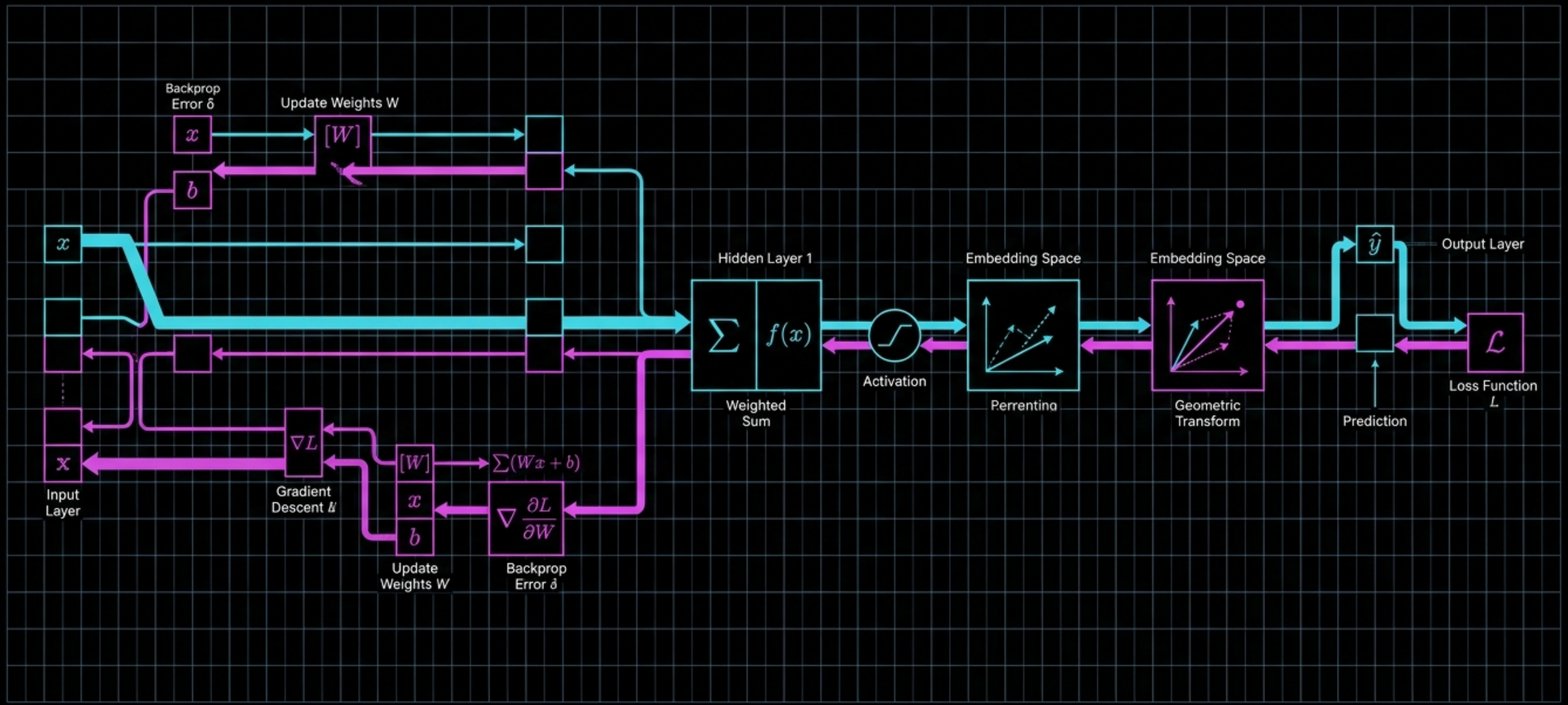
AI Basics — S01E01: Neural Networks, Backpropagation, Embeddings & Representation Geometry

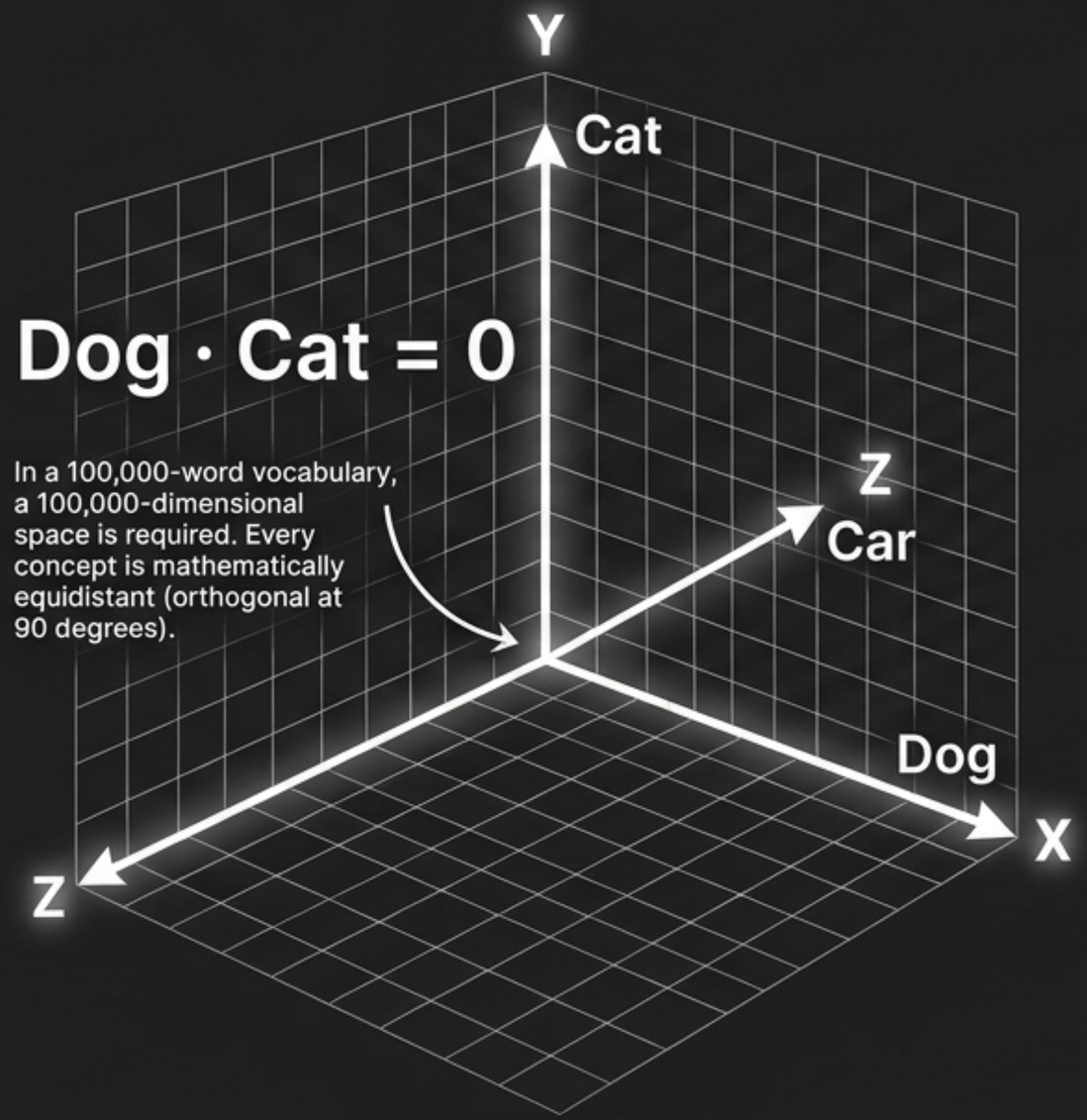
A Systems-Level Schematic of Modern Machine Learning



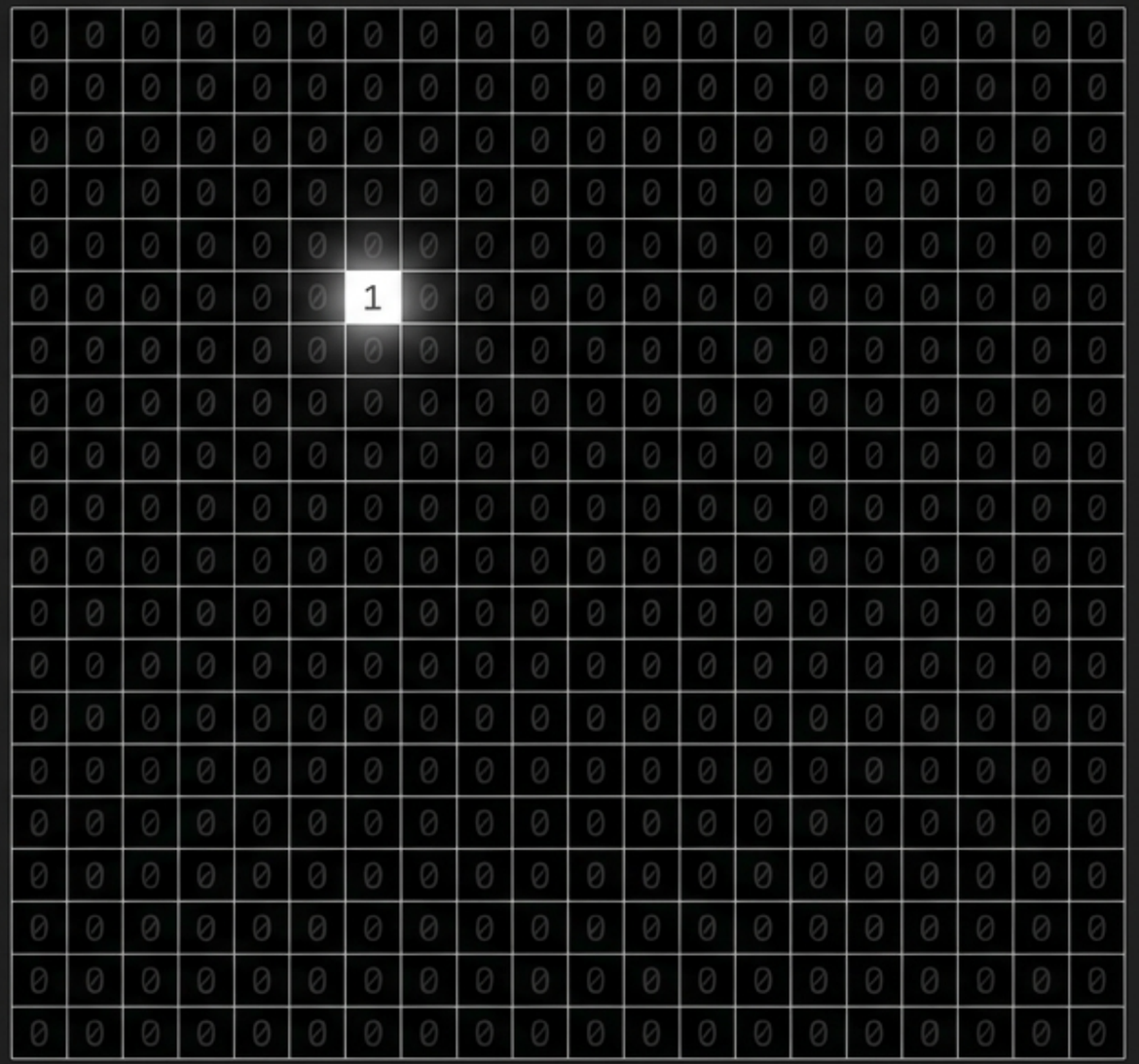
Gemini Notebook

THE MATHEMATICAL FAILURE OF DISCRETE SYMBOLS

One-hot encodings map every token to an orthogonal axis, resulting in zero relational geometry and an unscalable vocabulary explosion.



