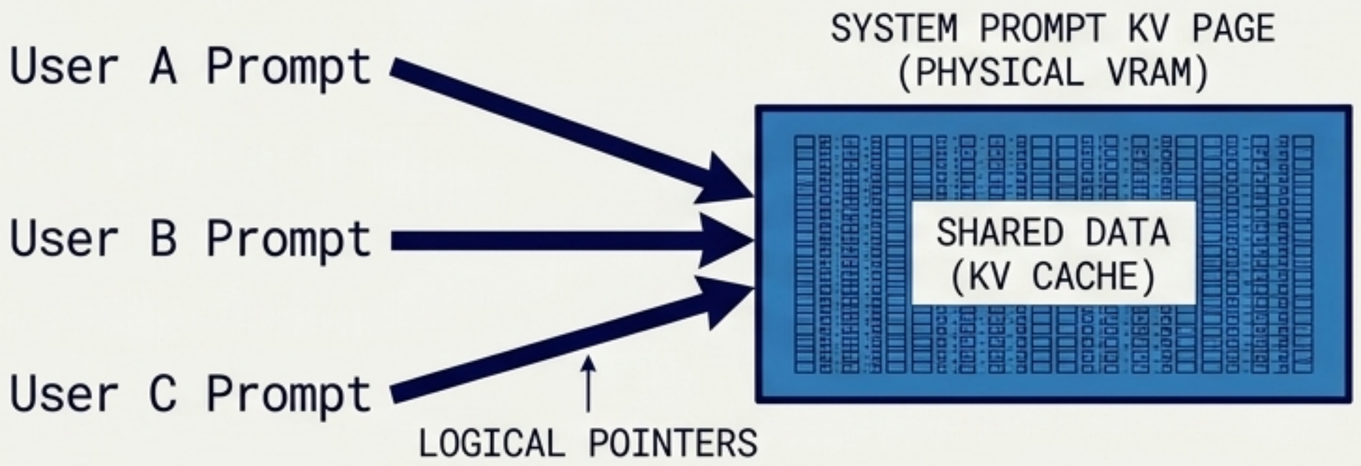
Gemini Notebook

Continuous Batching & Copy-on-Write

CONTINUOUS BATCHING: DYNAMIC REQUEST FLOW



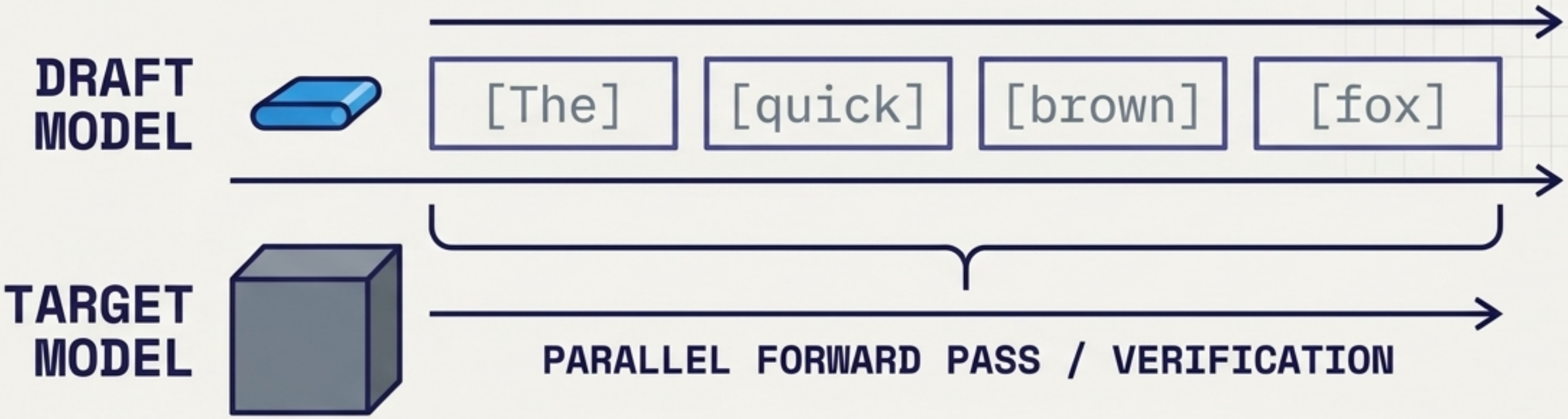
COPY-ON-WRITE: SHARED SYSTEM PROMPT



Paged block mapping enables complex sharing. If 100 concurrent users share the same system prompt, vLLM stores the KV cache for it exactly once and maps 100 logical pointers to it, saving massive memory bandwidth.

