

BME 6938 · ADVANCED BIOMEDICAL AI · SPRING 2027

LECTURE 4-1

Inside Modern LLMs

Architectures in 2026



Parisa Rashidi · Herbert Wertheim College of Engineering, University of Florida



Where We Are

Weeks 4 and 5 go inside the models that every later week uses.

Weeks 2 and 3

The 2017 transformer

Tokens, attention, encoder and decoder blocks

Week 4: today

Inside modern LLMs

What changed in the block;
memory for long records;
mixture of experts

Week 5

Post-training and reasoning

From pretrained model to clinical assistant

Why Architecture Matters in the Clinic

Three practical questions decide whether a model can be used on patient data.

Does it fit?

Weights and the memory for a long record must fit on the GPUs a hospital owns.

Sections 2 and 3

Can it read the whole record?

Years of notes, labs, and reports: tens of thousands of tokens or more.

Sections 2 and 4

What does each token cost?

Compute per token decides if a tool can run on every admission.

Section 3

Today's Plan

1 What changed since 2017

normalization, feed-forward, positions, attention heads

2 The KV cache

why long records need memory, and three ways to shrink it

3 Mixture of experts

huge models that use a small part of themselves per token

4 More 2026 changes

vocabularies, long context, multi-token prediction, MedGemma

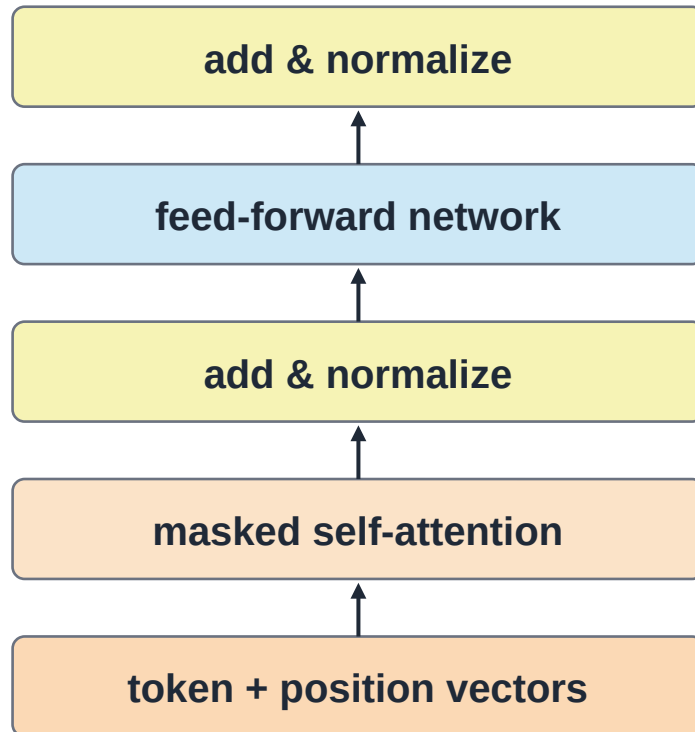


What Changed Since 2017

Recap: The 2017 Decoder Block

Attention, then a feed-forward network, each wrapped by add & normalize.

× N blocks, then predict the next token



normalization

after each sublayer, LayerNorm

feed-forward

two layers with ReLU

positions

added once, at the input

attention

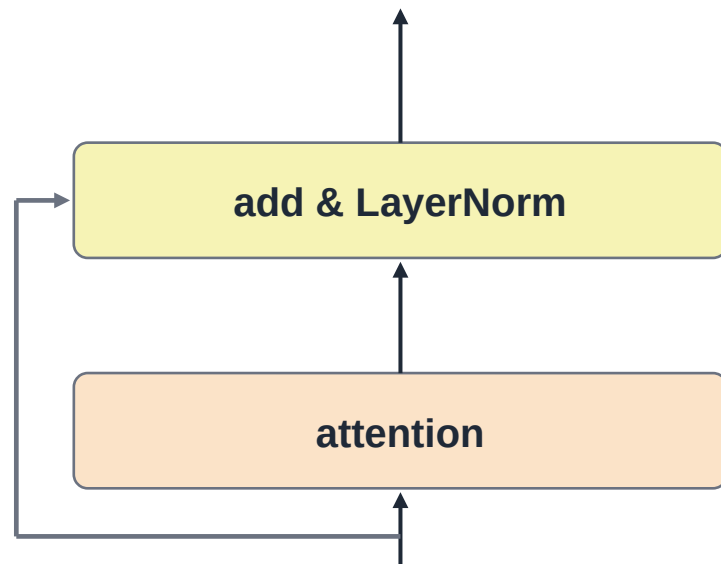
every head has its own keys and values

All four are replaced in a 2026 model. The block structure stays.