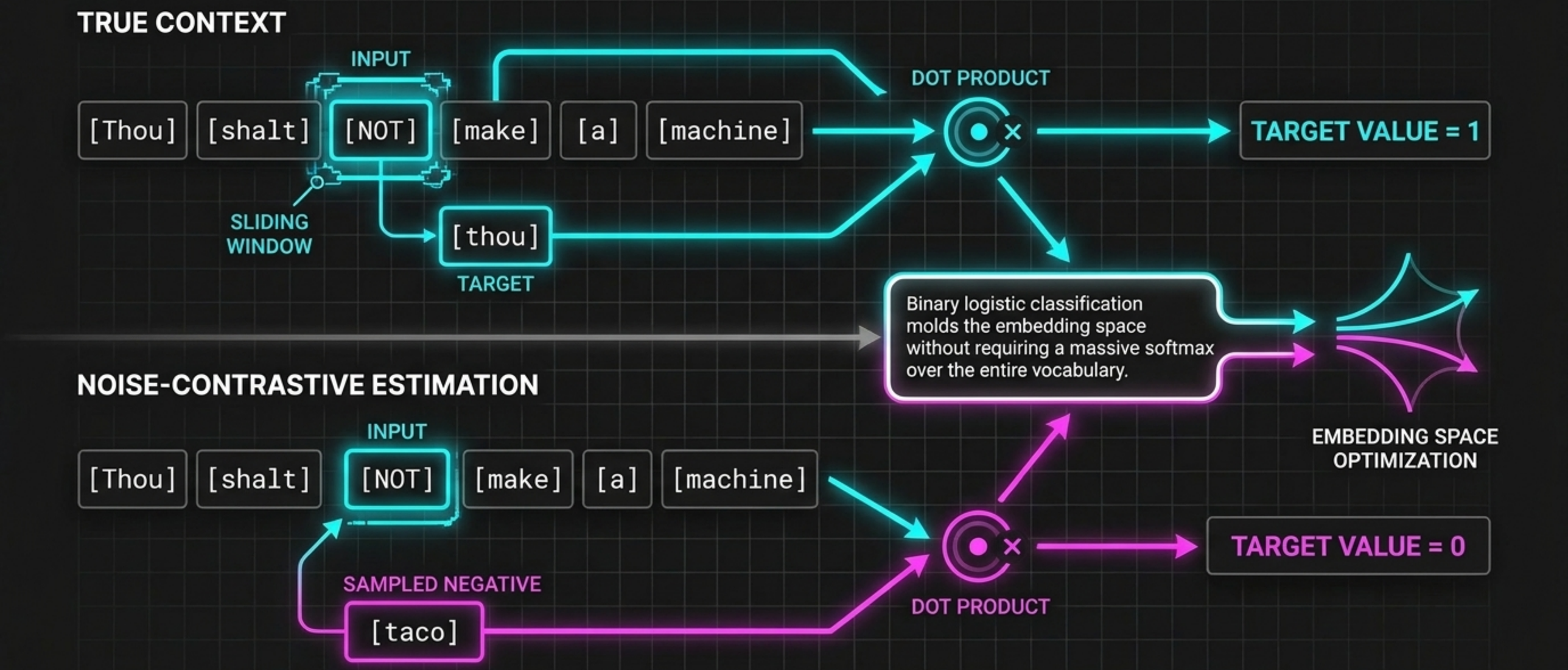
Gross inefficiency: Storing millions of zeros for a single 1.



Gemini Notebook

MAPPING DISCRETE TOKENS TO CONTINUOUS VECTOR SPACES

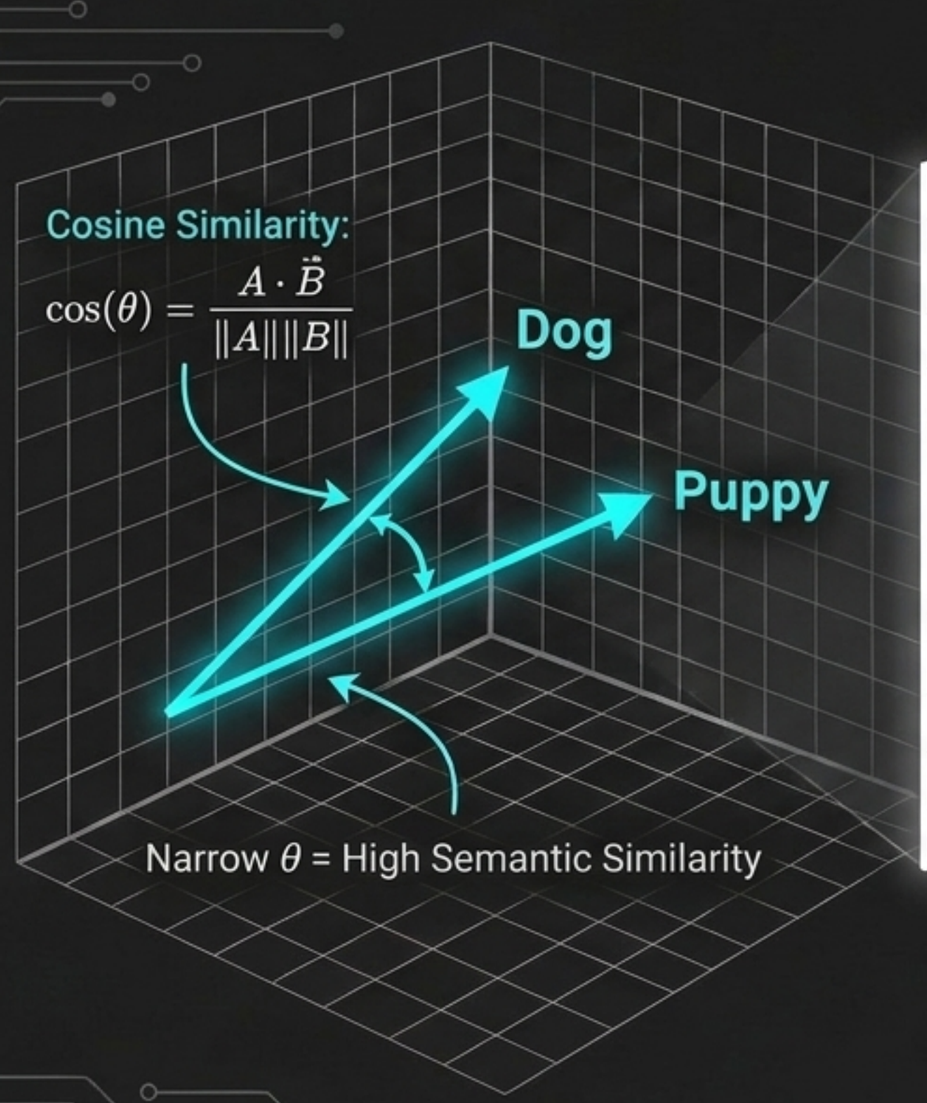
Dense embeddings project symbols into lower-dimensional continuous space, trained by contrasting actual contexts against sampled noise.

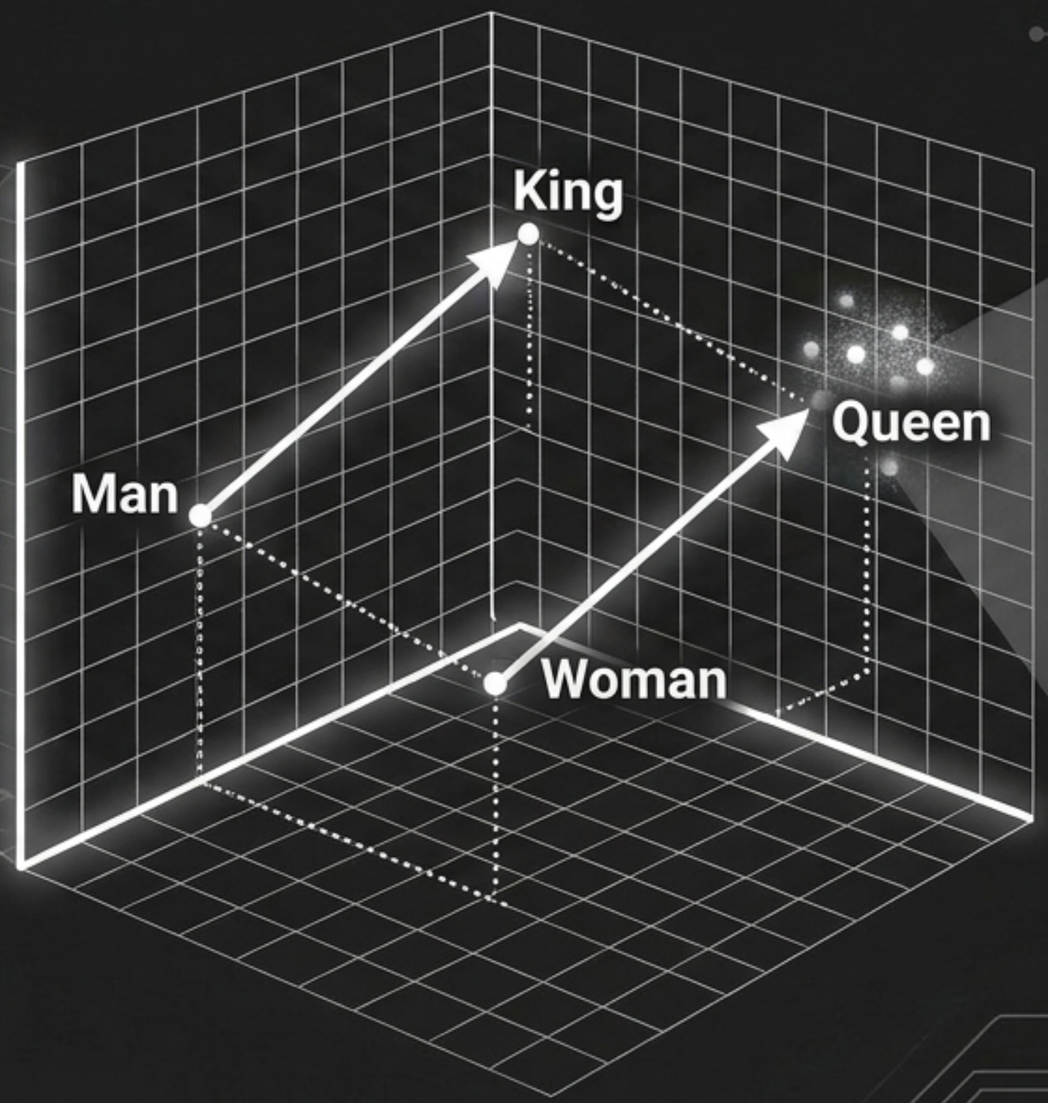


Gemini Notebook

DIRECTIONAL ALIGNMENT ENCODES SEMANTICS AND SYNTAX

In dense vector spaces, semantic relationships manifest as linear relational offsets, measurable via cosine similarity and dot products.



