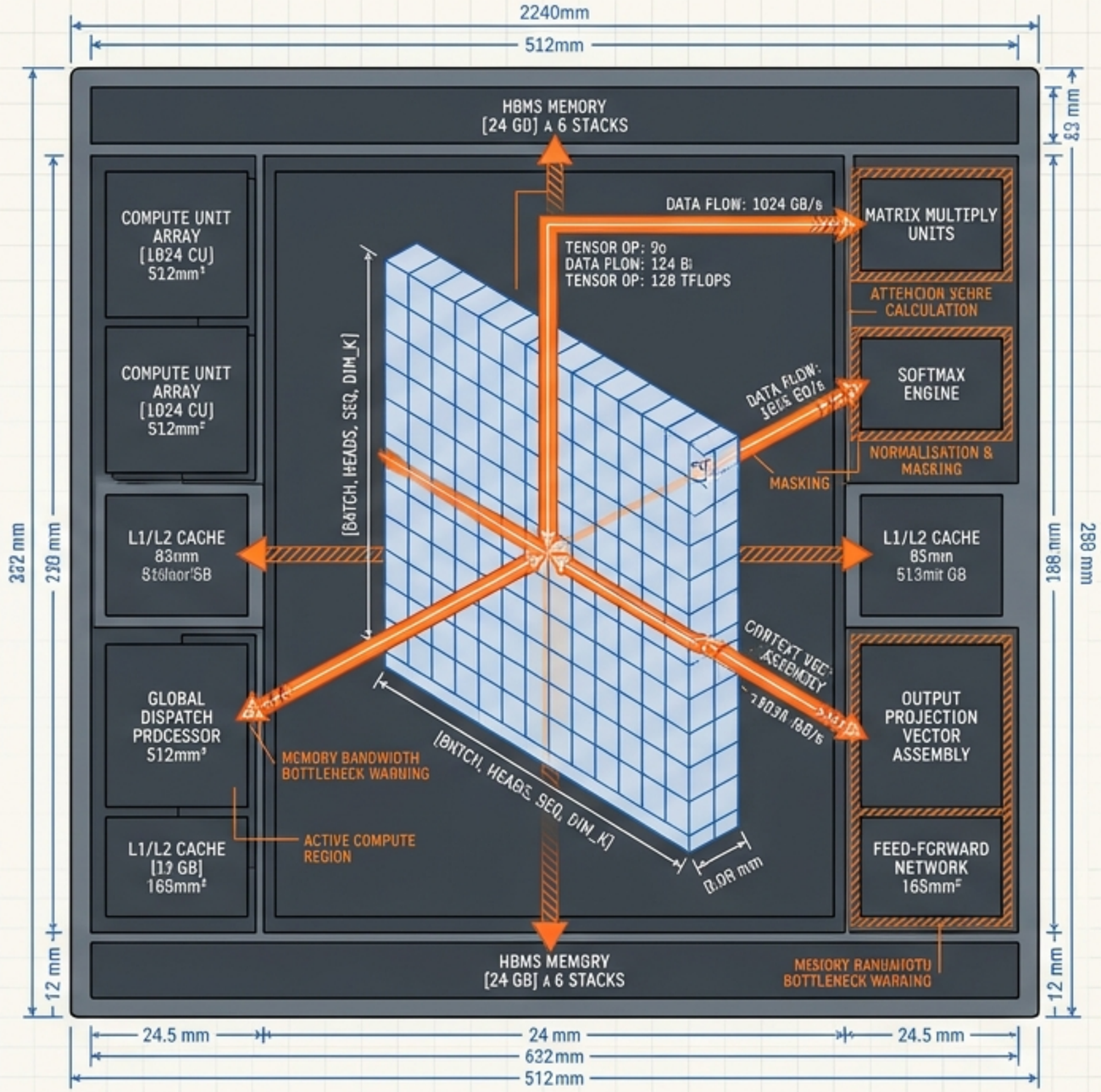
AI Basics — S01E02: Transformer Internals

A Deep-Dive into Self-Attention, Hardware Hierarchies, and Production Systems

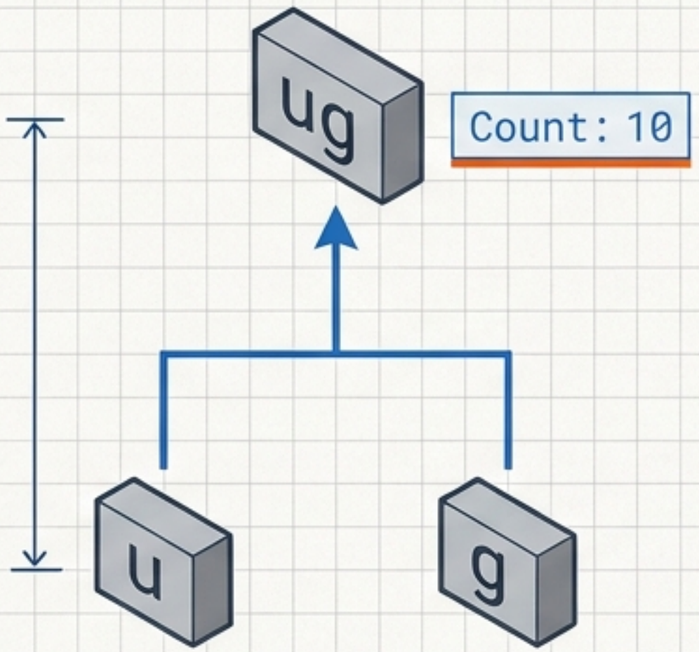


Gemini Notebook

THE SUBWORD ALGORITHM MATRIX

Byte-Pair Encoding (BPE)

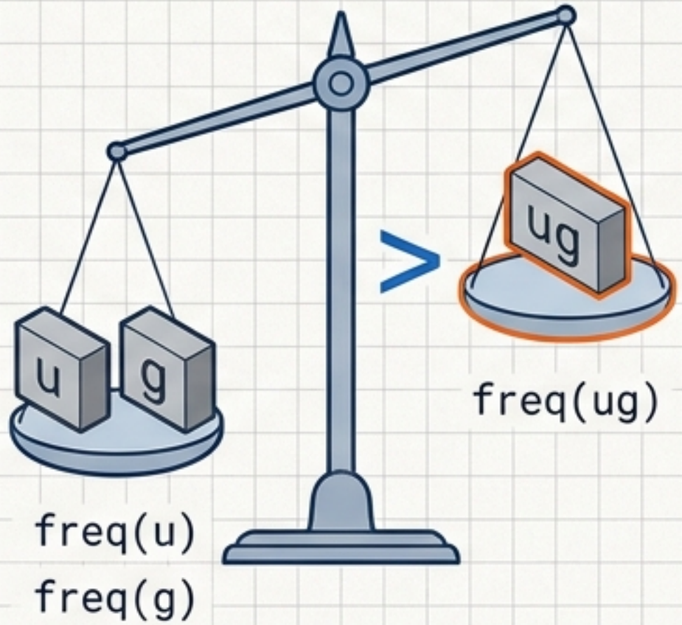
Mechanic: Frequency merges. Starts with base characters, iteratively merges the most frequent adjacent pairs.



WordPiece

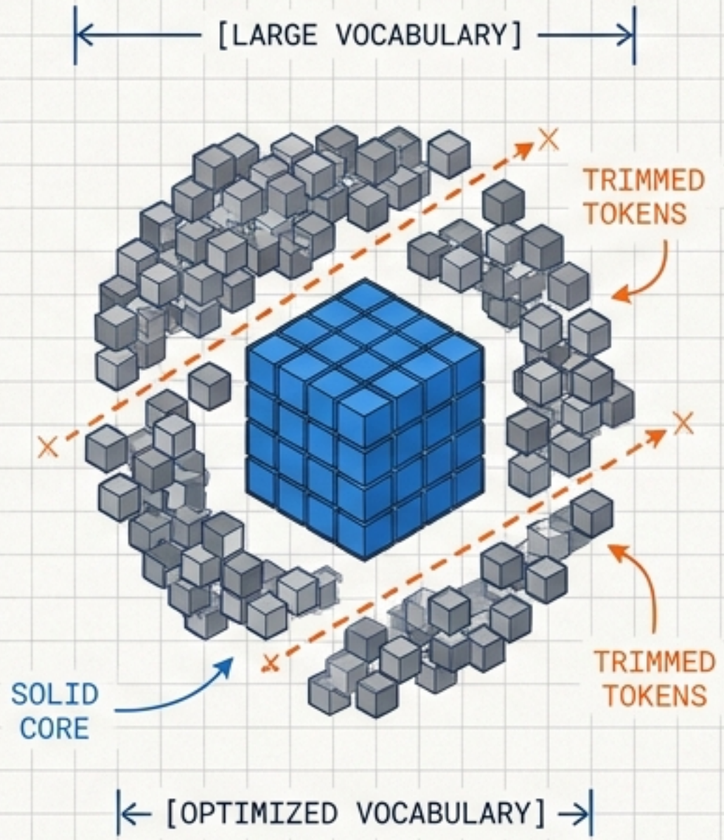
Mechanic: Likelihood maximization. Evaluates pairs by score. Favors merges where combined token appears more often than chance.

$$\text{score} = \frac{\text{freq(ug)}}{\text{freq(u)} \times \text{freq(g)}}$$



Unigram

Mechanic: Probabilistic trimming. Starts with a massive vocabulary and iteratively removes tokens that contribute least to overall likelihood.