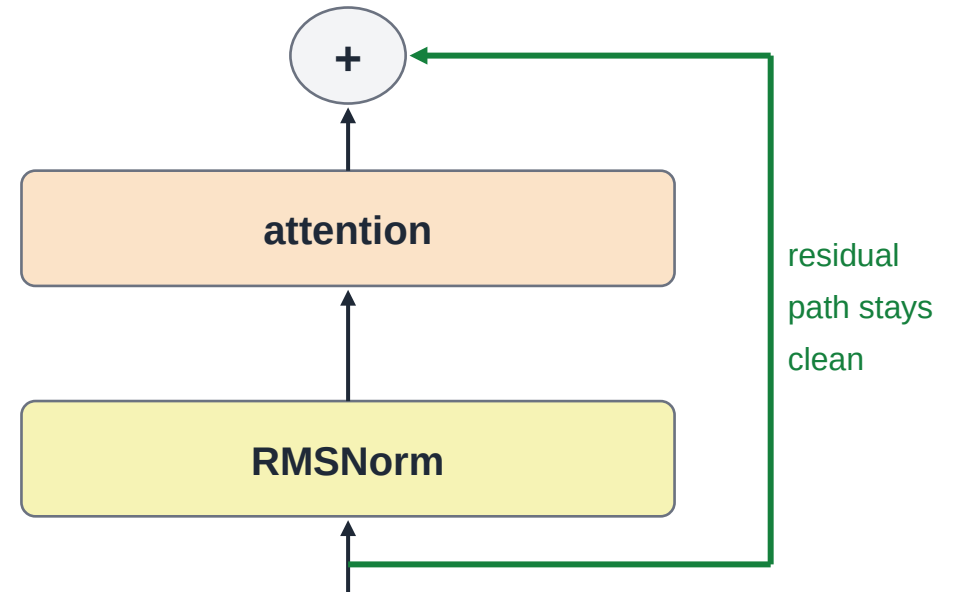
Change 1: RMSNorm, Before Each Sublayer

Normalize the input to a sublayer, not its output, and only rescale.