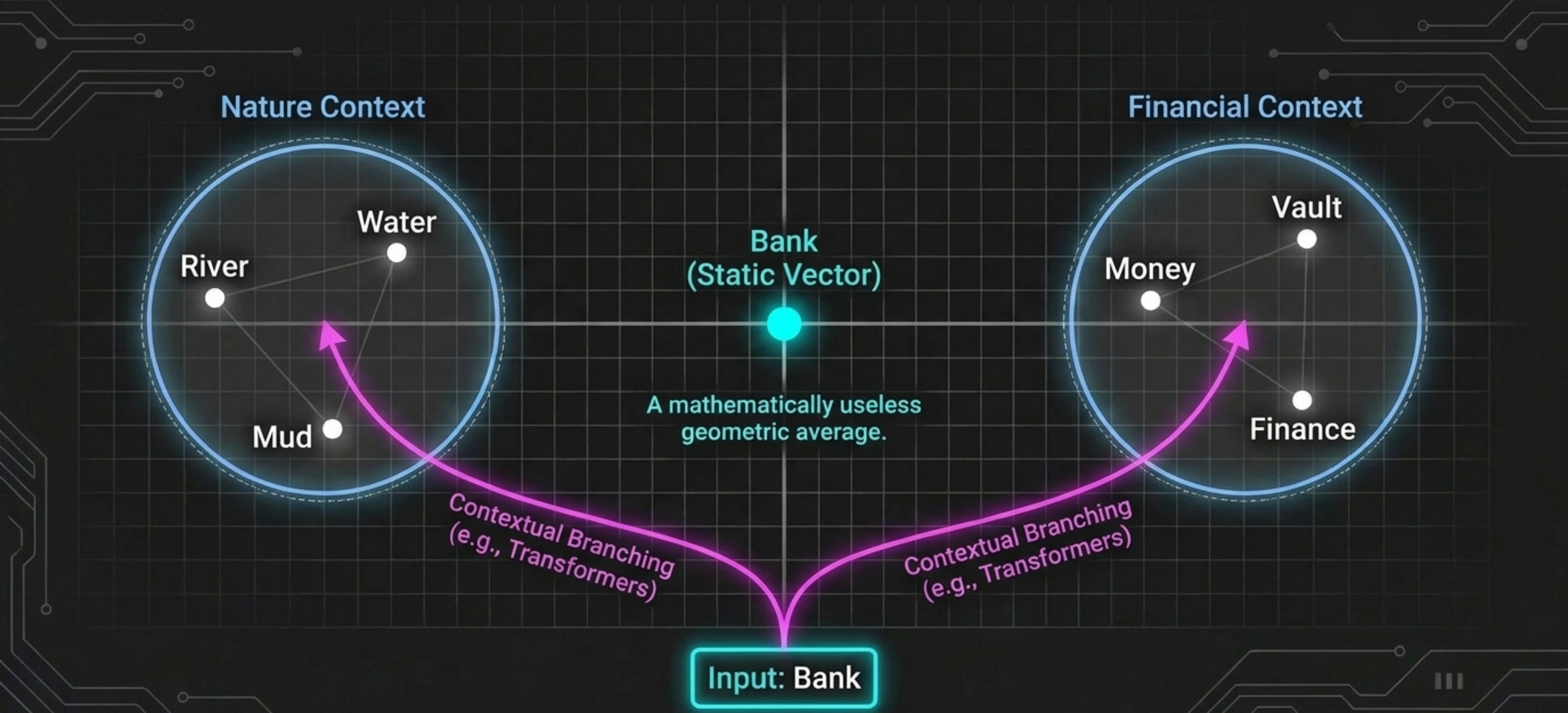


King
- Man
+
Woman
≈
Queen

Gemini Notebook

THE POLYSEMY BOTTLENECK IN STATIC REPRESENTATIONS

Static single-vector mappings collapse distinct contextual meanings into a single geometric average, degrading representational fidelity.

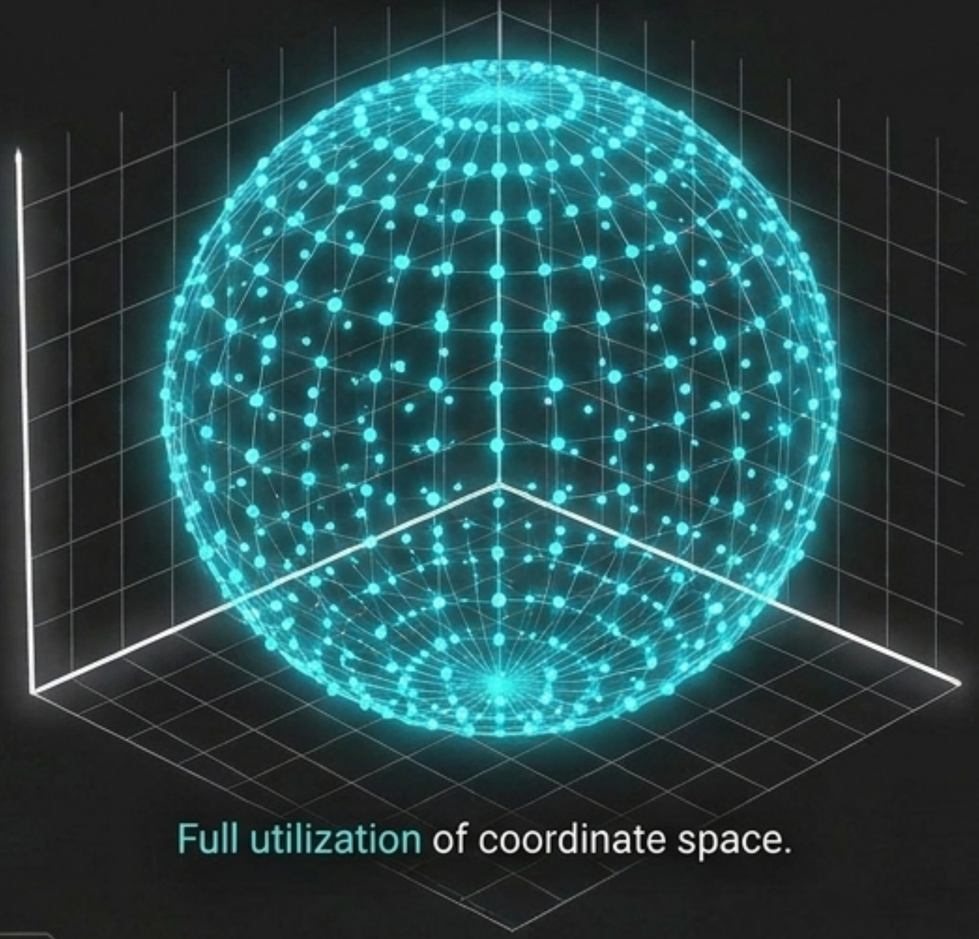


Gemini Notebook

HIGH-DIMENSIONAL GEOMETRY PITFALLS: THE CONE OF ANISOTROPY

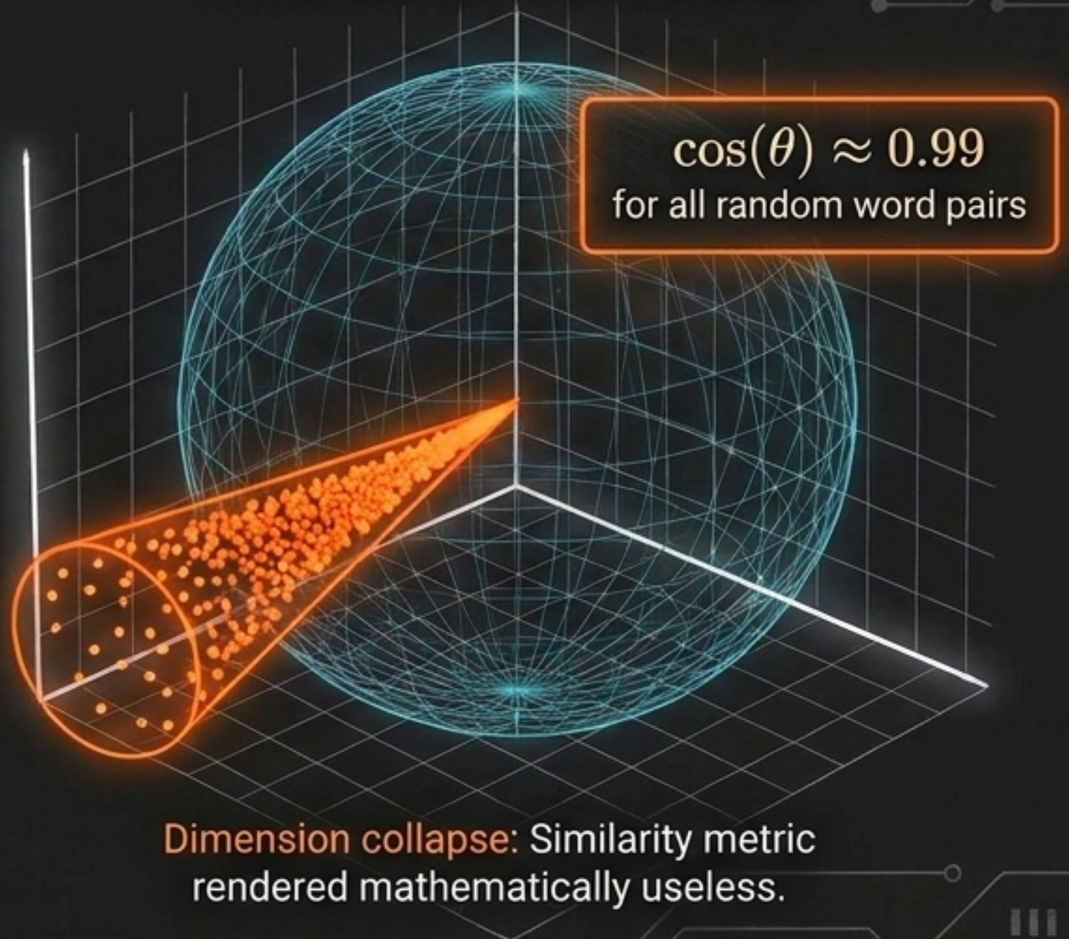
Without strict regularization, high-dimensional embeddings tend to cluster into a narrow cone, causing dimension collapse and destroying cosine discriminability.

Ideal (Isotropic)



Full utilization of coordinate space.

Pathological (Anisotropic)



Dimension collapse: Similarity metric rendered mathematically useless.

Gemini Notebook

SYSTEM TRADE-OFFS: DIMENSIONALITY VS. RAM AND LATENCY

Increasing embedding dimensions yields diminishing returns in expressivity while scaling hardware costs linearly.

Dimension	Expressivity	VRAM per 100k Vocab (FP32)	Search Latency
d = 384	Sufficient for most semantic tasks	150 MB	Ultra-low (Fast)
d = 4096	Marginal gains in complex relations	1.6 GB	High (Compute bound)

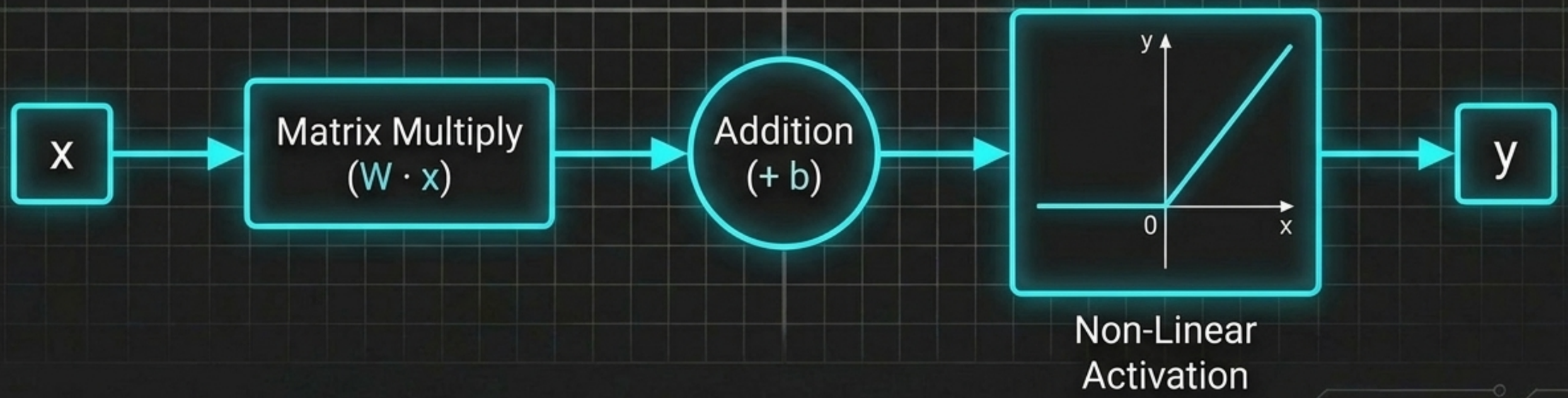
```
Static Lookup Table Memory Footprint:
100,000 tokens × 4096 dimensions × 4 bytes (FP32) ≈ 1.6 GB
```

Gemini Notebook

NEURAL LAYERS AS COMPOSED MATRIX TRANSFORMATIONS

A layer is a directed acyclic graph (DAG) applying a linear affine transformation followed by a non-linear continuous squash.