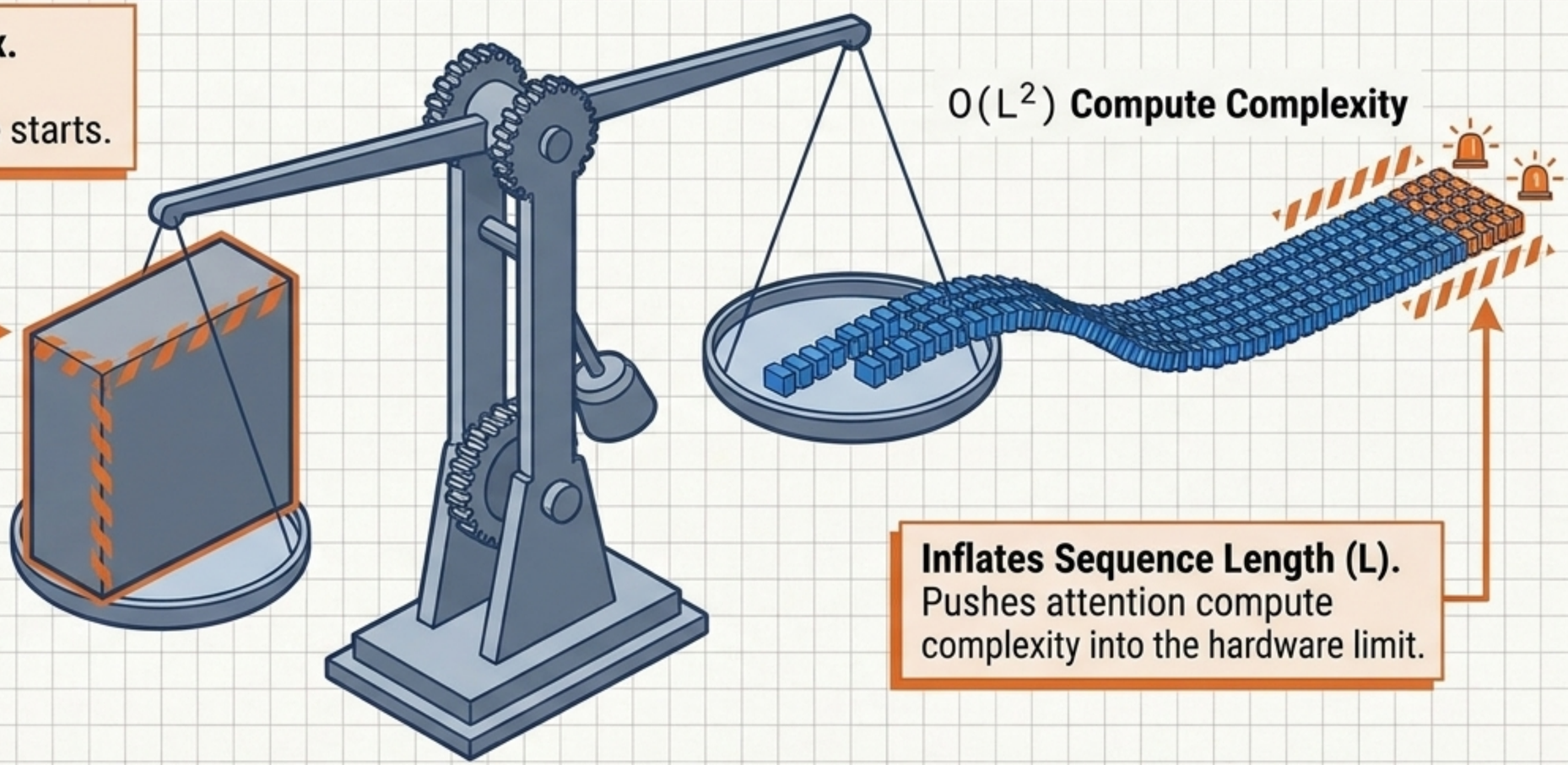


Gemini Notebook

THE VOCABULARY VRAM TRADE-OFF

Bloats the Embedding Matrix.
Consumes finite GPU VRAM immediately before inference starts.

$V \times d_{model}$



$O(L^2)$ **Compute Complexity**

Inflates Sequence Length (L).
Pushes attention compute complexity into the hardware limit.

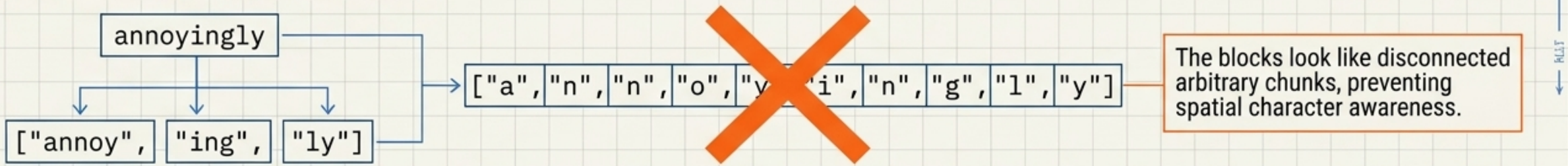
Synthesis: The VRAM **vs. FLOPS Equilibrium:** A large vocabulary trades sequence length compute limits for immediate embedding matrix VRAM consumption.

Gemini Notebook

PRODUCTION REALITY: TOKENIZATION FAILURES

CHARACTER ARITHMETIC

Why LLMs fail at exact math and spelling.



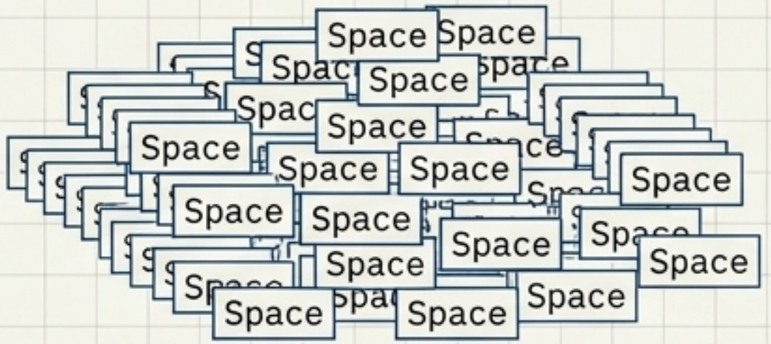
The blocks look like disconnected arbitrary chunks, preventing spatial character awareness.

WHITESPACE SENSITIVITY

Code generation indentation bloat.

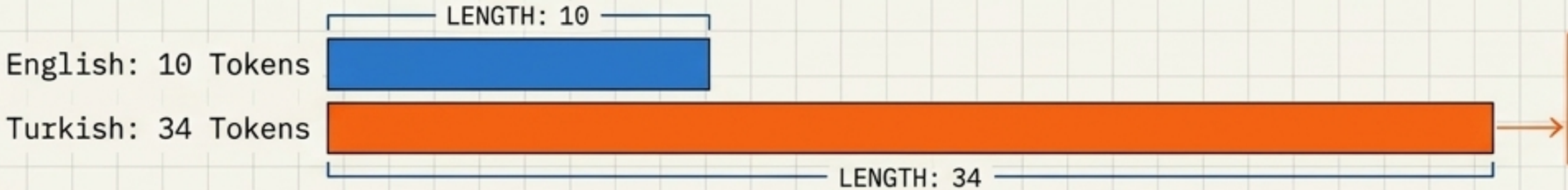
```
def example_function():
    if condition:
        do_something()
    else:
        do_something_else()
```

Leading spaces collapse irregularly or inflate into dozens of tokens, bloating context and confusing structural logic.



NON-ENGLISH FERTILITY PENALTIES

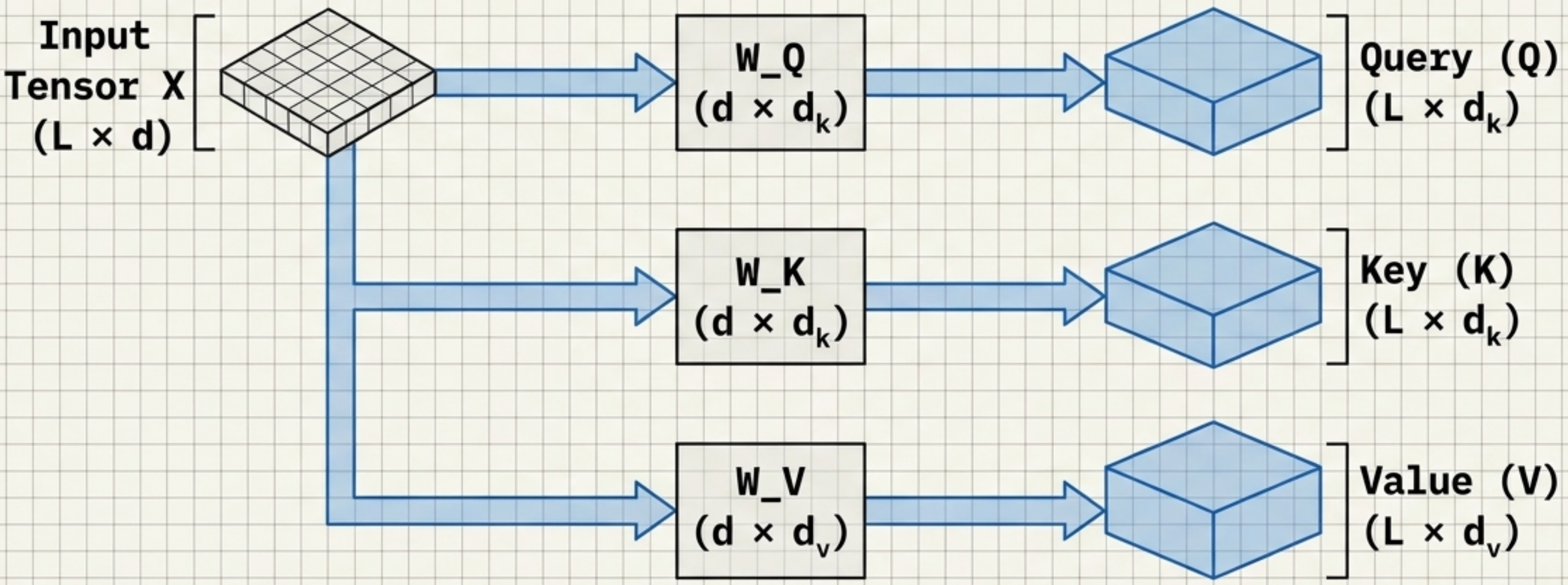
Suboptimal subword merges.



Non-English languages burn through context windows 2x–4x faster due to fragmented subword merges.