2017: normalize after



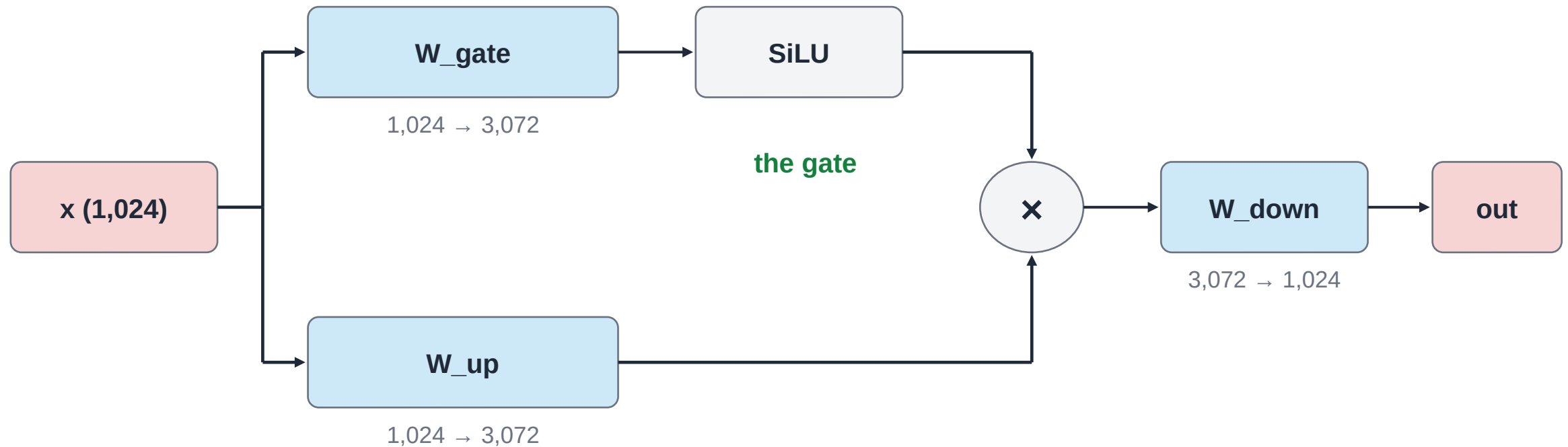
2026: normalize before



$$\text{RMSNorm}(x) = g \cdot x / \sqrt{(\text{mean}(x^2))} \quad x = [3, -1, 1, -3] \rightarrow [1.34, -0.45, 0.45, -1.34]$$

Change 2: A Gated Feed-Forward Network (SwiGLU)

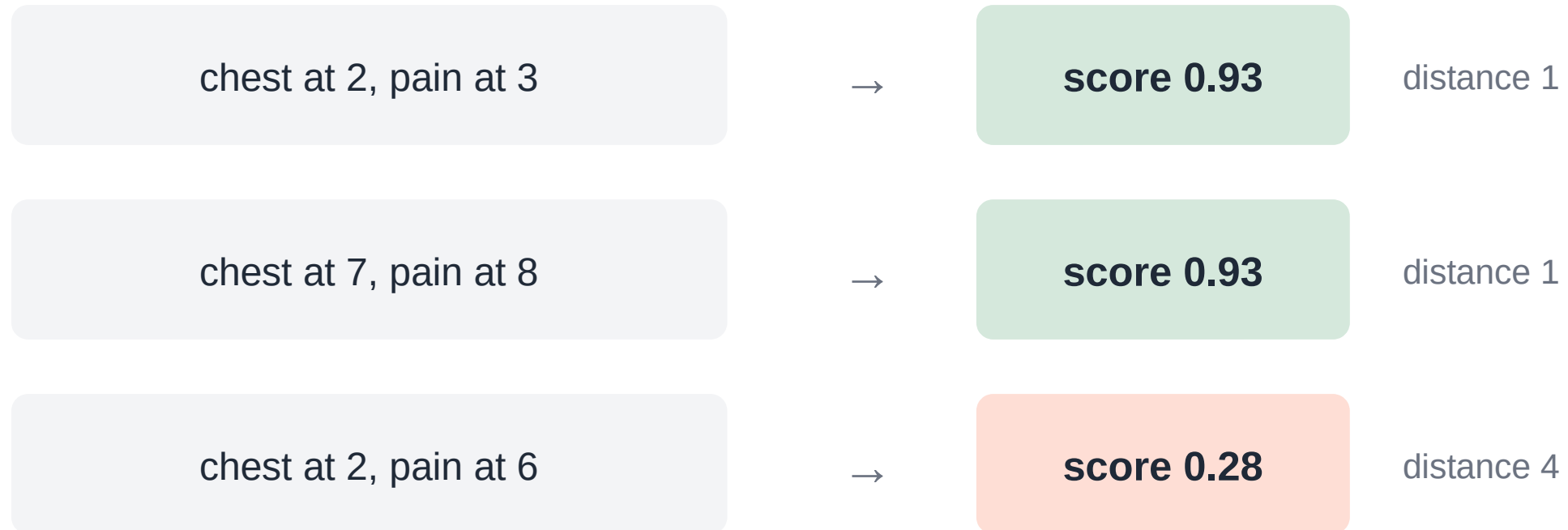
Project the token twice; one projection gates the other.



$$\text{FFN}(x) = W_{\text{down}} (\text{SiLU}(W_{\text{gate}} x) \odot W_{\text{up}} x) \quad 2017: W_2 \text{ReLU}(W_1 x)$$

Change 3: Positions Move Inside Attention

RoPE rotates each query and key by an angle set by its position (Week 2, clip 4).