$$y = \max(0, Wx + b)$$

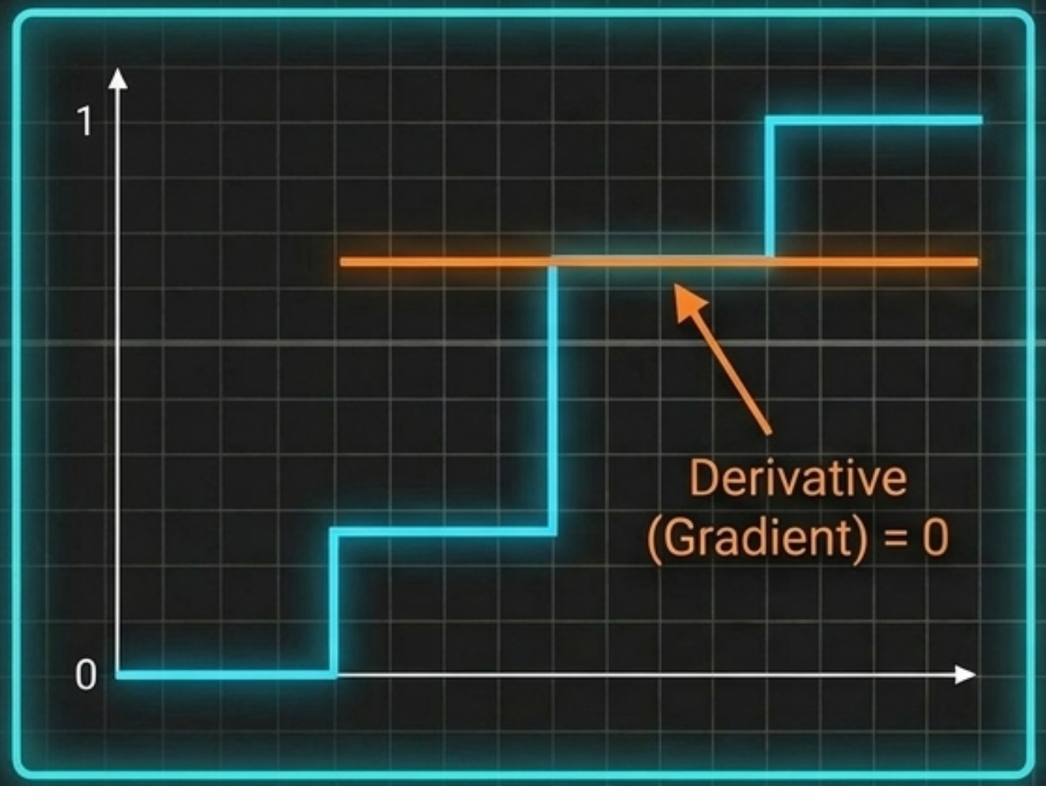


Gemini Notebook

THE NON-DIFFERENTIABILITY OF ACCURACY

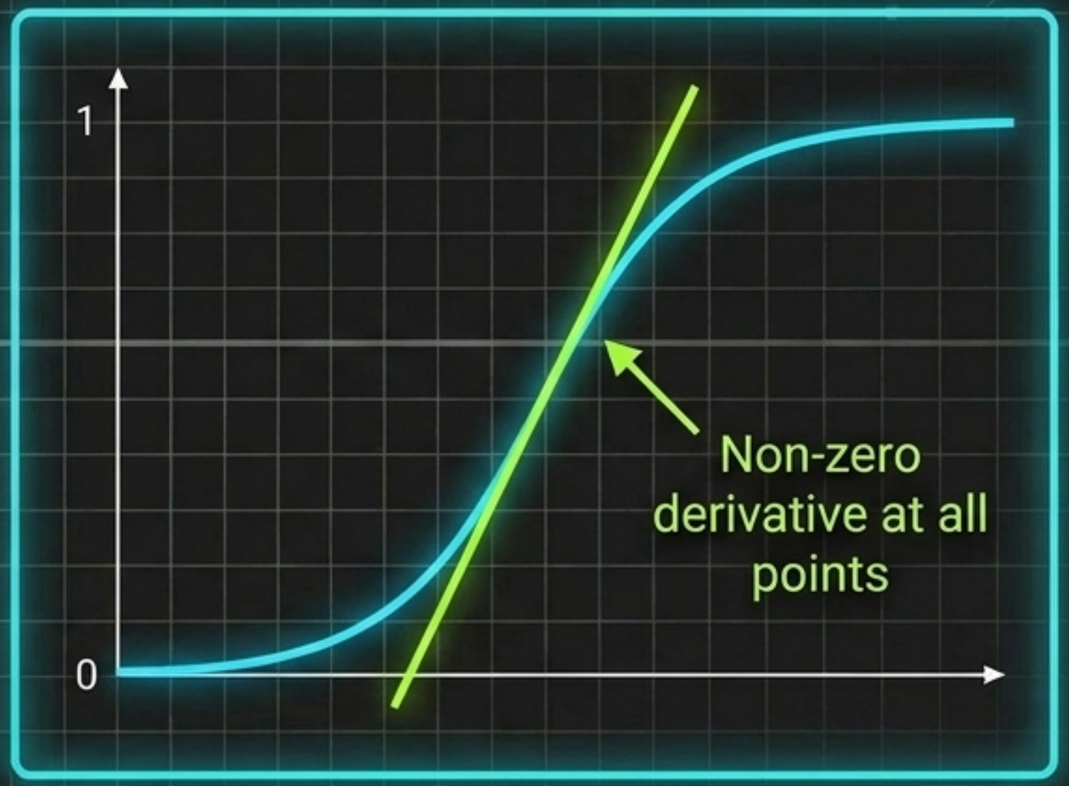
Accuracy is a discrete step-function with a gradient of zero almost everywhere. Optimization requires continuous, smooth surrogate functions.

Accuracy (Discrete)



Zero gradient = zero learning.

Smooth Surrogate (e.g., Logistic/Sigmoid)

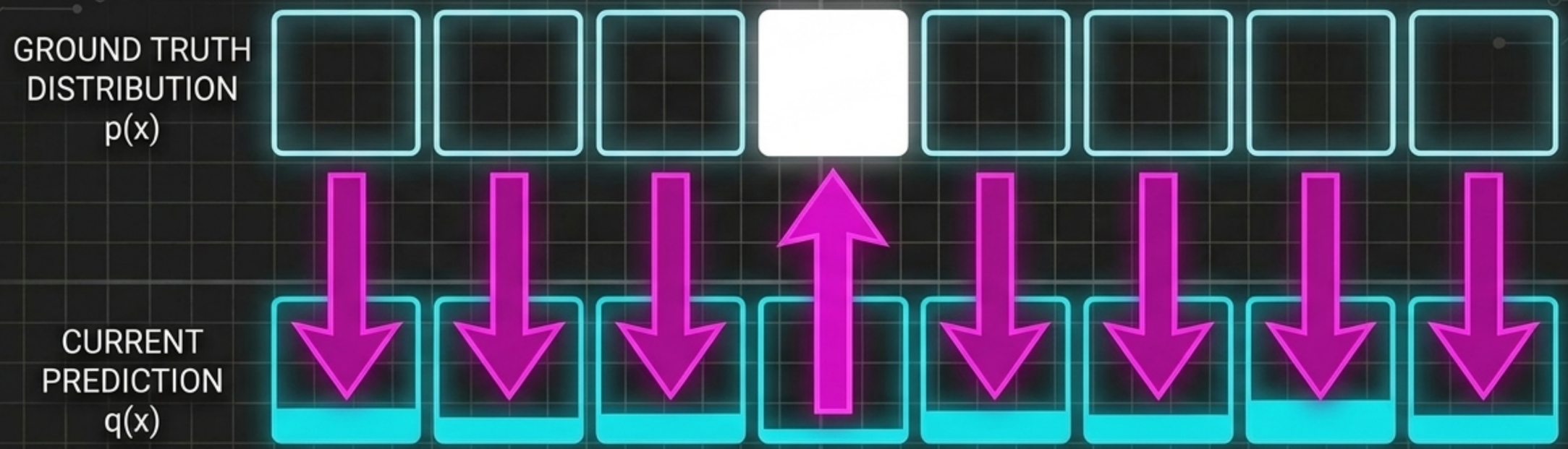


Provides a continuous directional vector toward the minimum.

Gemini Notebook

INFORMATION-THEORETIC LOSS: SHIFTING PROBABILITY MASS

Cross-Entropy loss penalizes the divergence between the true data distribution p(x) and the model's predicted probability distribution q(x).



$$H(p, q) = -\sum p(\mathrm{x}) \log(q(\mathrm{x}))$$

Loss function forces q(x) probability mass to reshape and match p(x) over successive iterations.

Gemini Notebook

THE FORWARD PASS: GRAPH EXECUTION AND ACTIVATION CACHING

As data flows forward, intermediate activations must be calculated and locked into GPU VRAM to allow for subsequent gradient calculation.

State-Transition Directed Acyclic Graph (DAG), time t=1

$h_1 = W_1 x$ → $h_2 = \text{ReLU}(h_1)$ → $L = \text{CrossEntropy}(h_2, y)$



Memory Heap

Intermediate matrix states are saved and consume VRAM for the duration of the training step.

Gemini Notebook

THE BACKWARD PASS: CHAIN RULE AND JACOBIAN-VECTOR PRODUCTS

Gradients flow in reverse from the scalar loss, multiplying local derivatives with upstream gradients to compute parameter updates.

State-Transition Directed Acyclic Graph (DAG), time t=1 (Reverse Flow)

$h_1 = W_1 x$

$\frac{dL}{dW_1} = \frac{dL}{dh_1} \cdot \frac{dh_1}{dW_1}$

$h_2 = \mathrm{ReLU}(h_1)$

$\frac{dL}{dL} = 1$



Memory Heap

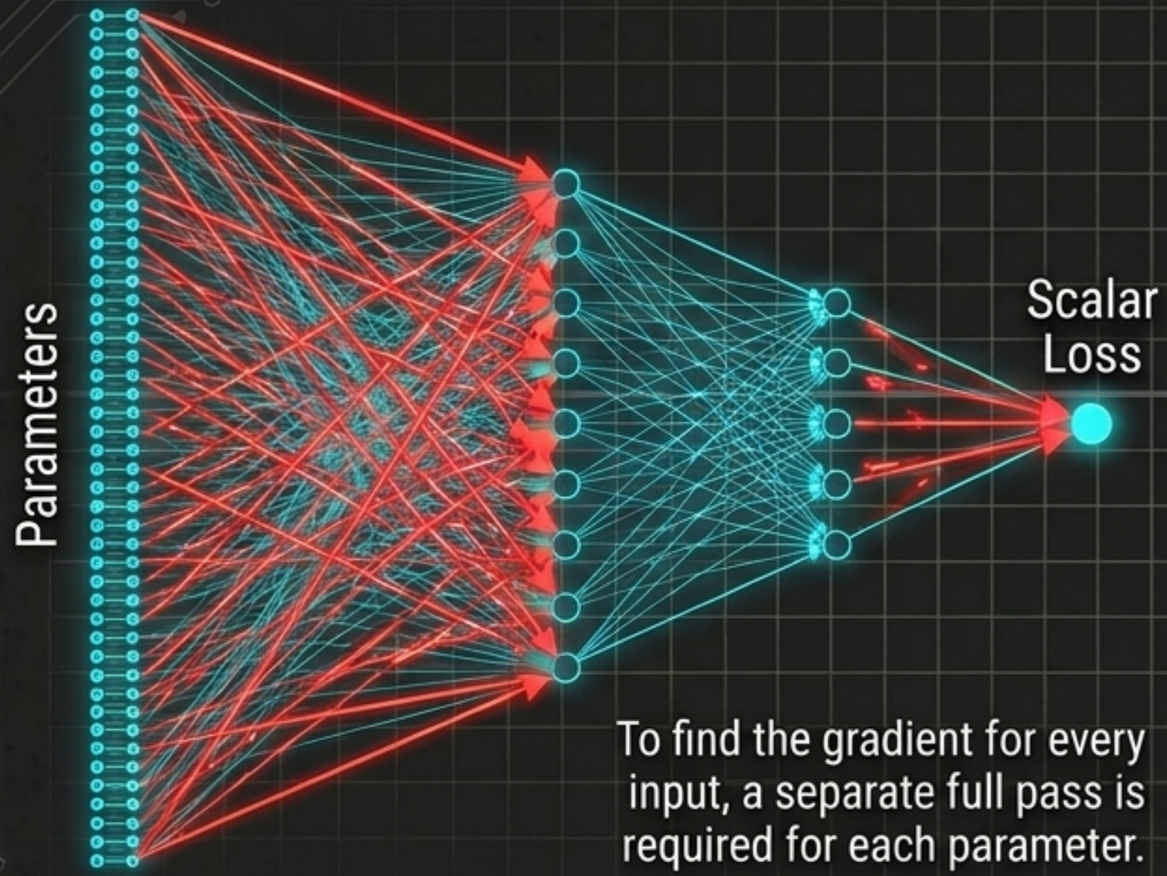
Cached activations are retrieved from VRAM to compute local derivatives, then memory is freed.