Gemini Notebook

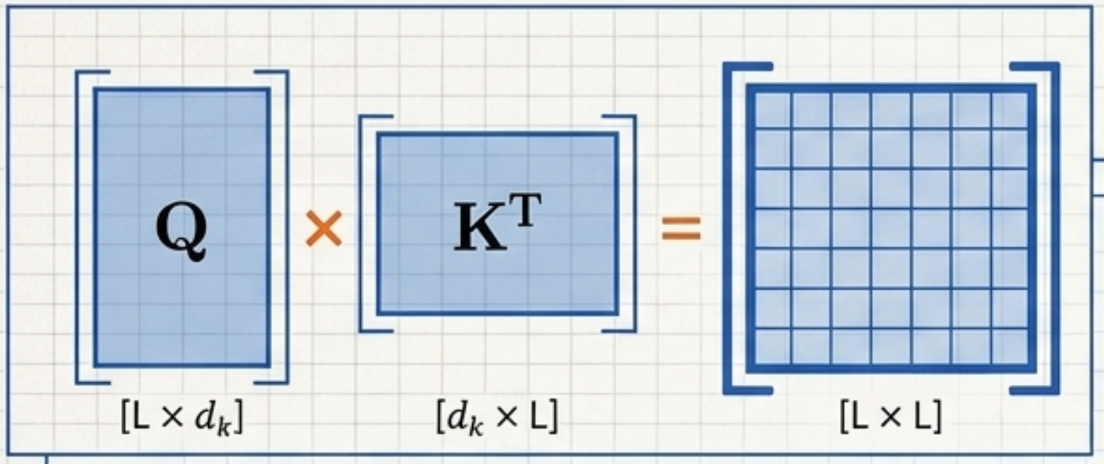
Exact Q, K, V Linear Projections



The Filing Cabinet Metaphor:
Query (Q) is the sticky note topic you are searching for.
Key (K) is the folder label. Value (V) is the actual document inside.

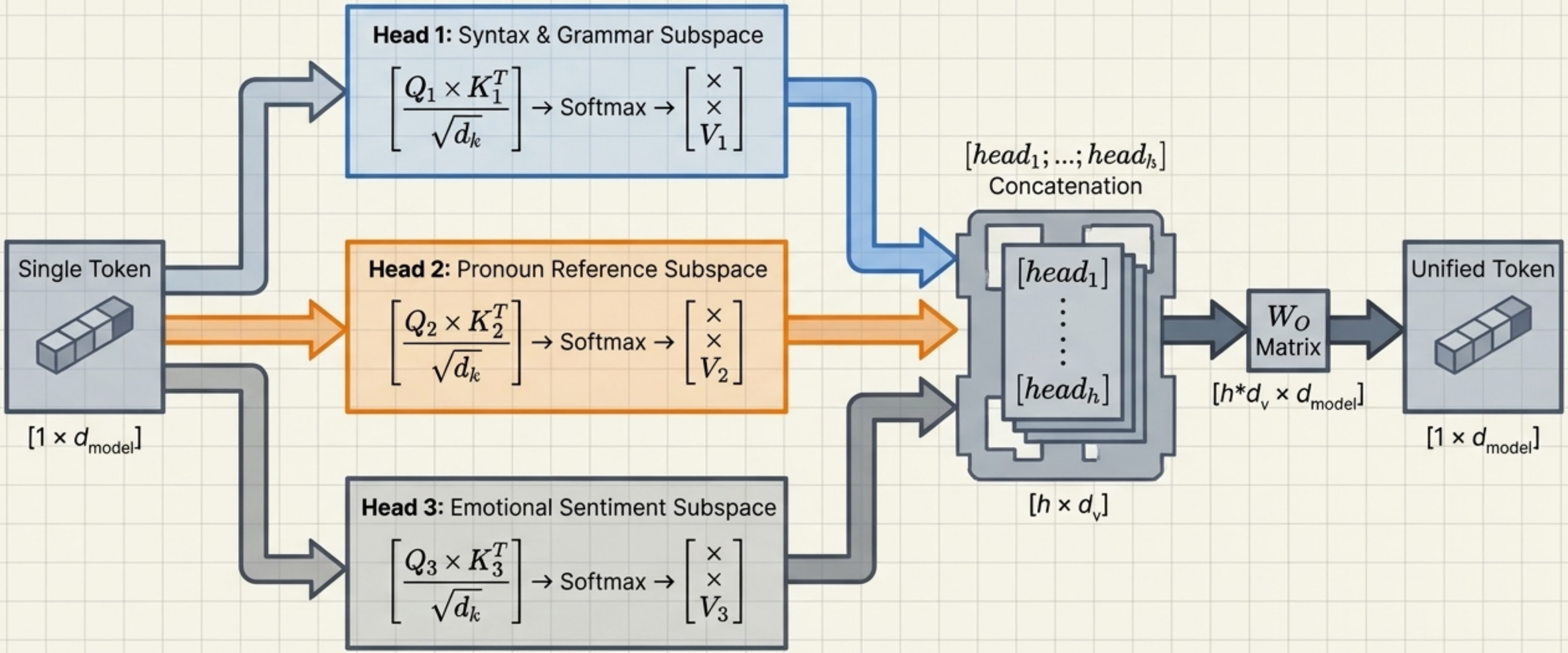
Gemini Notebook

Softmax Saturation & the $\sqrt{d_k}$ Scaler



Gemini Notebook

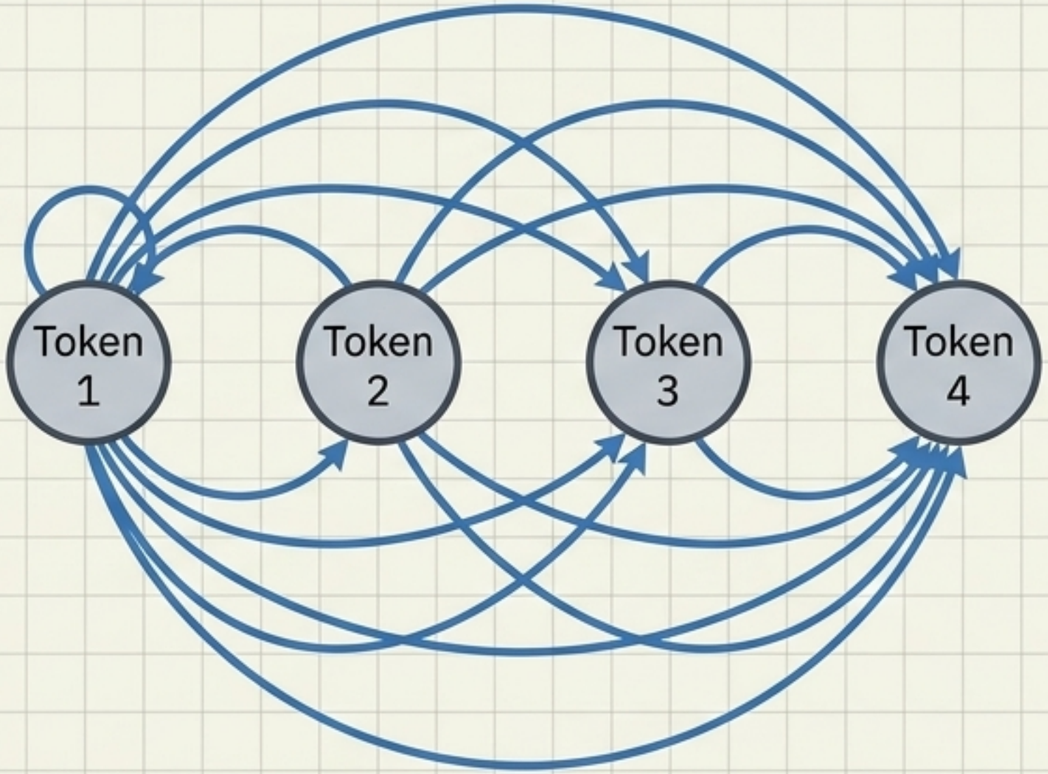
Multi-Head Subspaces



Gemini Notebook

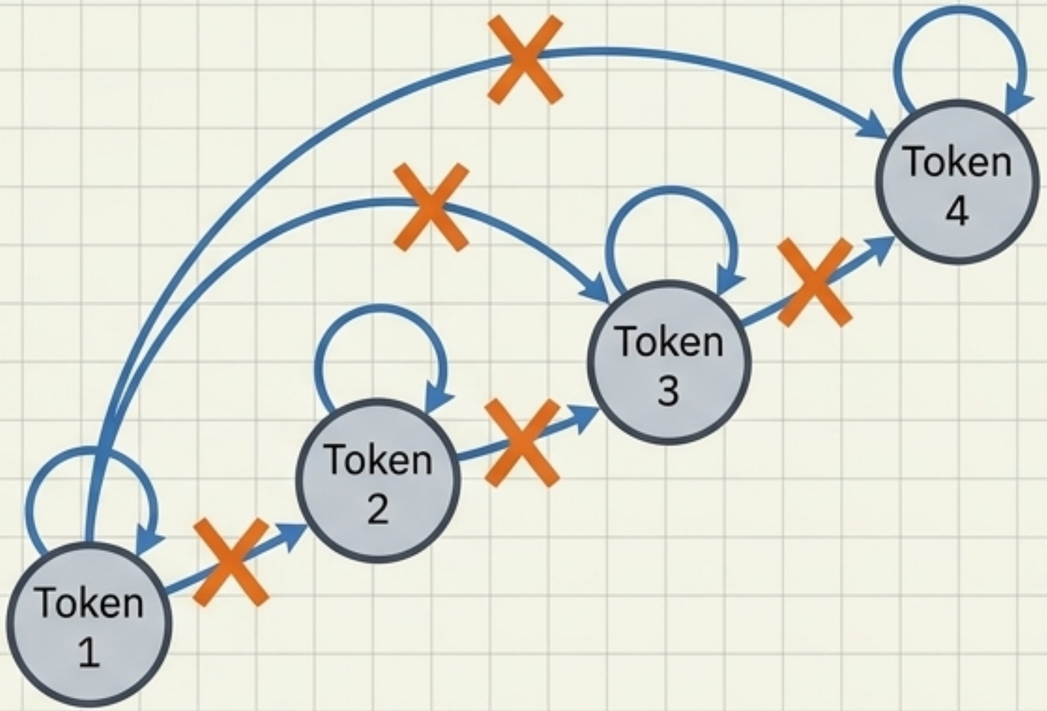
Bidirectional vs. Autoregressive Dataflow

Encoder / Bidirectional Dataflow



Full-context representation. Information flows seamlessly left-to-right and right-to-left.