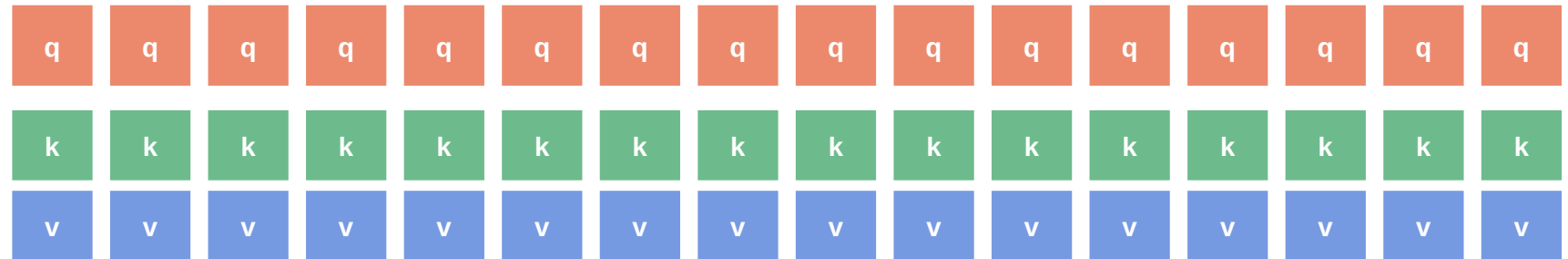


Same distance, same score: position is relative. No position vector is added at the input.

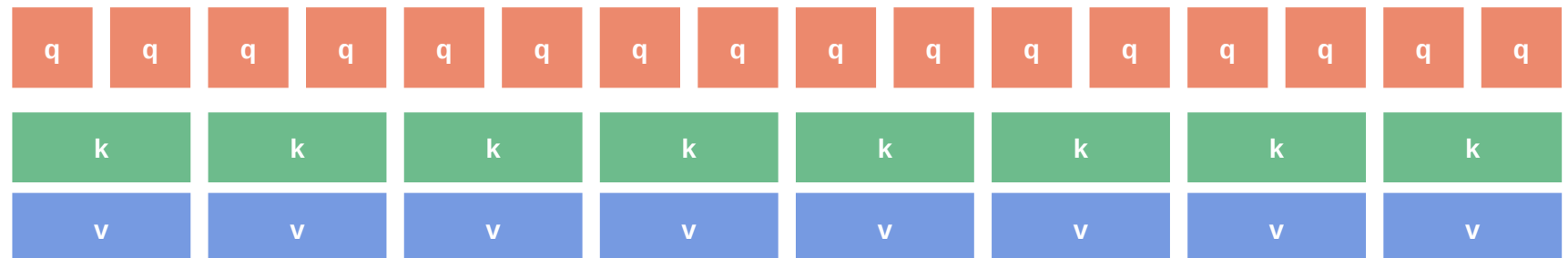
Change 4: Grouped-Query Attention

Several query heads share one key head and one value head.

2017: multi-head



Qwen3-0.6B: grouped-query



16 query heads in both rows; 16 key-value heads at the top, 8 at the bottom

Animation: What Changed Since 2017

Animation clip: week04-clip1-modern-block.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

The 2026 Block at a Glance

	2017 transformer	Qwen3-0.6B (2025)	DeepSeek-V3 (2024)
Normalization	LayerNorm, after	RMSNorm, before	RMSNorm, before
Feed-forward	ReLU, dense	SwiGLU, dense	SwiGLU, mixture of experts
Positions	sinusoidal, added	RoPE in attention	RoPE in attention
Attention	multi-head	grouped-query (16 q, 8 kv)	multi-head latent
Blocks	6	28	61
Parameters	about 65M	0.6B	671B (37B per token)