Gemini Notebook

SCALING EFFICIENCY: WHY REVERSE-MODE AUTODIFF?

For networks mapping millions of parameters to a single scalar loss, reverse-mode computes all gradients in O(1) passes.

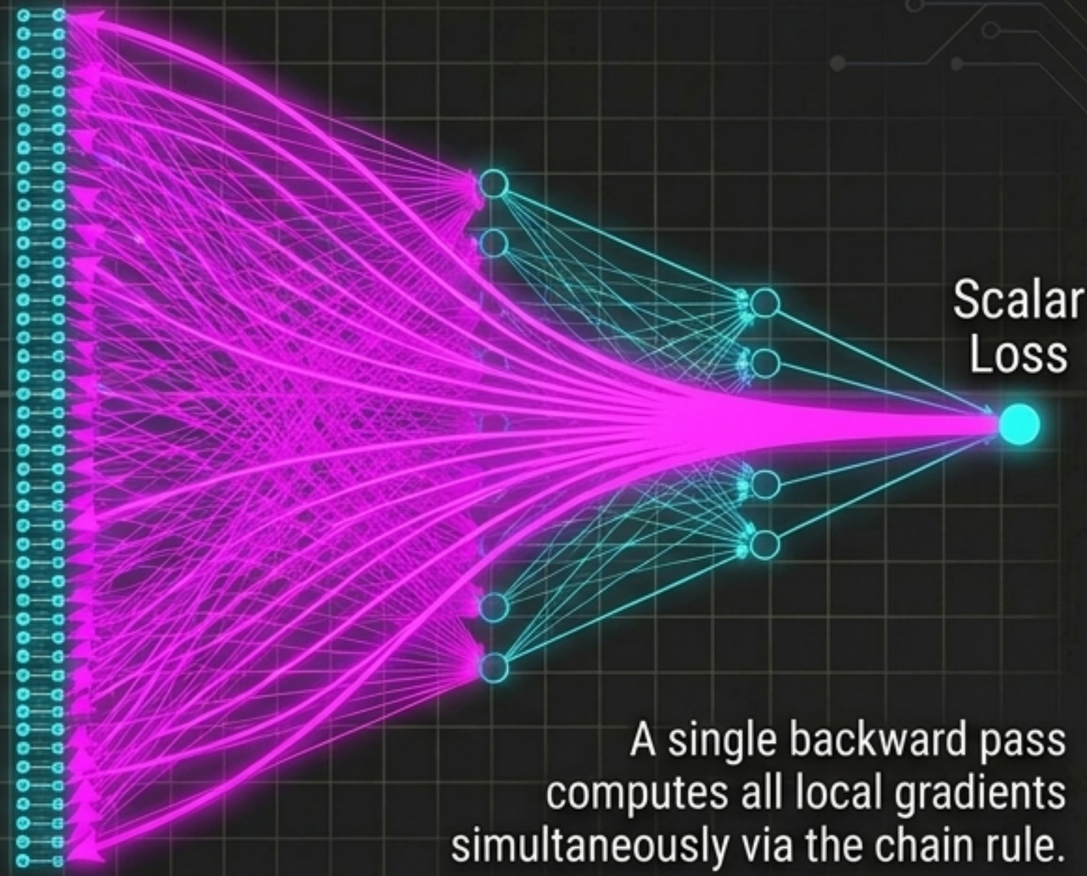
FORWARD-MODE AUTODIFF



To find the gradient for every input, a separate full pass is required for each parameter.

COST: N PASSES

REVERSE-MODE / BACKPROPAGATION



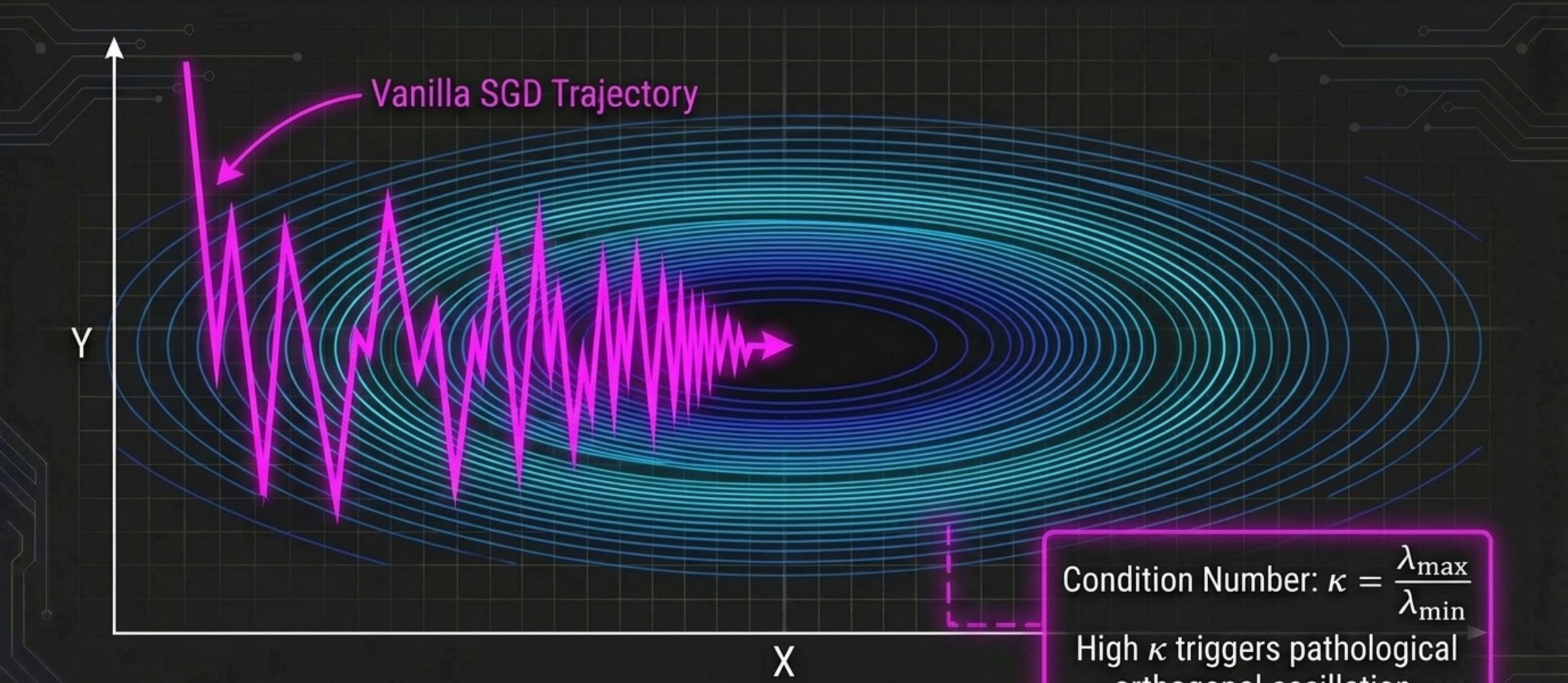
A single backward pass computes all local gradients simultaneously via the chain rule.

COST: 1 PASS. A 1,000,000x SPEEDUP.

Gemini Notebook

NAVIGATING PATHOLOGICAL LOSS LANDSCAPES

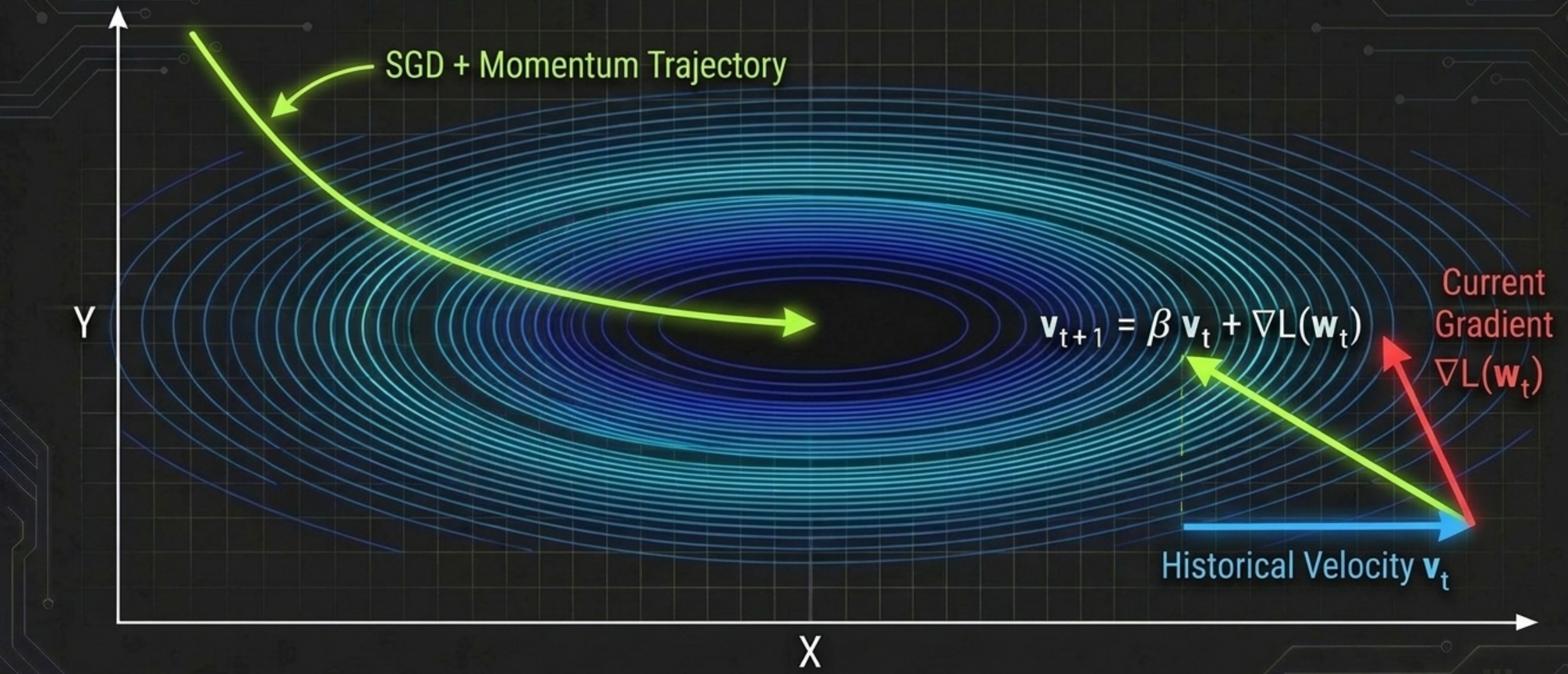
Vanilla Stochastic Gradient Descent (SGD) fails in ill-conditioned Hessian ravines, wasting computation on severe orthogonal oscillation.