Gemini Notebook

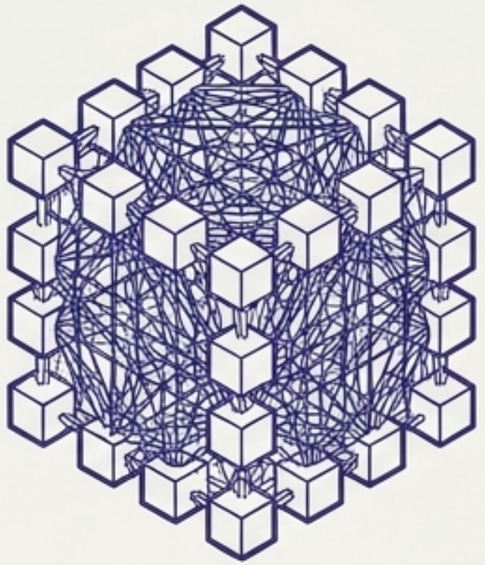
Breaking Sequential Bottlenecks: Speculative Decoding



Exploiting excess compute bandwidth during memory-bound decoding. The draft model guesses; the target model verifies. Accelerates decoding 2x-3x while mathematicaly guaranteeing lossless, exact target distribution.

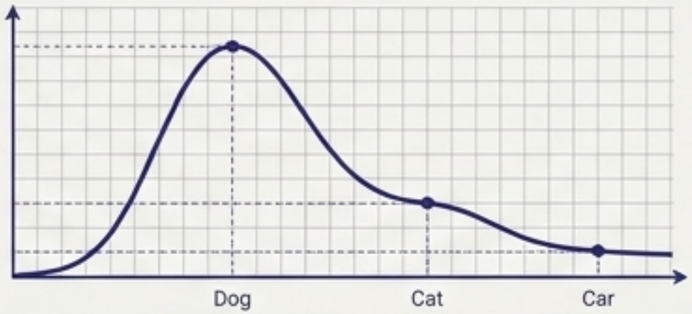
Gemini Notebook

Knowledge Distillation into Small Language Models (SLMs)