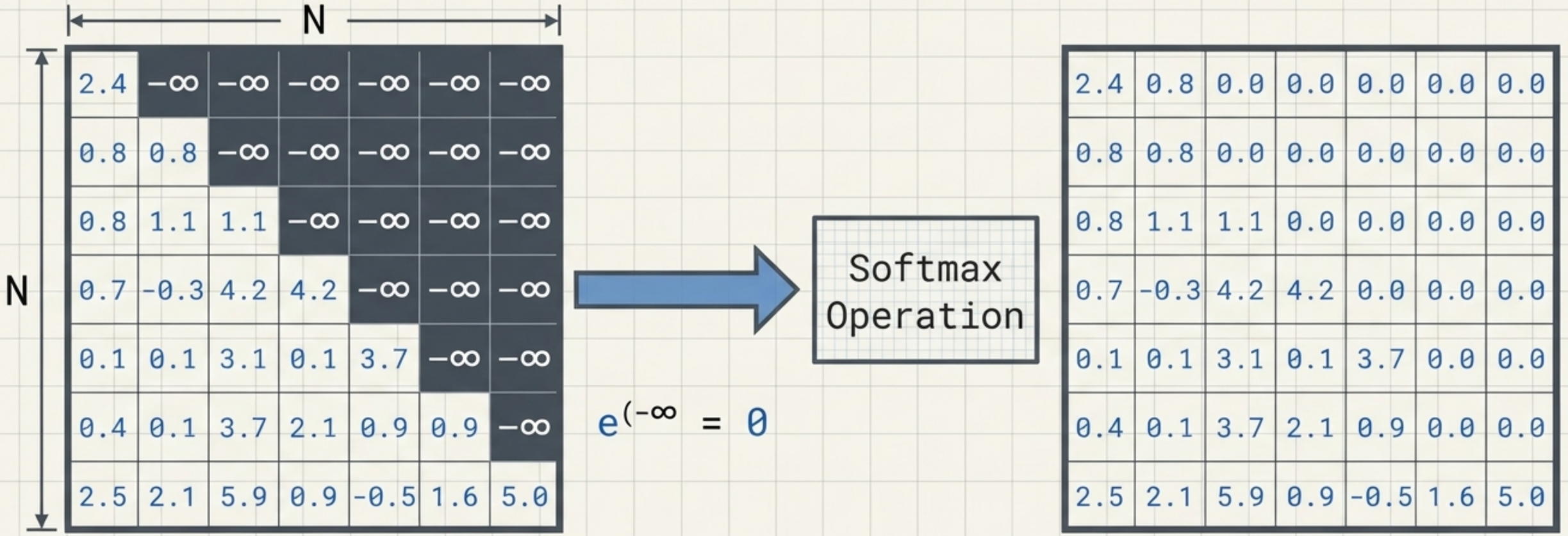
Decoder / Autoregressive Dataflow



Strictly past-context attention for sequence generation.

Gemini Notebook

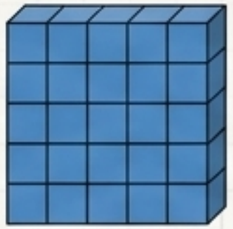
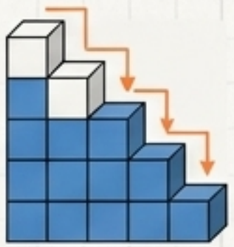
Upper-Triangular Causal Masking



The Causal Mask: Forcing the model to predict the **future** without physically being able to look at it during parallel training.

Gemini Notebook

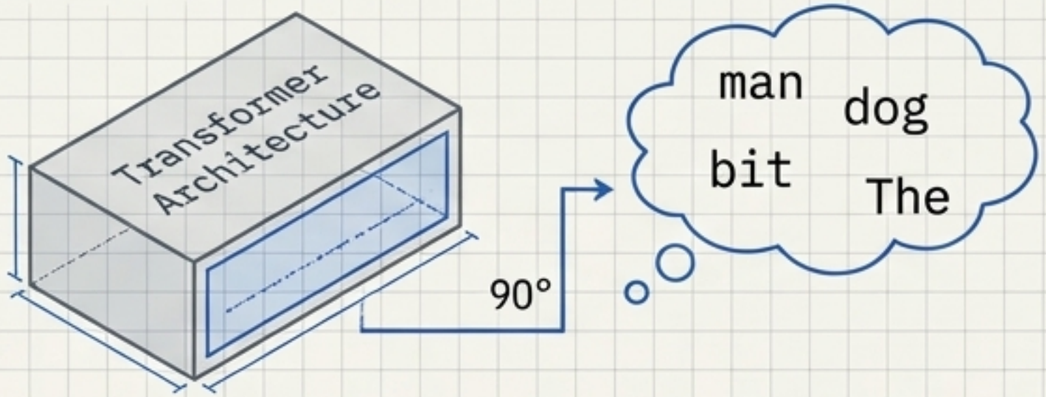
Topological Selection Guide

BERT-class (Encoders)			
Architecture: Bidirectional, unmasked.	**Task Types:** Sequence Classification, Named Entity Recognition, Extractive QA, Semantic Search / Reranking	**When to Deploy:** When you have the entire sequence upfront and need deep, contextual understanding of the whole text simultaneously	
GPT-class (Decoders)			
Architecture: Autoregressive, upper-triangular masked.	**Task Types:** Text Generation, Chatbots, Code Completion, Agentic Reasoning	**When to Deploy:** When the output is open-ended and must be generated iteratively, token-by-token in a forward sequence.	

Gemini Notebook

The Permutation Invariance Problem

The Problem

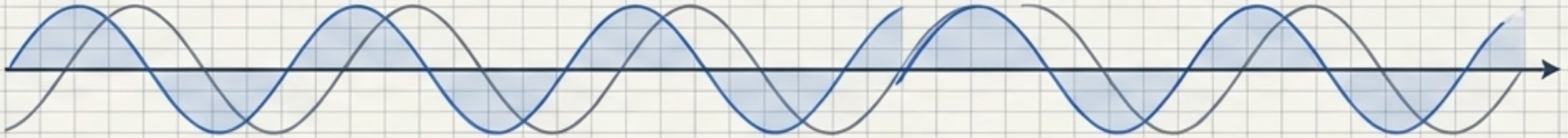


Without position IDs, **“The dog bit the man”** and **“The man bit the dog”** compute identically.

Attention is a set operation, not a sequential one.

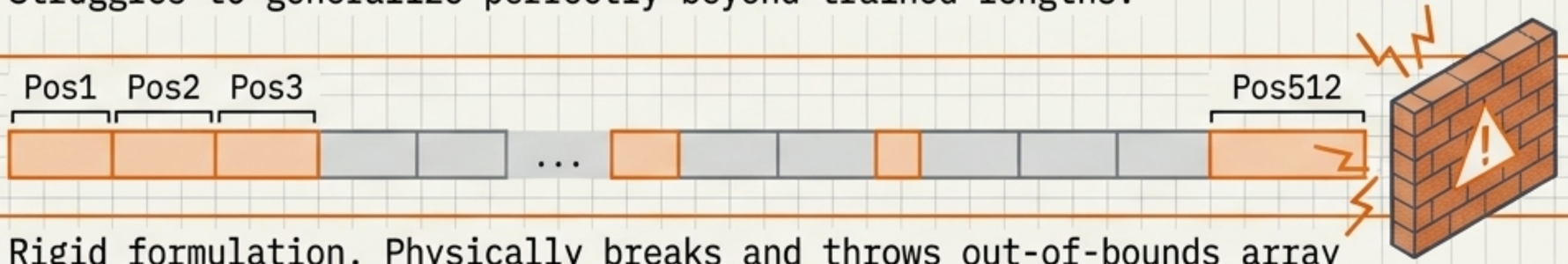
Legacy Solutions & Limits

Fixed Sinusoidal (Vanilla)



Struggles to generalize perfectly beyond trained lengths.

Learned Absolute (BERT)



Rigid formulation. Physically breaks and throws out-of-bounds array errors on sequences longer than the exact training window.

Gemini Notebook

Engineering Specification Manual

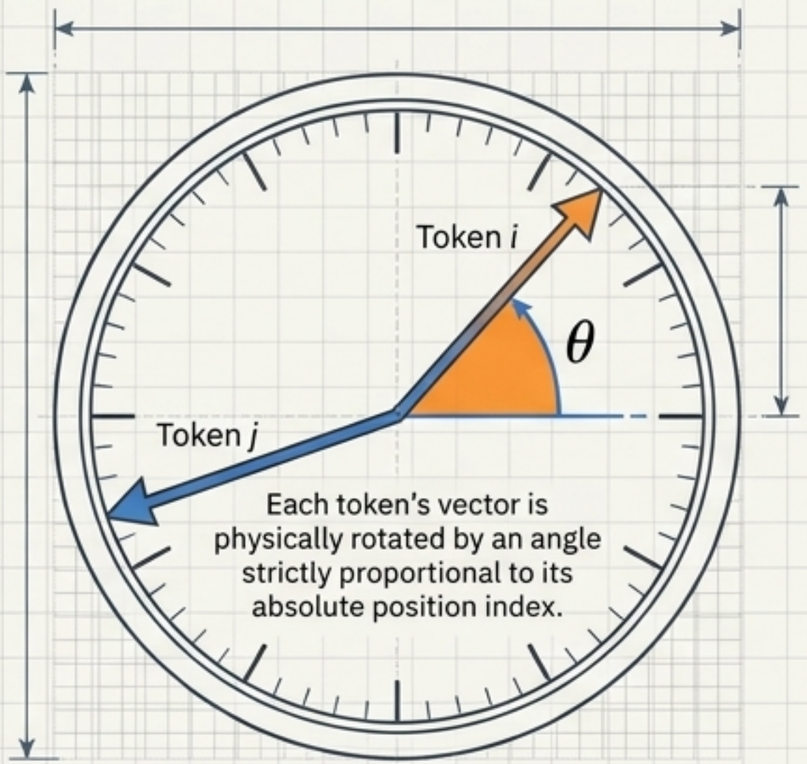
RoPE (Rotary Position Embedding) Mechanics

Math Space

$$R_{\Theta,m}^{d} W_q x_m$$

Instead of adding position as a static vector, RoPE multiplies by a rotation matrix in continuous space.