Gemini Notebook

STABILIZING DESCENT WITH EXPONENTIAL MOVING AVERAGES

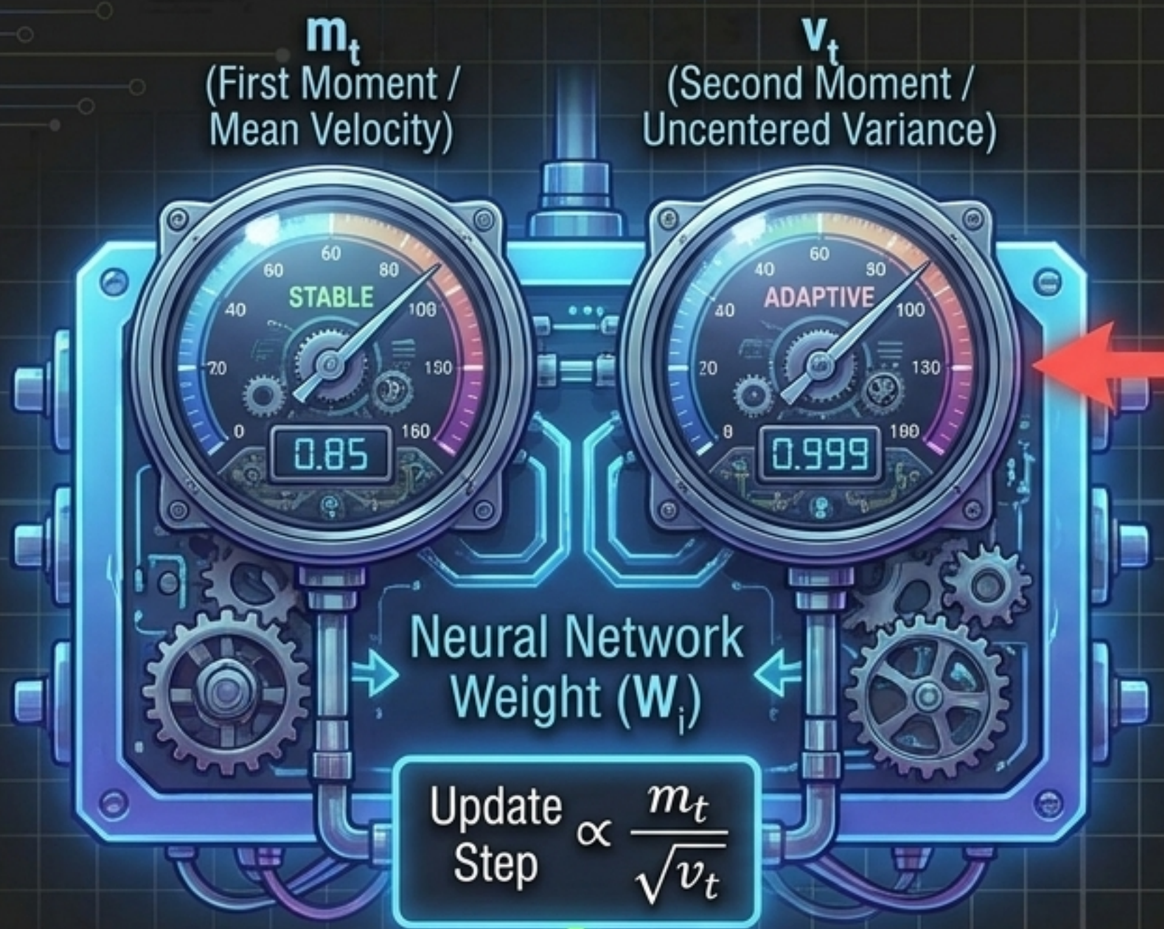
Momentum accumulates historical gradient velocity, dampening high-curvature oscillation while accelerating along the low-curvature optimal path.



Gemini Notebook

AdamW: Adaptive Moments and Decoupled Weight Decay

Adam tracks adaptive per-parameter first (velocity) and second (variance) moments, while AdamW decouples weight decay to prevent regularization distortion.



L2 Regularization (Incorrectly Mixed)

L2 Penalty ($\lambda||W||_2^2$)

Penalty Incorrectly mixed into v_t gauge calculation; leads to distortion.

AdamW Decoupled Weight Decay (Correct)

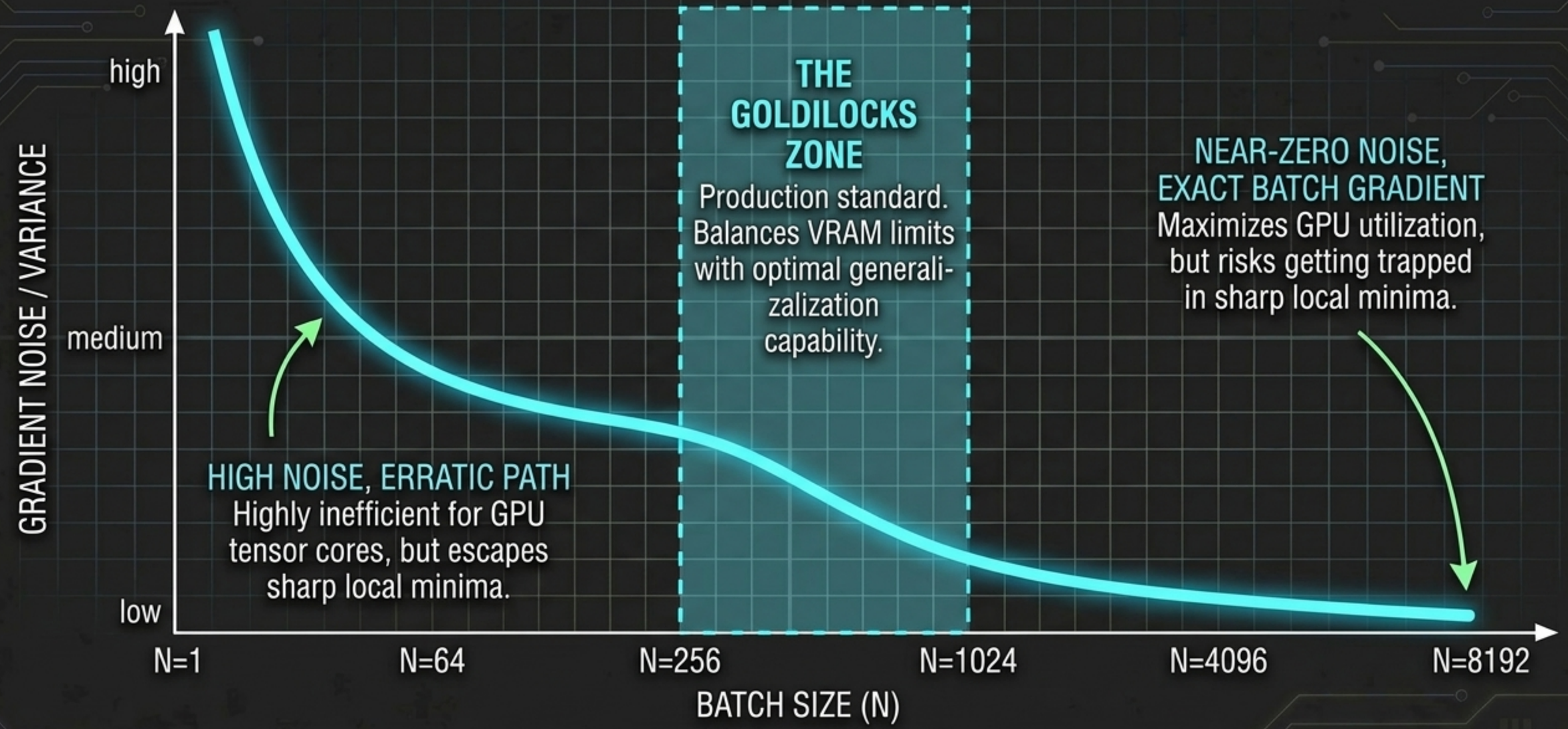
Decoupled Weight Decay Penalty ($\lambda * W$)

Correctly applied external to gradient moments. Prevents regularization distortion.

Gemini Notebook

DIAGNOSTIC TRADE-OFFS: BATCH SIZE VS. GRADIENT NOISE

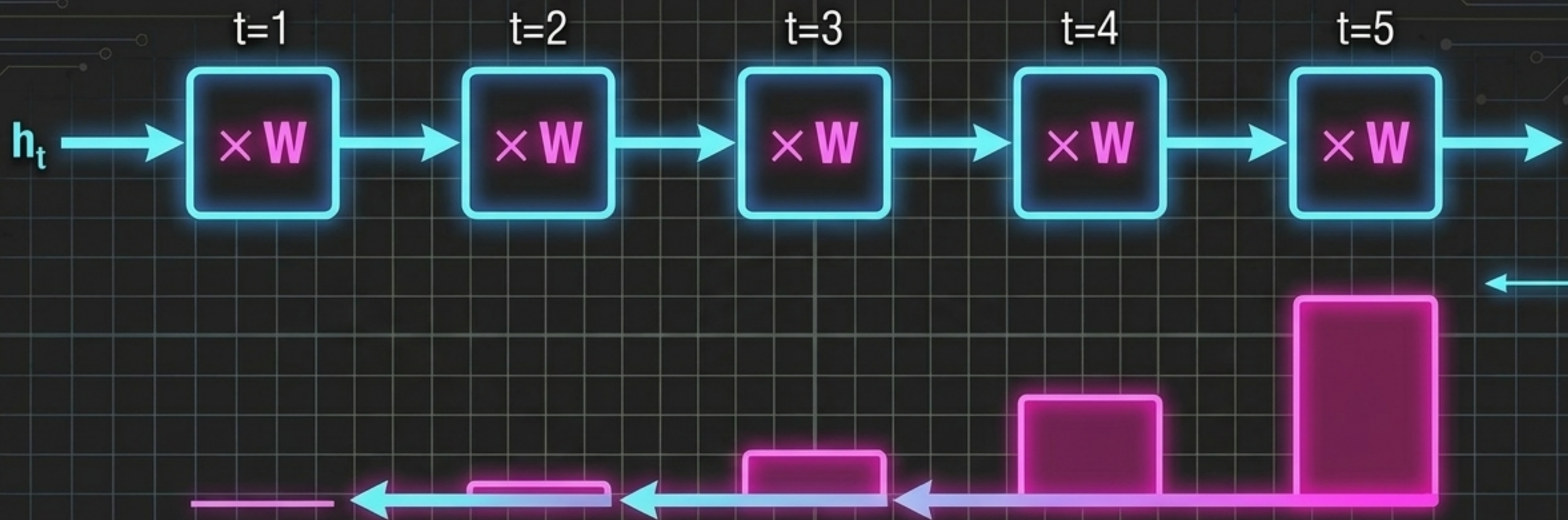
Balancing stochastic gradient noise (which aids generalization) against hardware-efficient parallel matrix multiplication.



Gemini Notebook

Deep Sequential Bottlenecks: Unrolled Networks

Vanilla RNNs force gradients through repeated multiplicative state transitions, guaranteeing that gradients will either exponentially vanish or explode.