Massive Teacher Model

Softmax logits: Dog 70%, Cat 20%, Car 10%



KL-Divergence

Dog 100%

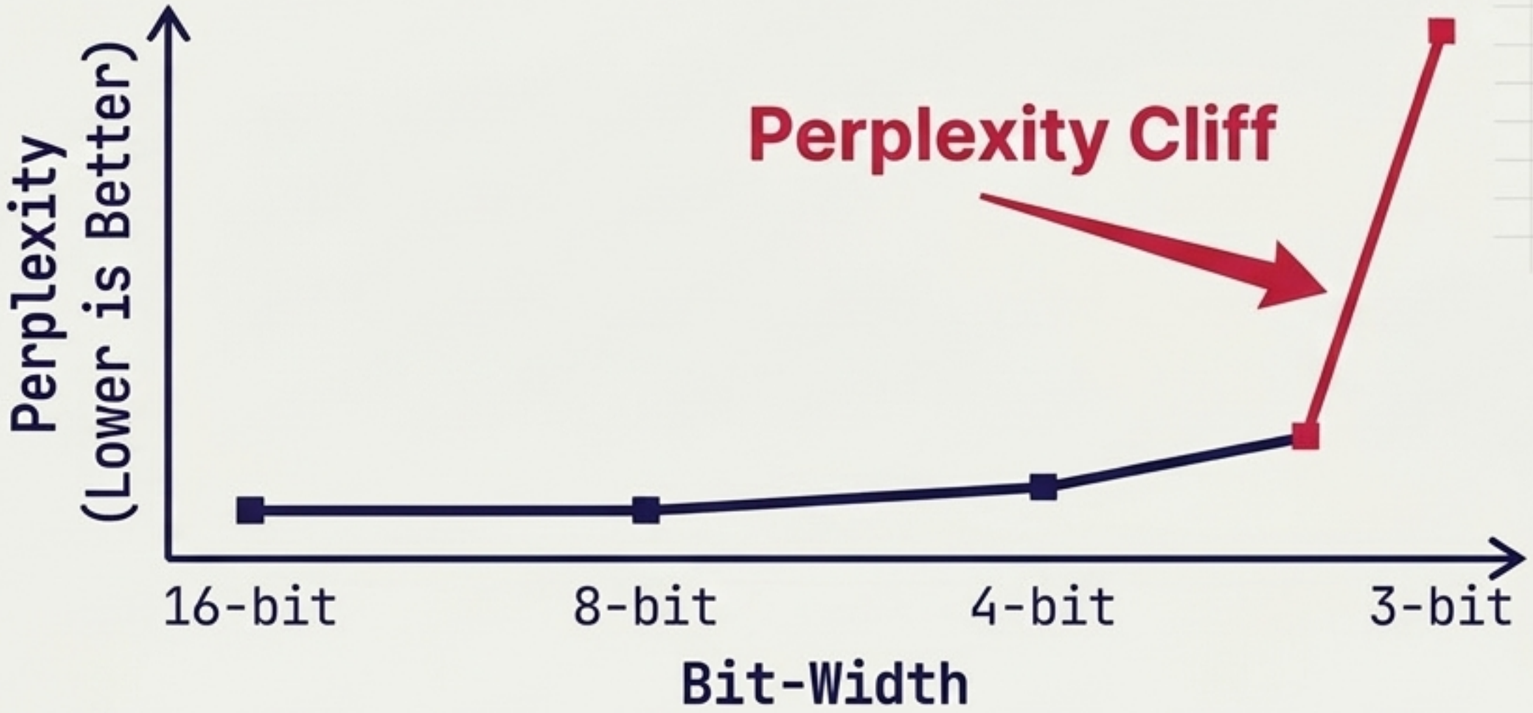


Student Model (3B)

Distillation transfers the dark knowledge—confusions, confidence, and sub-class relationships—from a massive teacher into a highly efficient 3B–8B parameter model, explicitly designed for edge hardware deployment.

Gemini Notebook

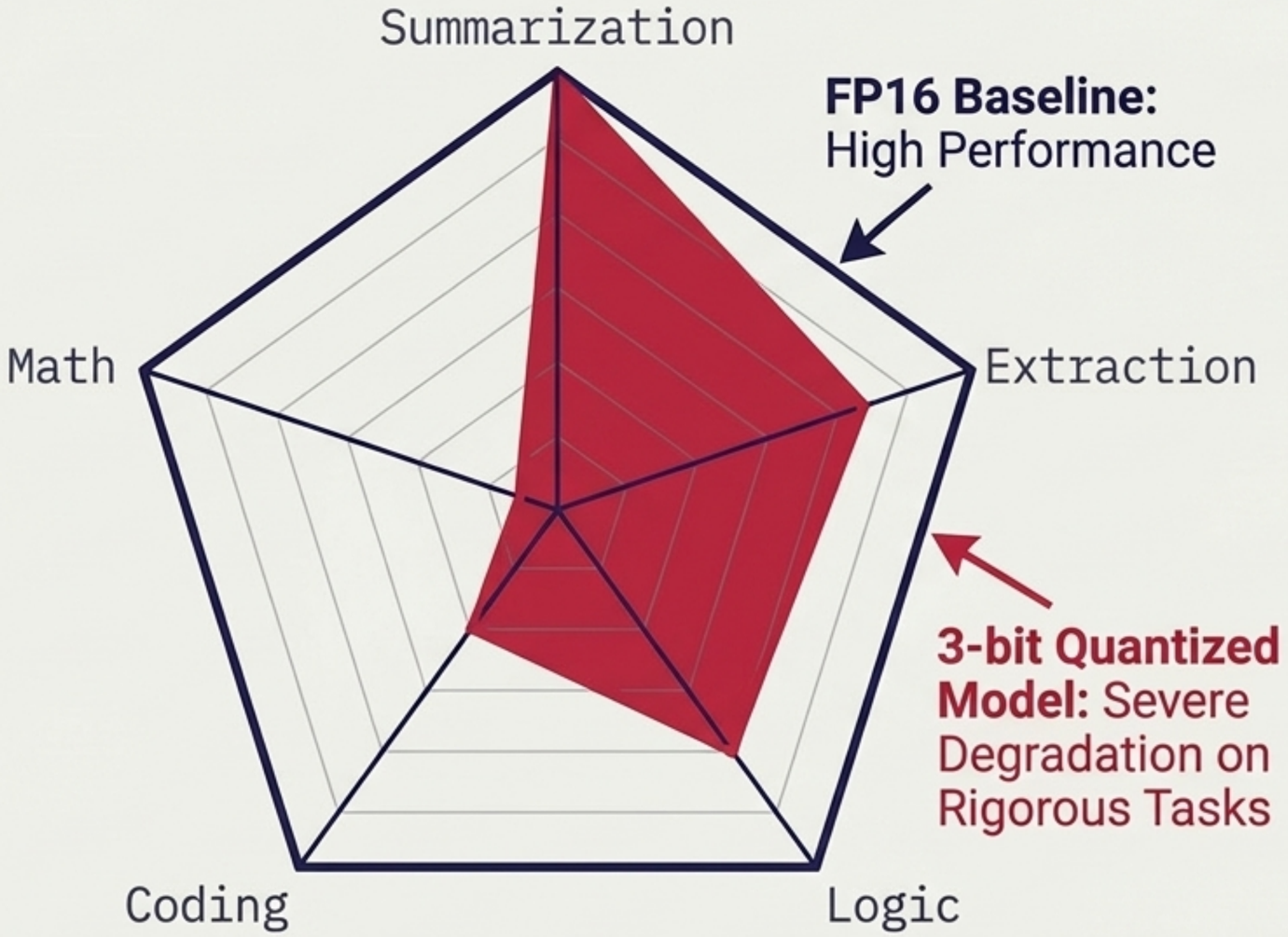
The Limits of Compression: The Perplexity Cliff



4-bit is the current Pareto-optimal frontier. Pushing hardware below 4 bits triggers catastrophic perplexity cliffs, severely degrading the model's coherent generative capability unless extreme architectural mitigations are deployed.