The KV Cache

Writing One Token at a Time

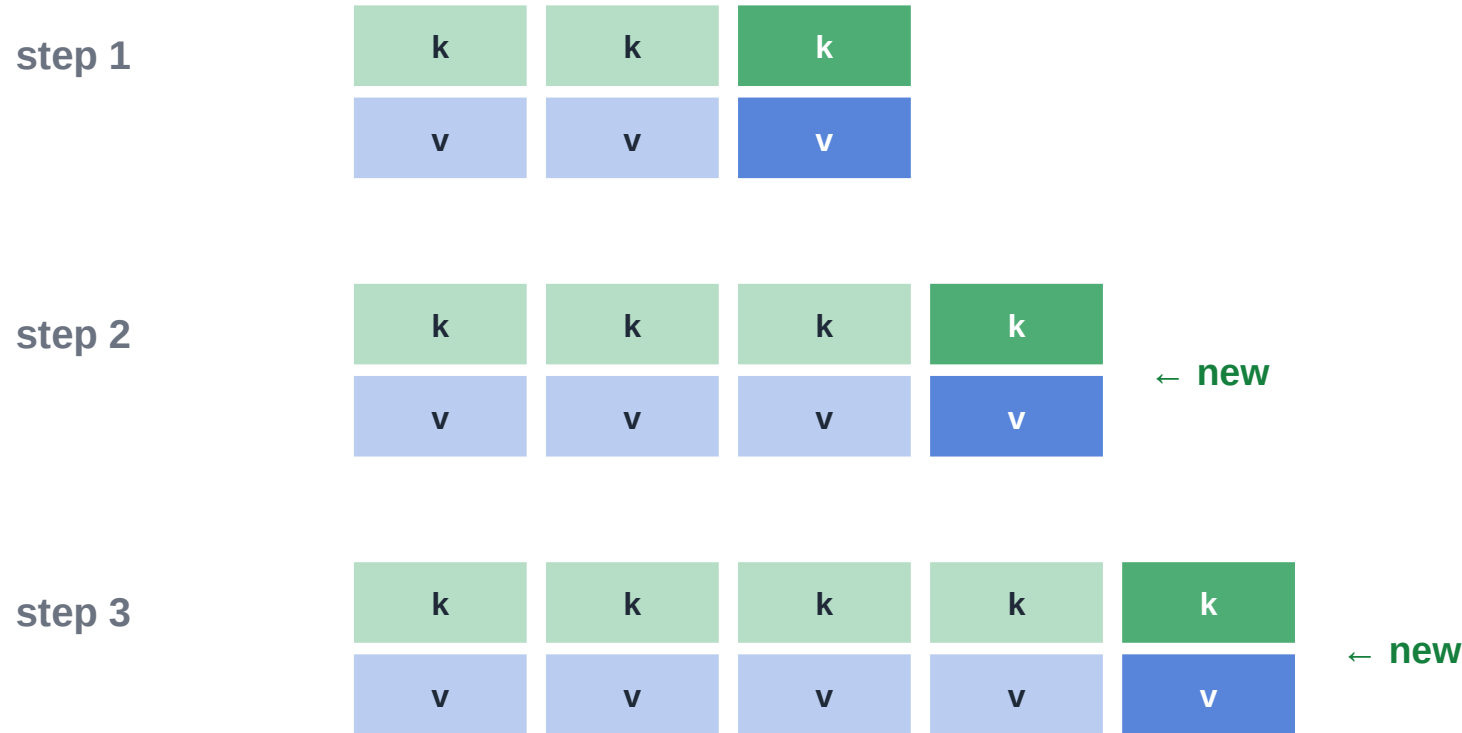
Each new token's query is compared with the keys of every earlier token.



The keys and values of earlier tokens never change.
So compute them once, store them, and reuse them at every step.

The KV Cache

Store every token's keys and values, in every layer and every head.



Cache size per token

2 (keys and values)