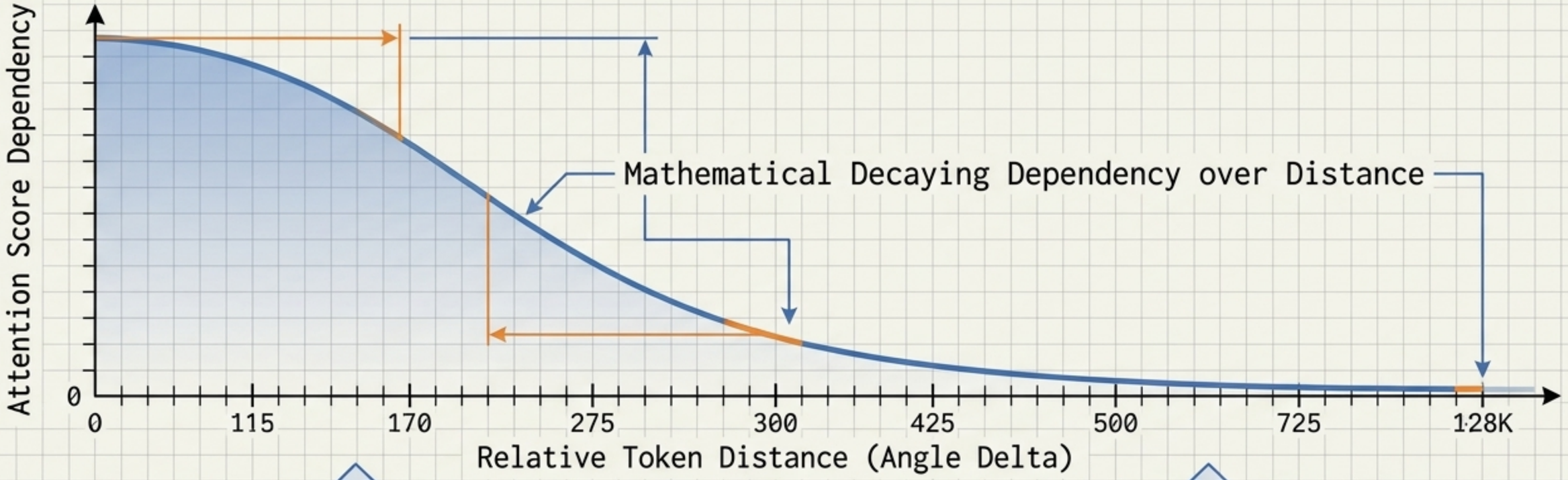
The Rotary Dial



The dot product between Query and Key natively relies only on their **relative angle** (distance).

Gemini Notebook

Scaling to Long Contexts with RoPE



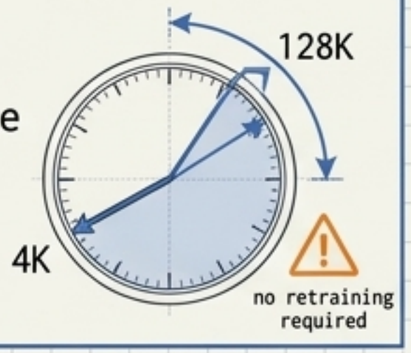
Zero-Shot Context Extrapolation

Models trained on 4K contexts can physically accept longer inputs without throwing out-of-bounds hardware errors because continuous angles do not have a hard array limit.



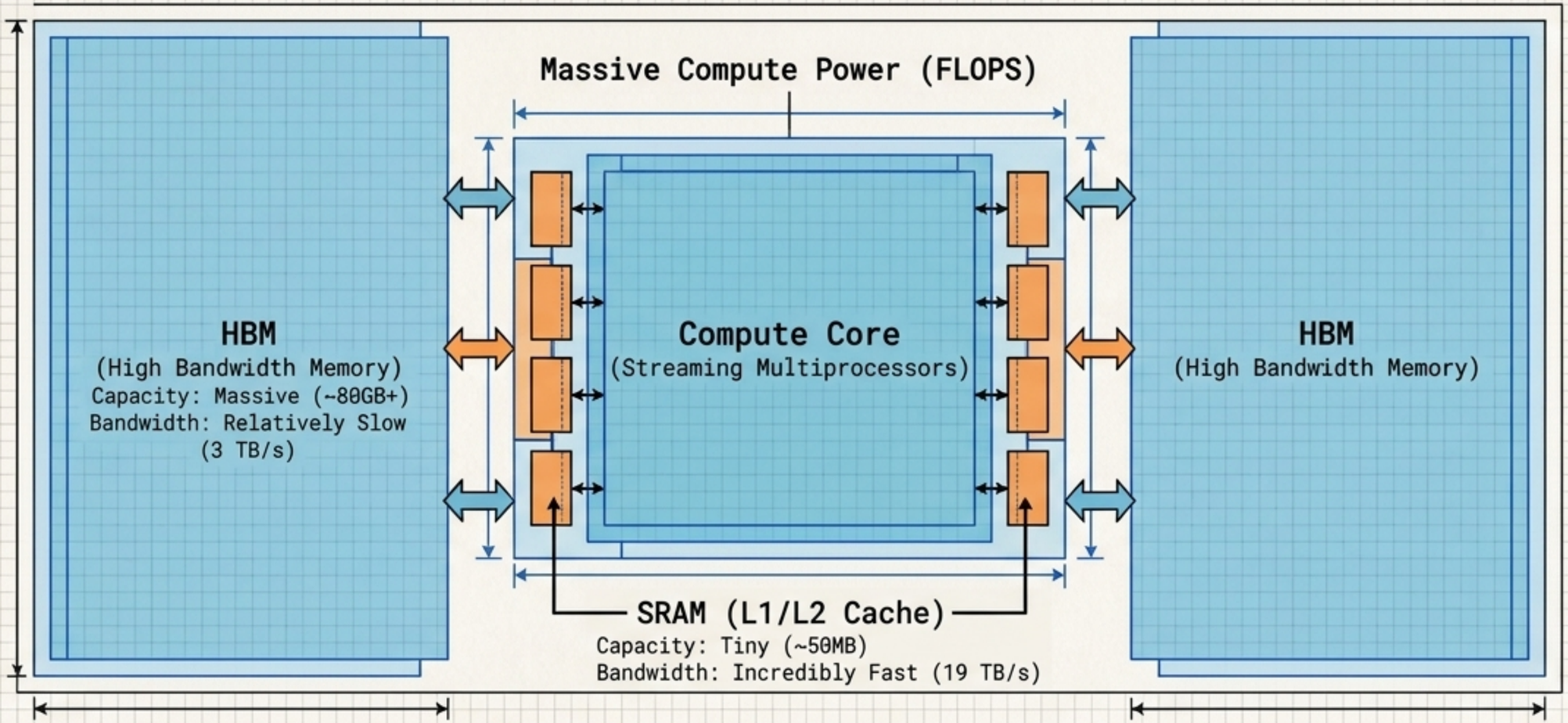
Dynamic RoPE Scaling (YaRN)

The rotational base can be algorithmically scaled at inference time, stretching a 4K trained model to perfectly process a 128K context window in production without retraining base weights.



Gemini Notebook

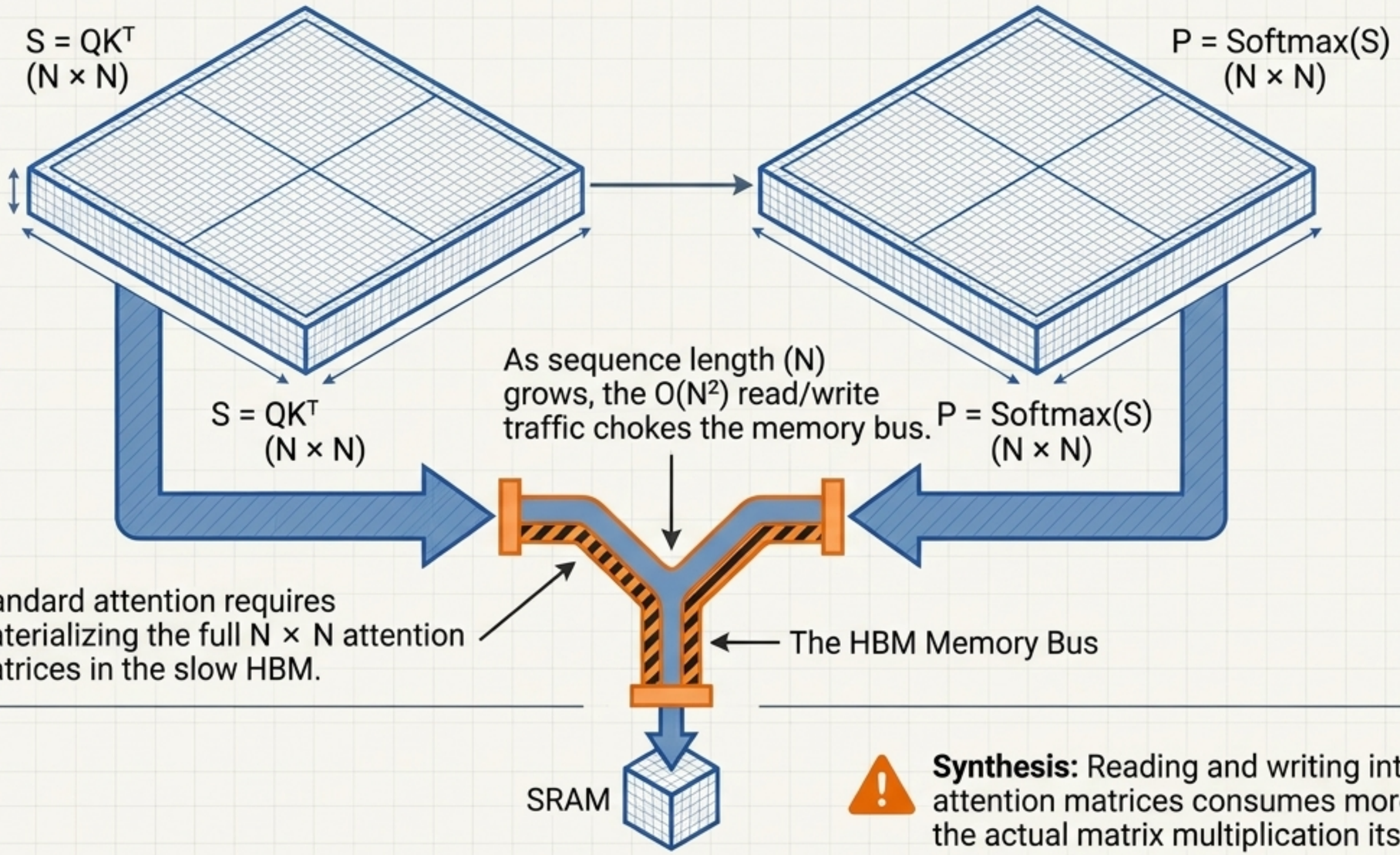
The Hardware Reality: FLOPS vs. Memory Bandwidth