The Vanishing Gradient Problem

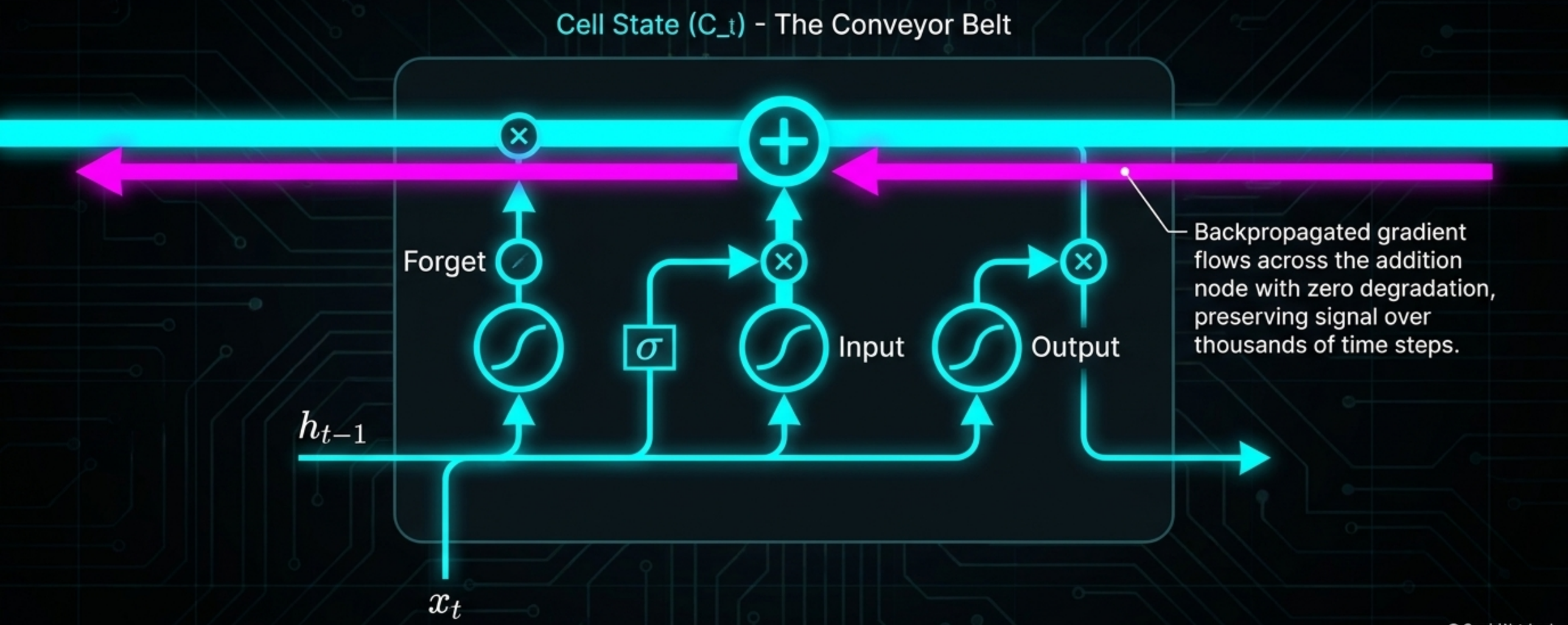
$$\frac{\partial h_t}{\partial h_1} = W^{t-1}$$

Because the dominant eigenvalue of W is < 1, multiplicative decay is mathematically guaranteed.

Gemini Notebook

LSTM: Additive Cell-State Highways

Long Short-Term Memory networks solve multiplicative decay by introducing an additive cell state, regulated by sigmoid gates, acting as an uninterrupted gradient highway.

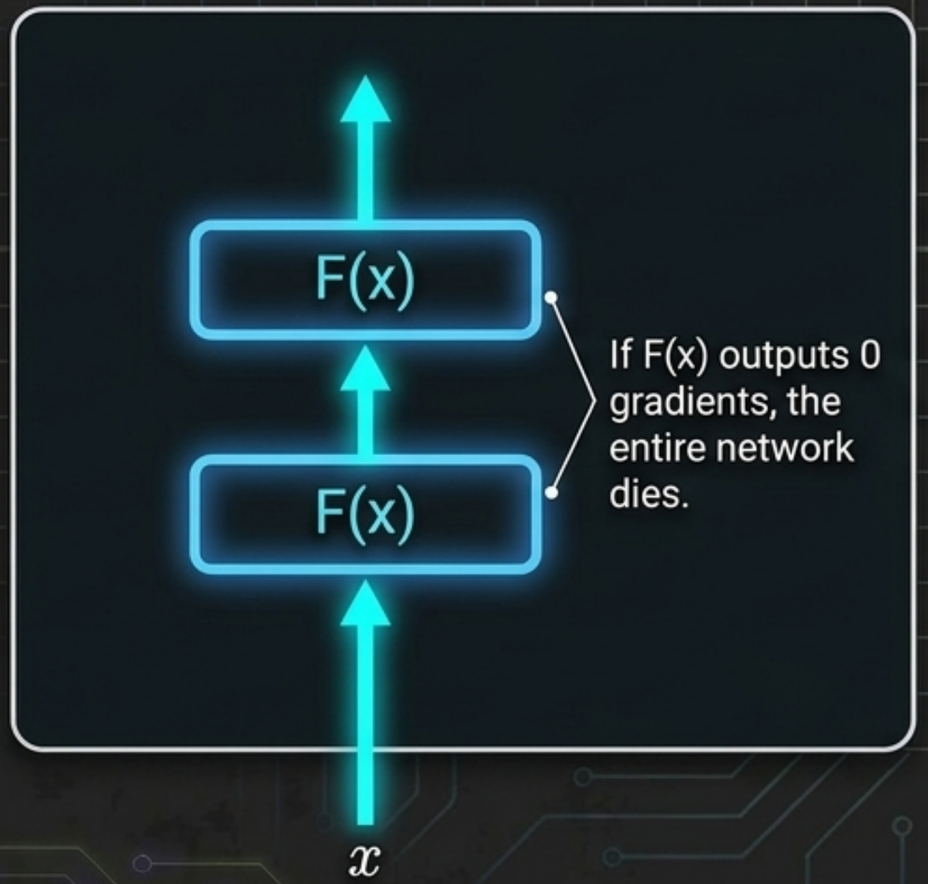


Gemini Notebook

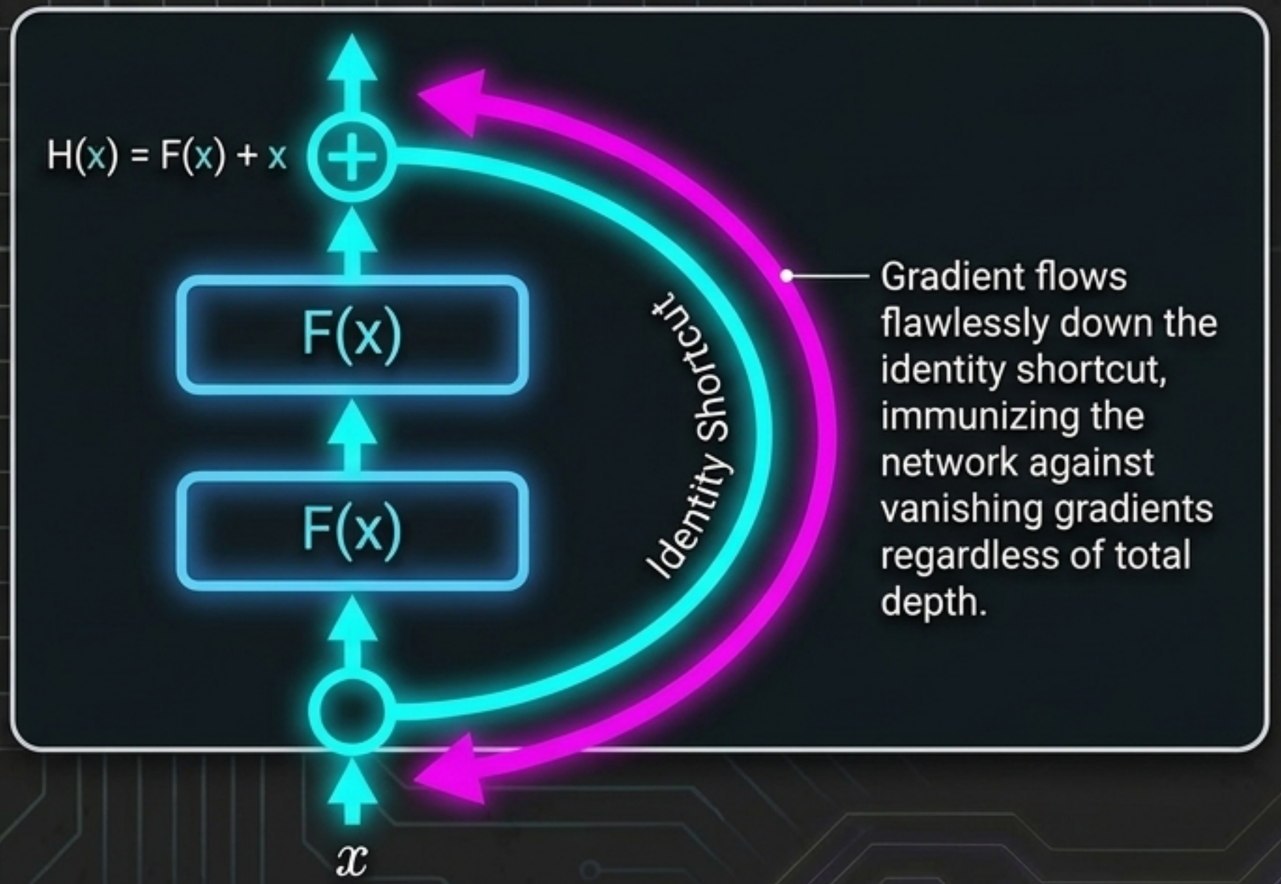
Residual Connections: Identity Shortcuts for Deep Topologies

By summing a layer's non-linear transformation with its original identity input, ResNets create stable gradient super-highways across ultra-deep networks.

Standard Feed-Forward



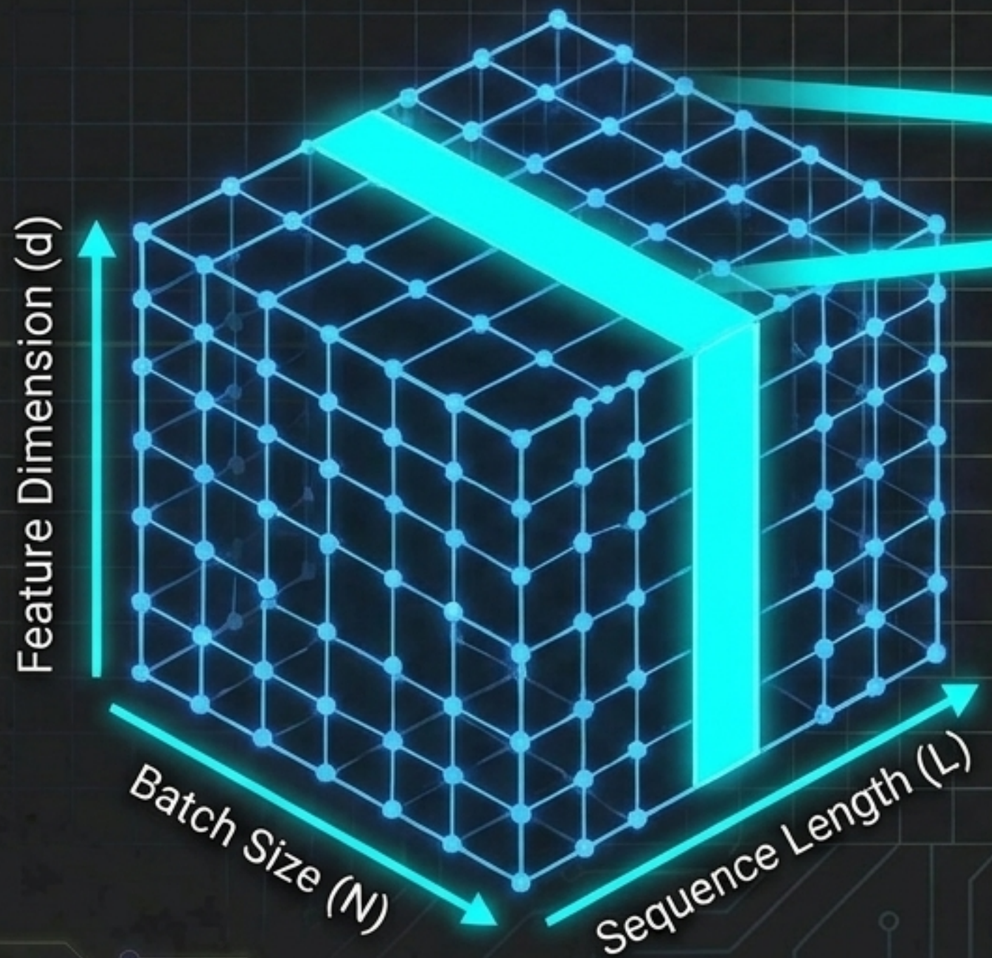
Residual Block



Gemini Notebook

Layer Normalization: Bounding Hidden-State Variance

To prevent catastrophic activation explosion in deep networks, LayerNorm computes and standardizes the mean and variance across the feature dimension per token.