× layers

× key-value heads

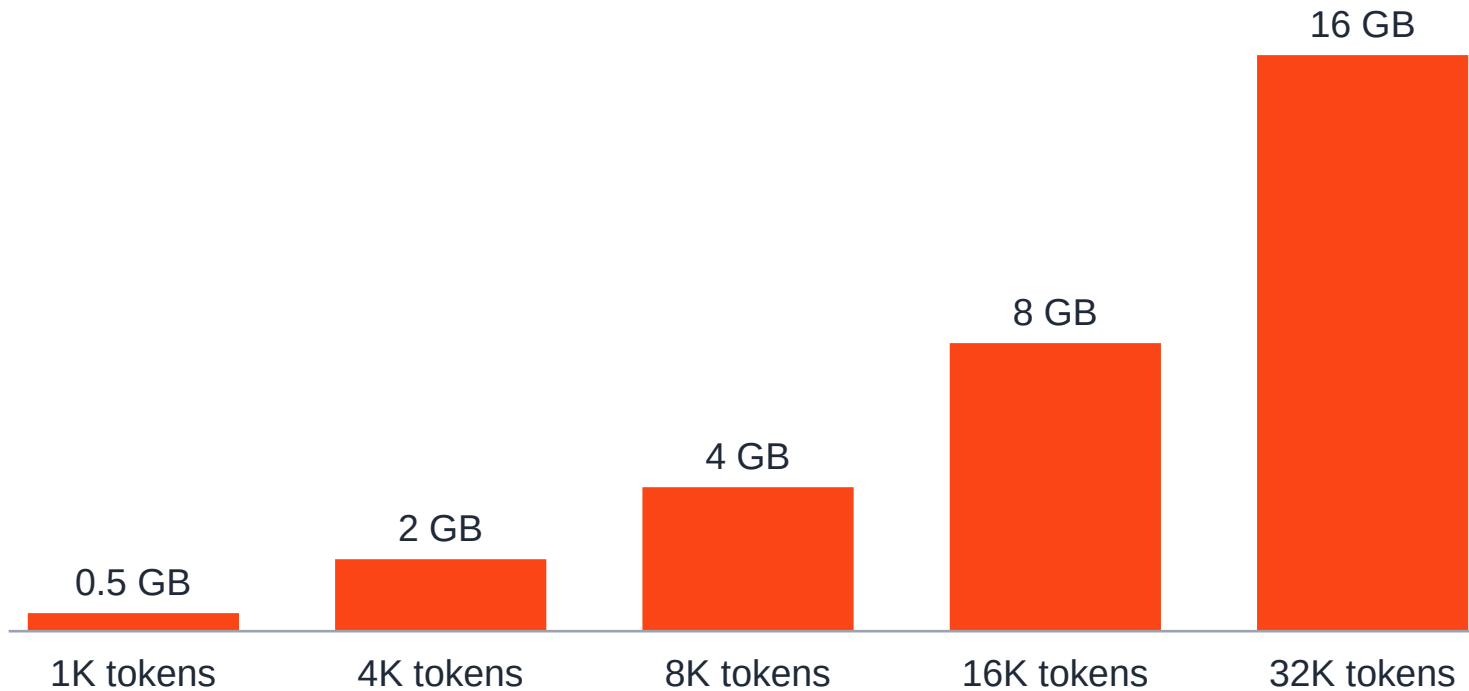
× head size

× bytes per number

How Big Does the Cache Get?

A long patient record can need more memory for the cache than for the model.

$$2 \times 32 \text{ layers} \times 32 \text{ heads} \times 128 \times 2 \text{ bytes} = 0.5 \text{ MB per token}$$