Gemini Notebook

Task-Specific Quantization Degradation



Language processing and summarization are highly robust to aggressive low-bit compression.

Rigorous reasoning tasks (Math, Coding) degrade significantly faster.

Deploy **precision** based strictly on the use case.

Gemini Notebook

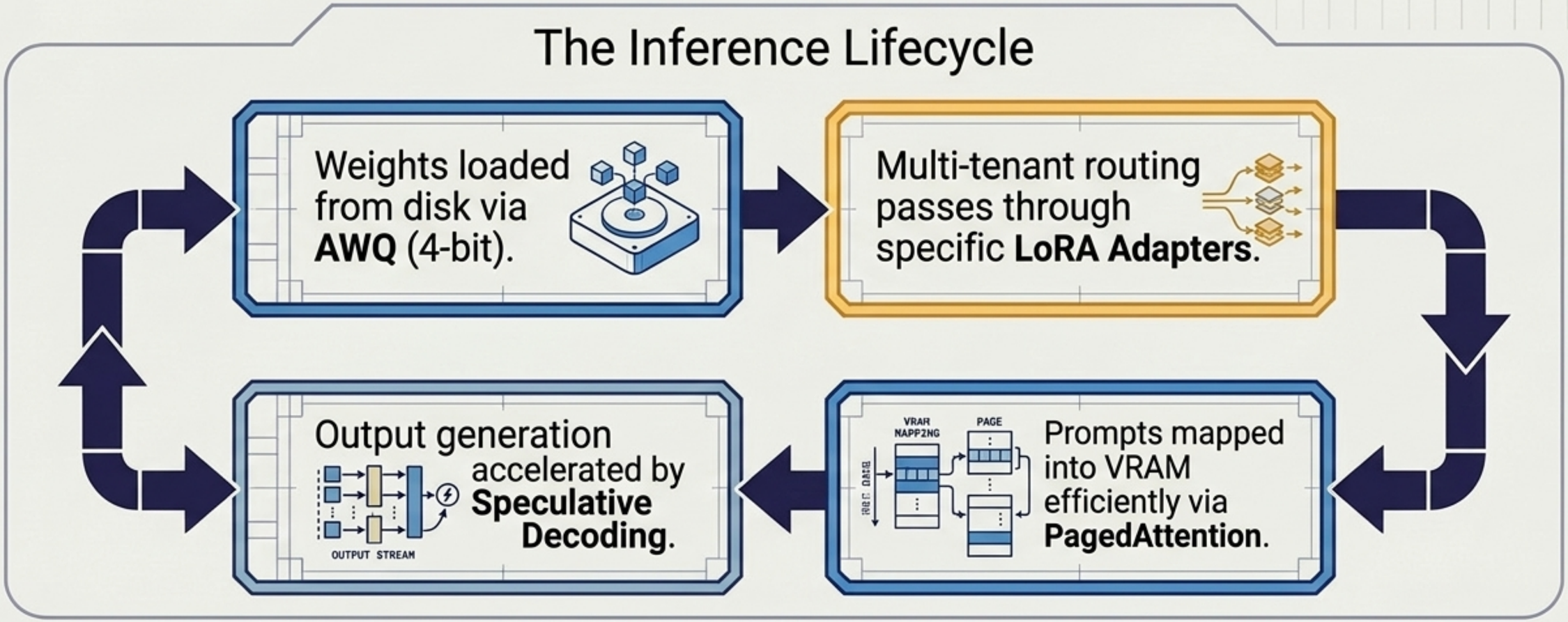
The Precision Playbook

Hardware	Strategy	Detail
Cloud H100s	Use **FP8**	(Native hardware support, no PTQ calibration overhead).
High-End Multi-GPU	Use **4-Bit AWQ / GPTQ**	(Optimal VRAM compression, high throughput).
Local Edge / Apple Silicon	Use **GGUF / Llama.cpp**	(Ideal for CPU/GPU hybrid offloading).
Fine-Tuning	Use **BF16 Compute** + **NF4 Storage**	via QLoRA stack.

Gemini Notebook

The Unified Architecture Stack

The Inference Lifecycle



Modern LLM deployment is not achieved via a single technique. It is a tightly integrated stack of bit-level arithmetic, OS-level memory mapping, and algebraic rank decomposition, dictated entirely by the physics of the memory wall.

Gemini Notebook