μ, σ^2 calculated along feature dimension **d**

μ

x → σ^2

Unlike Batch Normalization, LayerNorm is independent of other samples in the batch, making it mandatory for variable-length sequence models.

$$y = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \cdot \gamma + \beta$$

Gemini Notebook

Production Memory Arithmetic: Training vs. Inference

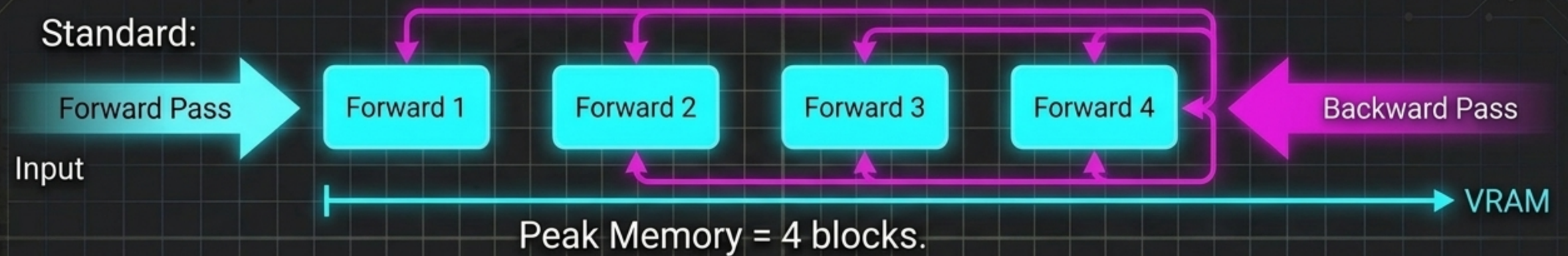
Training a model consumes 4*x to 8x more VRAM than inference due to the necessity of storing gradients, optimizer states, and forward activations.

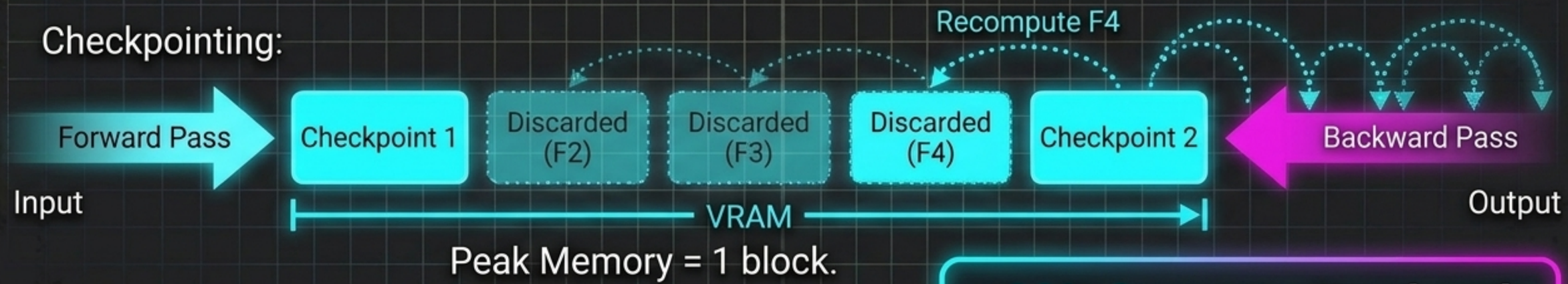


Gemini Notebook

Activation Checkpointing: Trading Compute for Memory

Instead of keeping all forward activations resident in VRAM, checkpointing discards intermediate states and recomputes them on-the-fly during the backward pass.



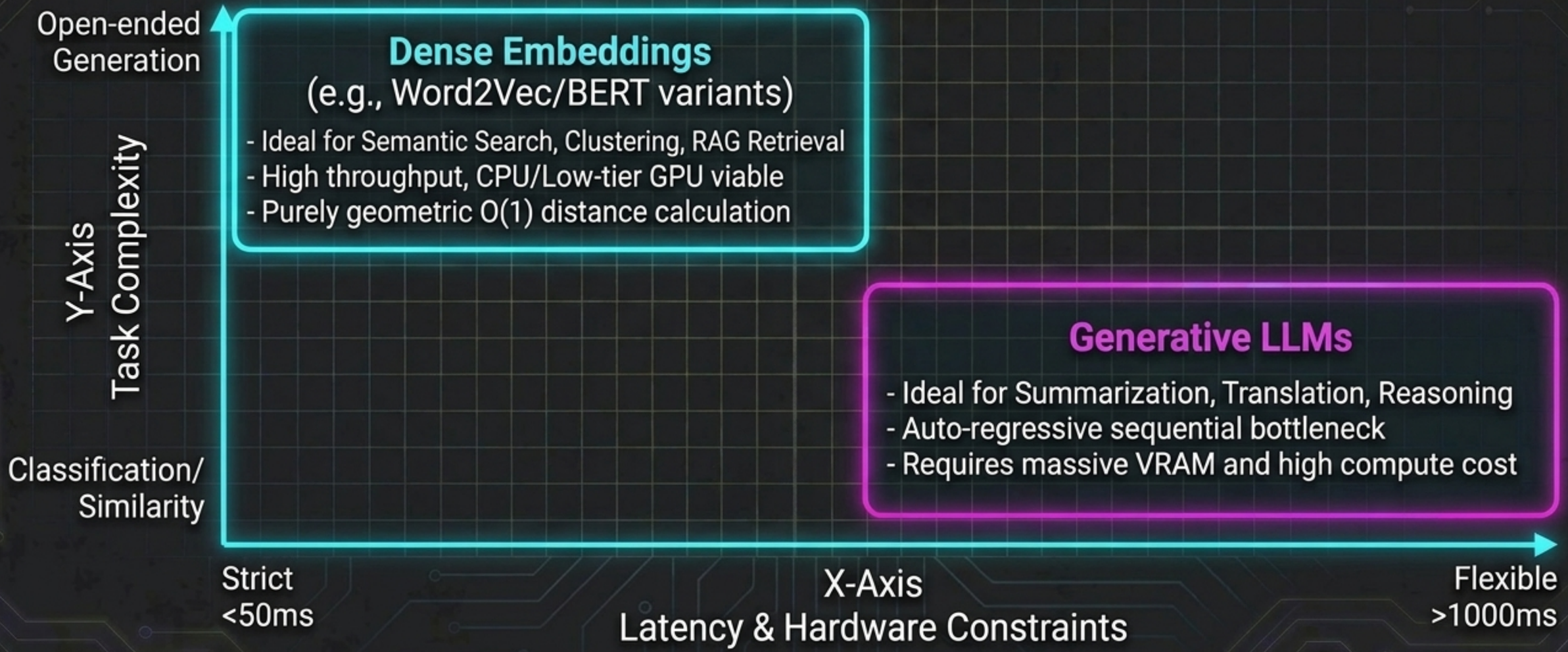


+20% Compute Overhead
for **-70%** Memory Footprint.

Gemini Notebook

Architectural Decision Boundaries: Embeddings vs. Generative LLMs

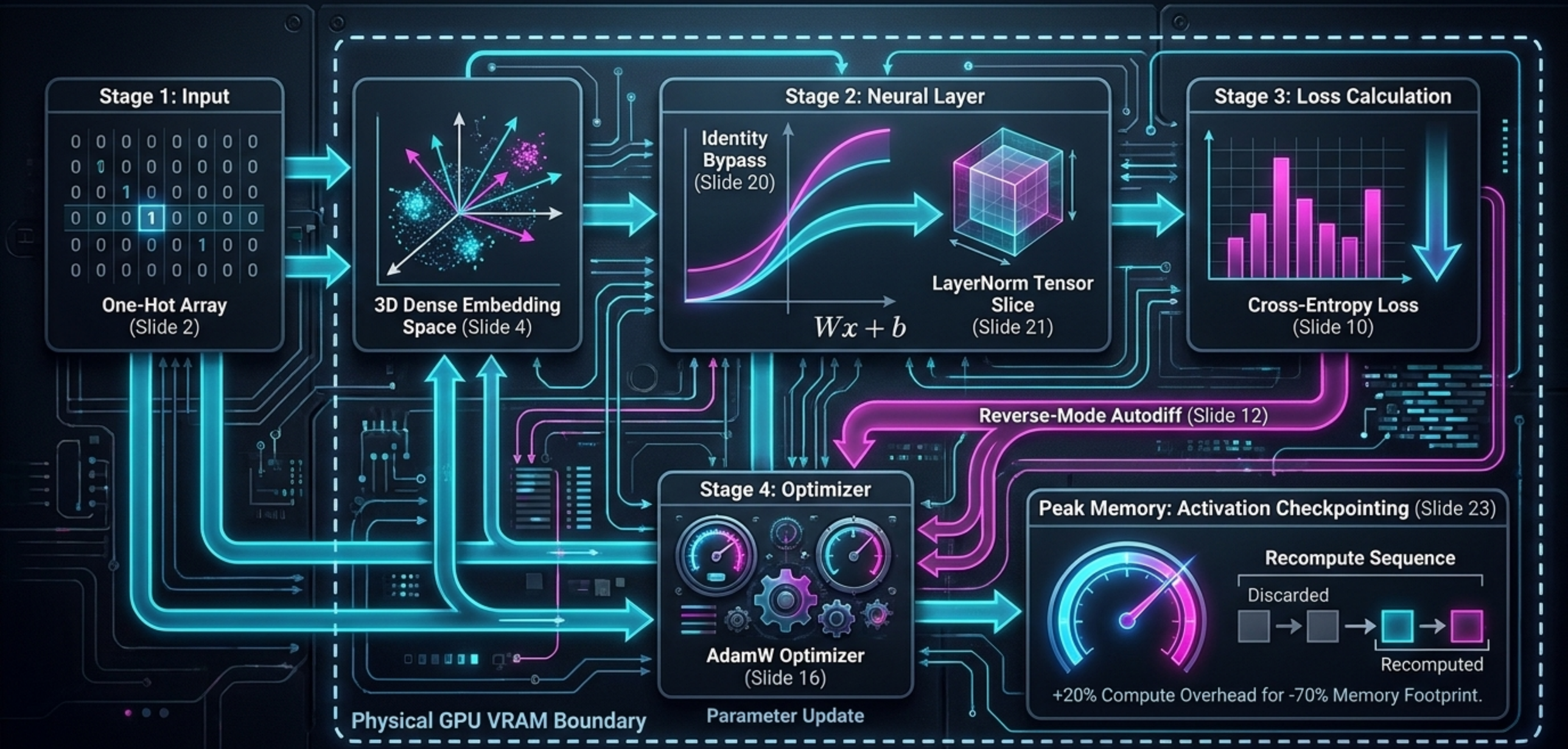
Match the systemic complexity to the production workload: use lightweight dense embeddings for geometric comparisons and LLMs for generative synthesis.



Gemini Notebook

Synthesis: The Complete Neural Architecture Blueprint

From sparse symbol to hardware optimization, the modern AI pipeline is a continuous assembly line of geometric projection, differentiable matrix calculus, and VRAM management.



Gemini Notebook