The Silicon Reality: Modern AI is rarely bound by how fast the GPU can do math (FLOPS). It is bound by how fast we can physically move data across the silicon from HBM to SRAM.

Gemini Notebook

The $O(N^2)$ HBM Bottleneck

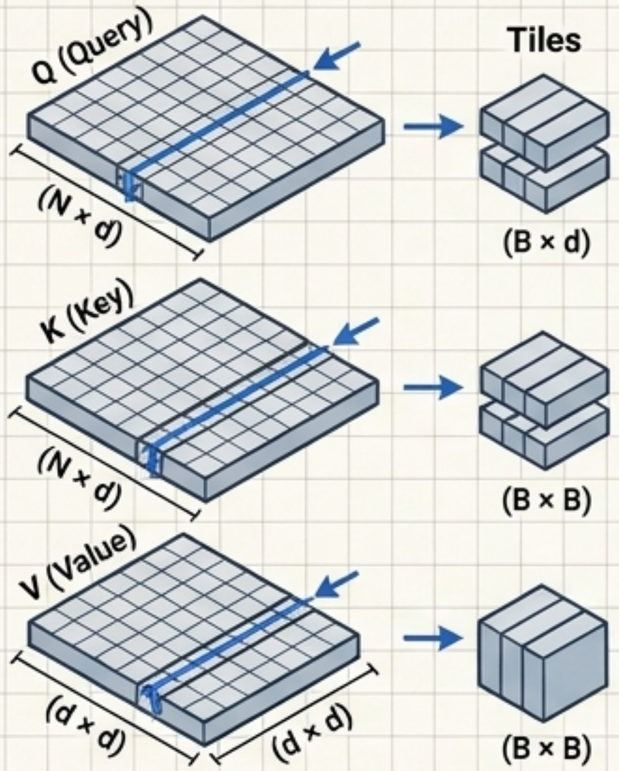


Synthesis: Reading and writing intermediate attention matrices consumes more time than the actual matrix multiplication itself.

Gemini Notebook

FlashAttention: SRAM Tiling & Online Softmax

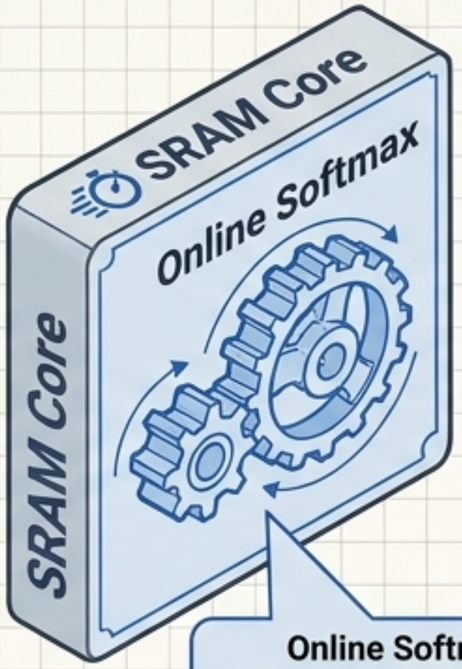
Step 1: Split into Tiles



Q, K, V matrices are segmented into small, fixed-size tiles.

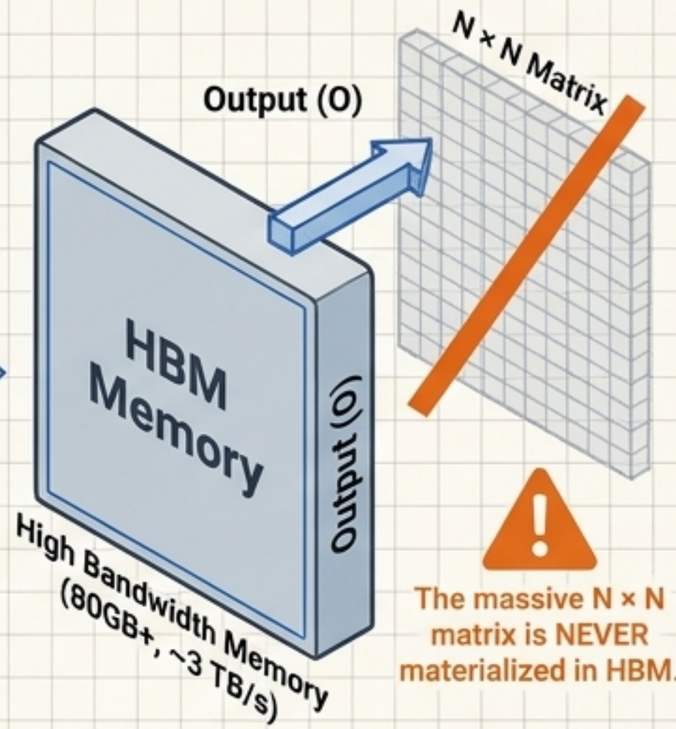
Step 2: Direct Load to SRAM

Only a single set of small blocks is loaded at a time.



Online Softmax:
Computes exact softmax incrementally block-by-block without needing the whole row.

Step 3: Write Output Once



The massive N × N matrix is NEVER materialized in HBM.

Result written back to HBM in a single pass.

By fusing operations in SRAM, FlashAttention turns an IO-bound memory nightmare into a compute-bound, lightning-fast operation.

Gemini Notebook

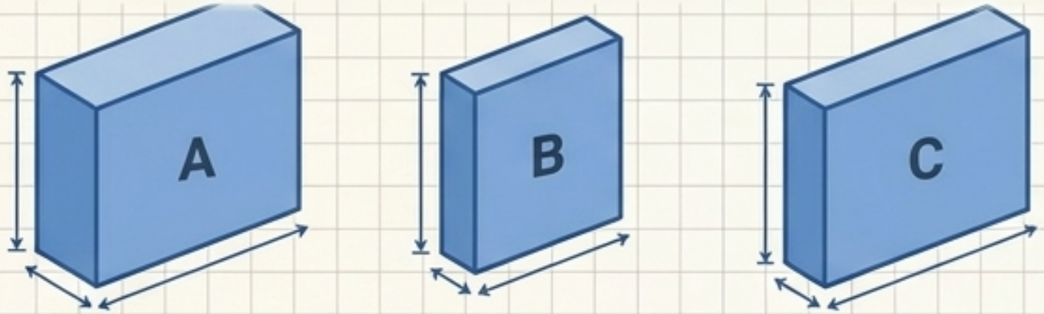
Prefill vs. Decode Dataflow

Prefill Phase

Action: Processing the user's entire prompt at once in a massive parallel batch.

Math:

$$[A]_{batch \times seq_len} \times [B]_{seq_len \times hidden_dim} = [C]_{batch \times hidden_dim}$$



Hardware Profile:



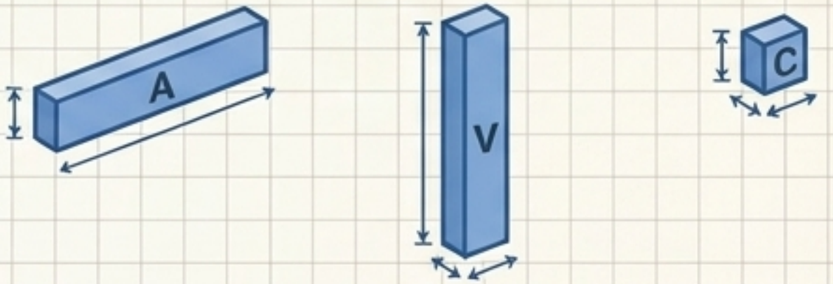
GPU math cores are efficiently churning through FLOPS at near maximum utilization.

Decode Phase

Action: Generating the response auto-regressively, token-by-token.

Math:

$$[A]_{1 \times hidden_dim} \times [V]_{hidden_dim \times 1} = [C]_{1 \times 1}$$



Hardware Profile:



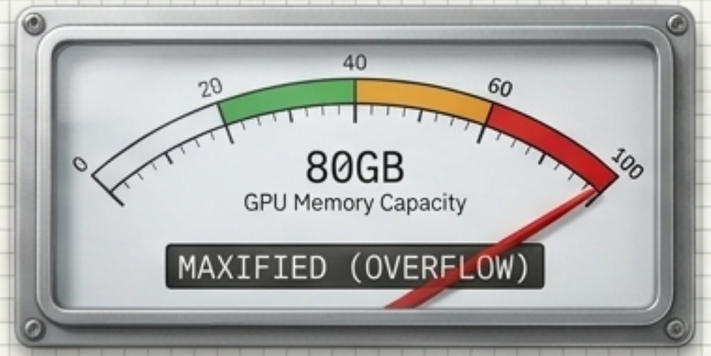
GPU math cores sit idle waiting for weights and KV cache blocks to load from slow HBM for every single generated token.