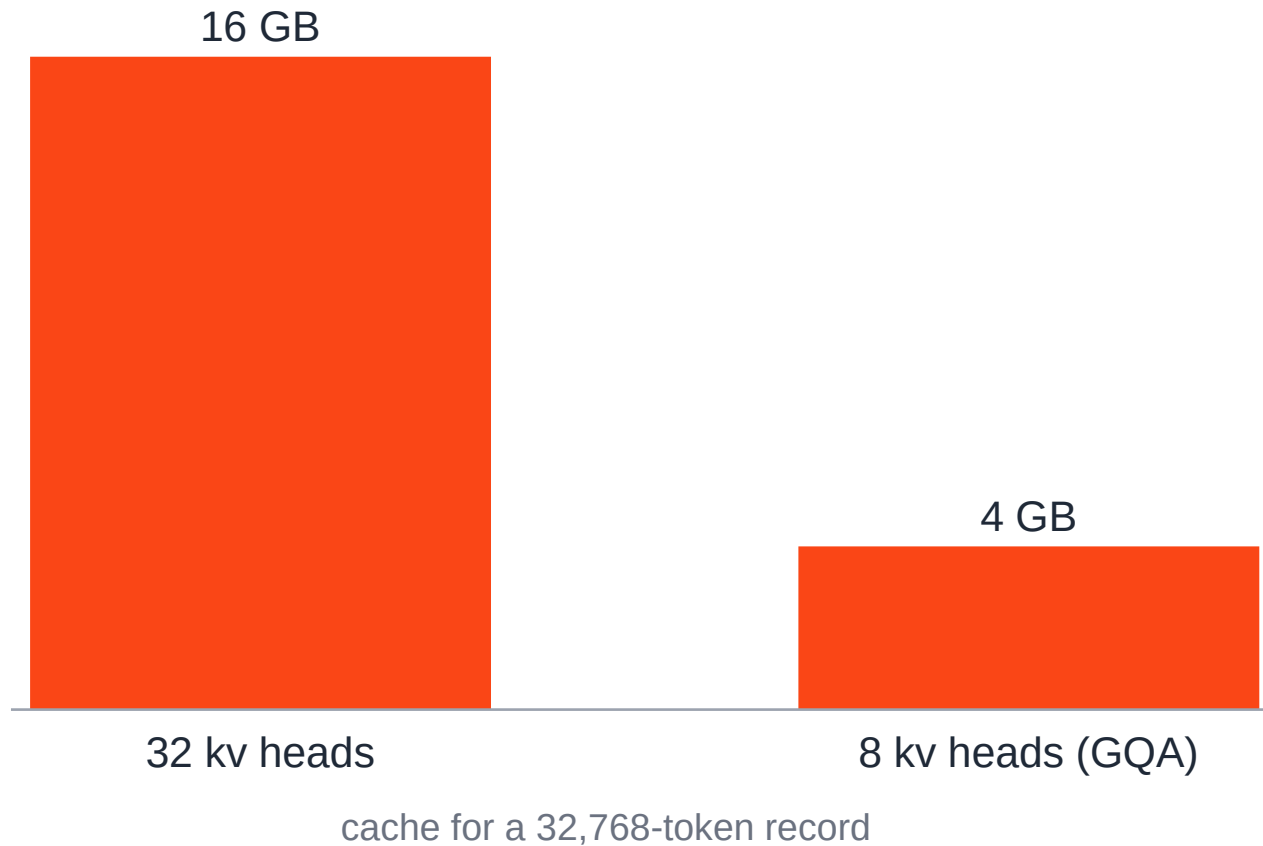


KV cache for one record, 32 key-value heads, 16-bit

for comparison
the model weights
in 16 bits:
16 GB

Fix 1: Grouped-Query Attention

Store 8 key-value heads instead of 32: the cache shrinks 4 times.



Llama 3 8B

32 query heads

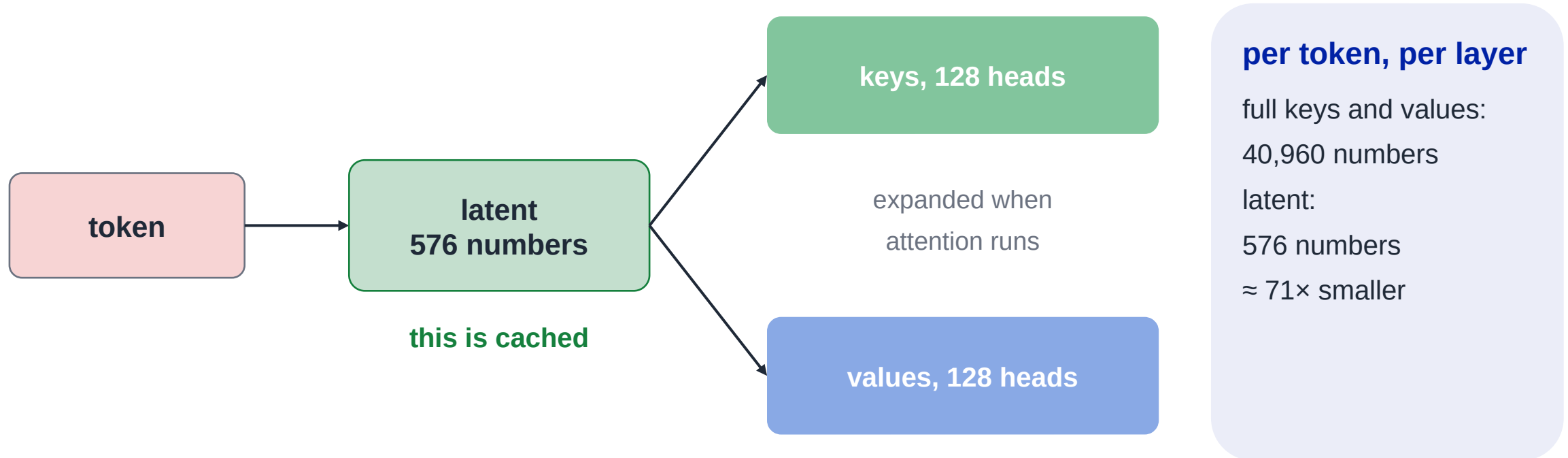
8 key-value heads

4 query heads per group

0.125 MB per token

Fix 2: Multi-Head Latent Attention (DeepSeek-V3)