Gemini Notebook

Exact KV Cache VRAM Math

```
KV_Cache_Size = 2 * batch_size * seq_len * num_layers * num_heads * head_dim * bytes_per_param
```

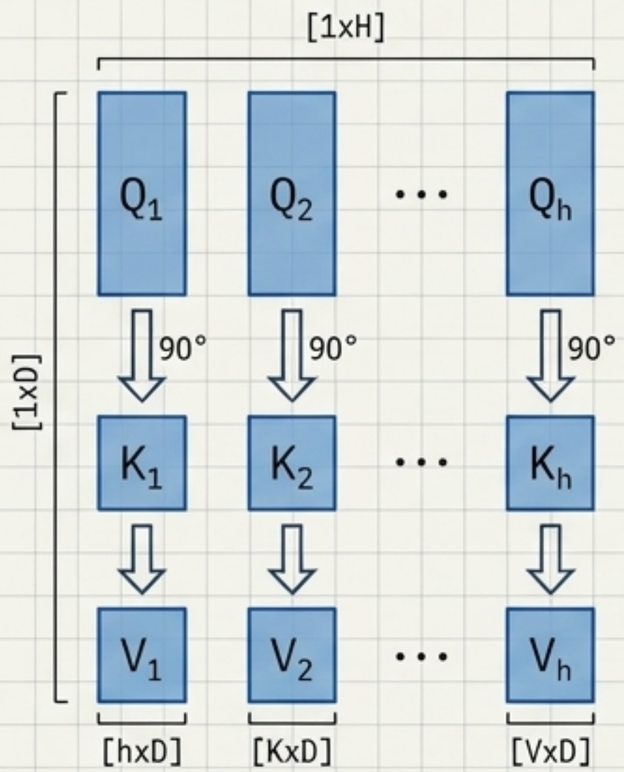
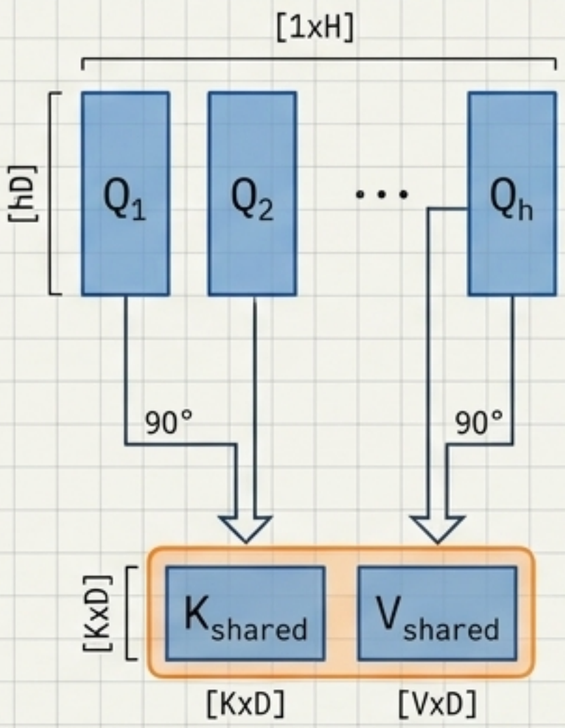
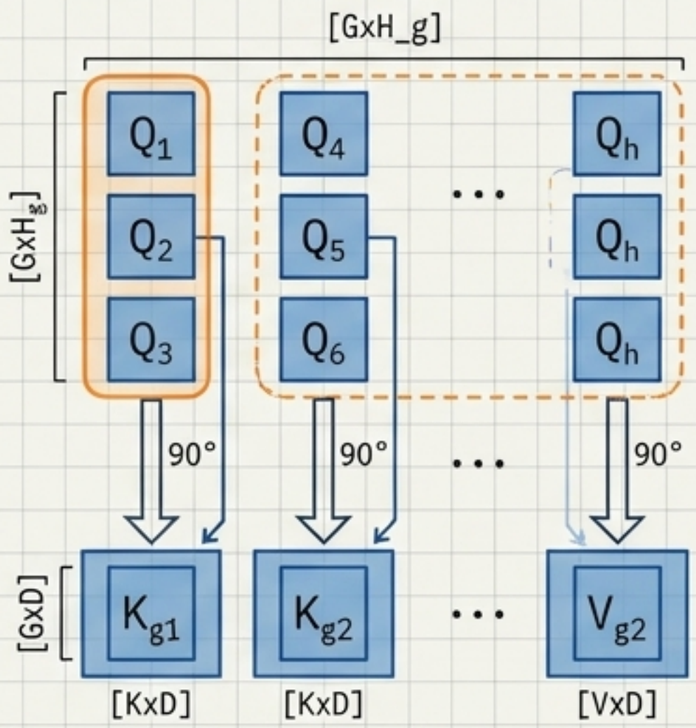
- **2**: Multiplier for both the Key and Value matrices.
- **batch_size**: The silent killer. Grows linearly with every single generated token in the output.
- **seq_len**: Multiplies the entire footprint drastically for concurrent users.



Diagnostic: A 100k context window for a 70B model exceeds an 80GB GPU's total capacity just in cache, before a single parameter weight is even loaded.

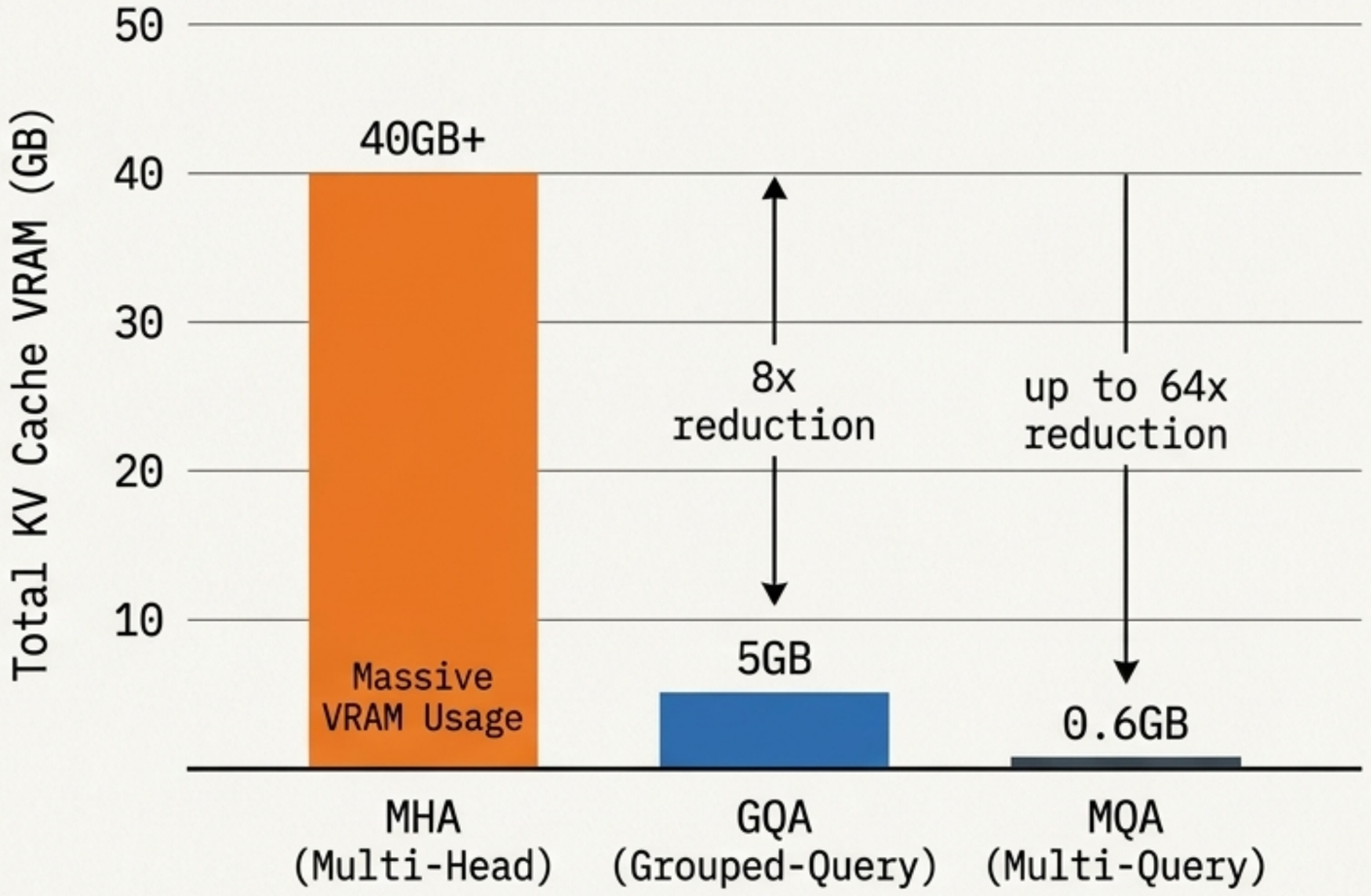
Gemini Notebook

Architectural Interventions: MQA & GQA

Multi-Head Attention (MHA)	Multi-Query Attention (MQA)	Grouped-Query Attention (GQA)
		
Status text	Status text	Status text
Maximum memory footprint.	Extreme memory savings, but potential quality degradation on highly complex mathematical/reasoning tasks.	The optimal production compromise (e.g., Llama). Slashes KV footprint massively with near-zero quality loss.

Gemini Notebook

Slashing Memory Footprints



Synthesis

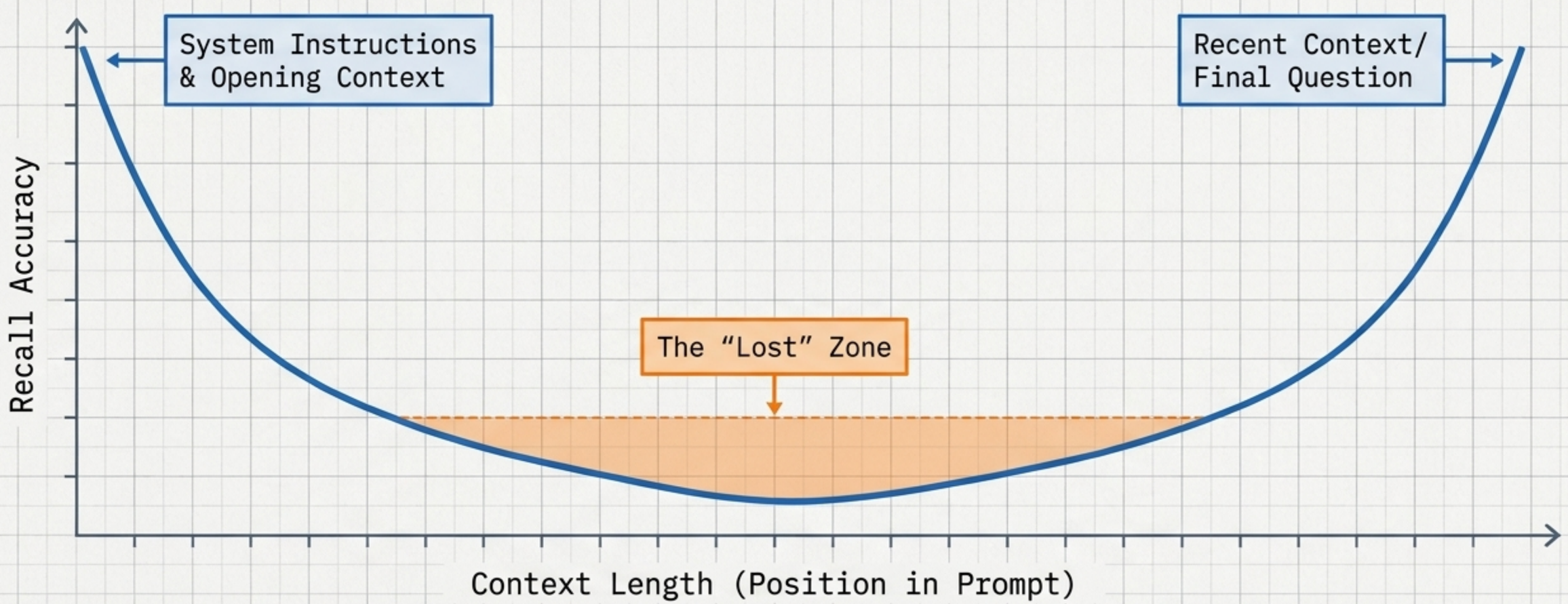
Production Impact:

By slashing the memory footprint, GQA allows the production server to physically load vastly larger concurrent batch sizes.

Because the Decode phase is strictly memory-bandwidth-bound, batching users together is the only engineering path to achieving high throughput and justifying hardware costs.

Gemini Notebook

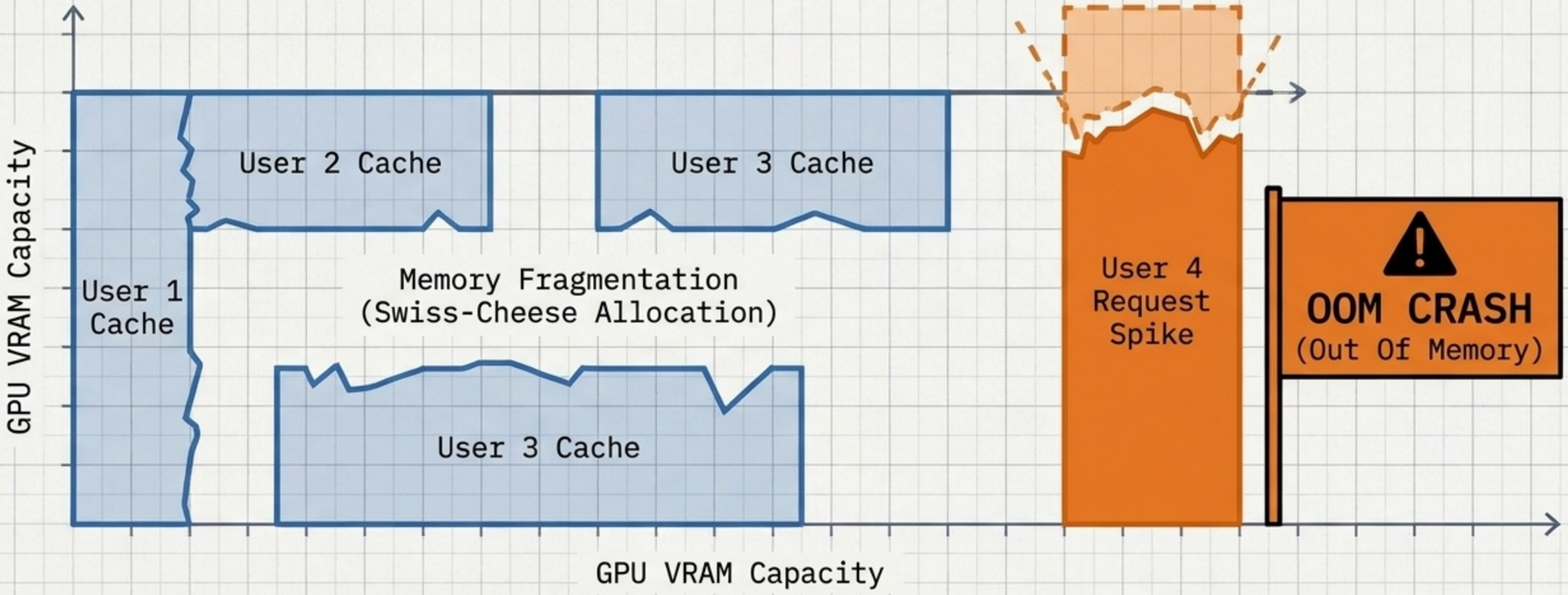
Context Degradation: “Lost in the Middle”



Even with RoPE and functionally infinite KV cache capacity, attention dilutes statistically. LLMs perfectly recall the exact start and end of a long document, while frequently dropping or hallucinating facts buried in the middle.

Gemini Notebook

Concurrent Batching & OOM Crashes



Without advanced paging (like PagedAttention), variable-length generations fragment memory memory rapidly. A sudden spike in a single user's request length hits the hard ceiling, crashing the entire GPU and dropping all concurrent users.

Gemini Notebook

The Hardware Sizing Matrix

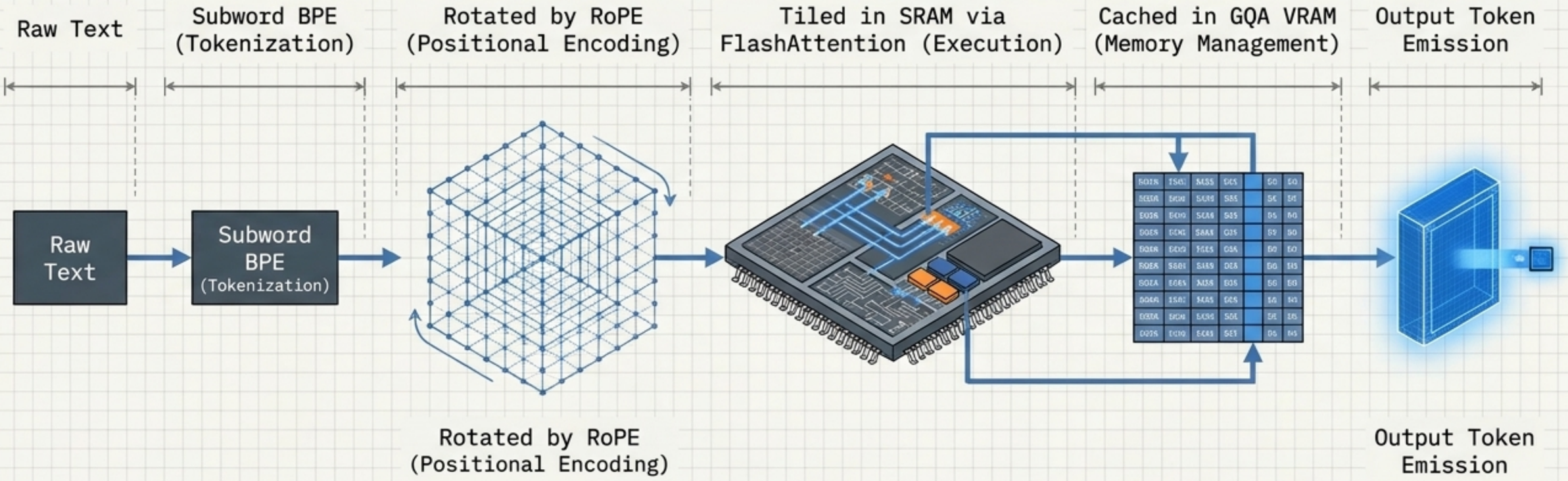
Model Parameters	Precision / Quantization	Target Context Length	Concurrent Batch Size	Required GPU Setup
8B	FP16	8K	1	1x 24GB (A10G)
8B	FP16	32K	16	2x 80GB (A100)
8B	INT4	128K	64	8x 80GB (H100)
70B	FP16	8K	1	2x 80GB (A100)
70B	INT4	32K	16	4x 80GB (A100)
70B	INT4	128K	1	8x 80GB (H100)
400B	INT4	8K	1	8x 80GB (H100)
400B	INT4	32K	1	16x 80GB (H100)
400B	FP16	8K	1	16x 80GB (A100)
400B	INT4	128K	16	Distributed Cluster (TBD)



Note: Bandwidth determines Tokens-Per-Second latency; VRAM capacity strictly dictates maximum concurrent batch sizing and sequence length limits.

Gemini Notebook

Conclusion & System Synthesis: The Full Stack



From Byte to Silicon:
The complete lifecycle of an AI token.

Gemini Notebook