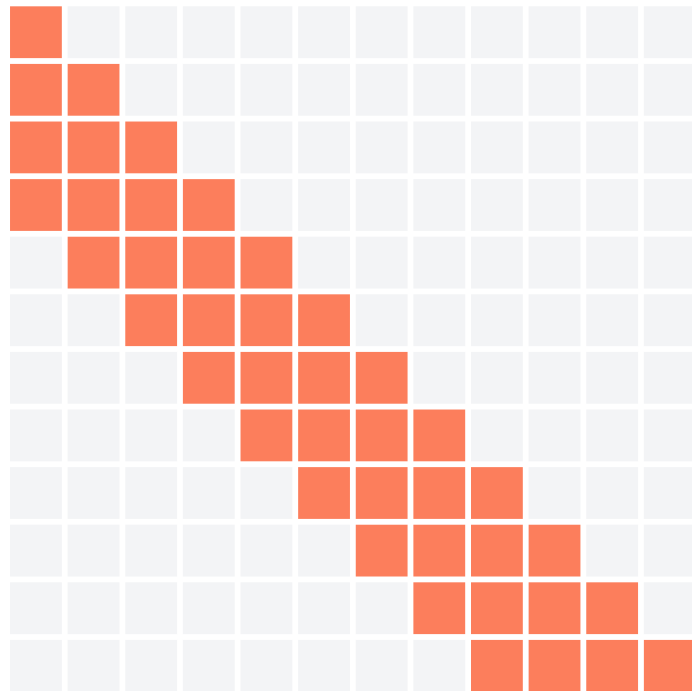
Cache one small compressed vector per token, and expand it when needed.



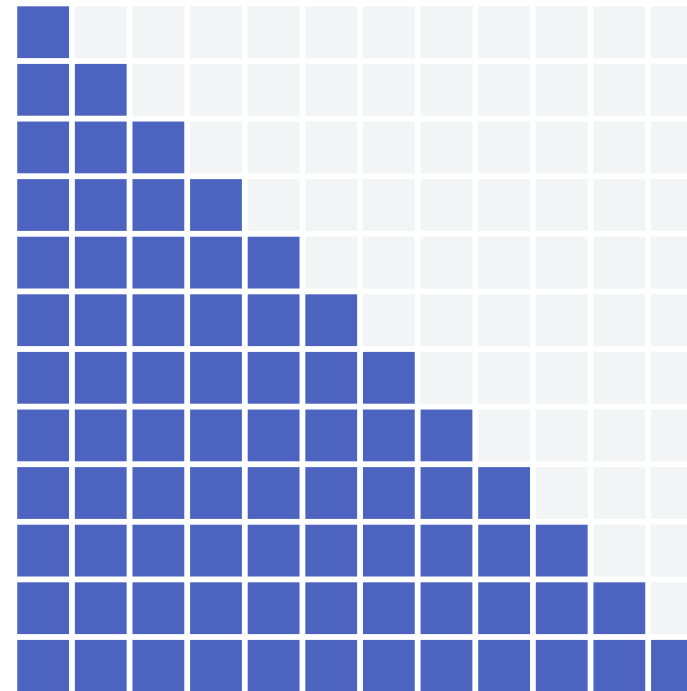
Fix 3: Sliding-Window Attention

Most layers look only at recent tokens; a few layers see everything.

local layer: last 4 tokens



global layer: all earlier tokens



Gemma 3

5 local layers
per global layer
window:
1,024 tokens

Animation: The KV Cache and How Models Shrink It

Animation clip: week04-clip2-kv-cache.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

2026: Sparse, Compressed, and Linear Attention

The newest models attack the same memory problem from more directions.

Sparse attention

DeepSeek-V3.2: each query attends only to the top-scoring earlier tokens

Compressed attention

DeepSeek-V4: compresses the cache along the sequence; about 10% of V3.2's cache at 1M tokens

Linear attention hybrids

Qwen3.5: 3 linear layers (Gated DeltaNet) for each full attention layer

Sharing the cache across layers

Gemma 4 small models reuse keys and values from earlier layers: about half the cache

The common goal

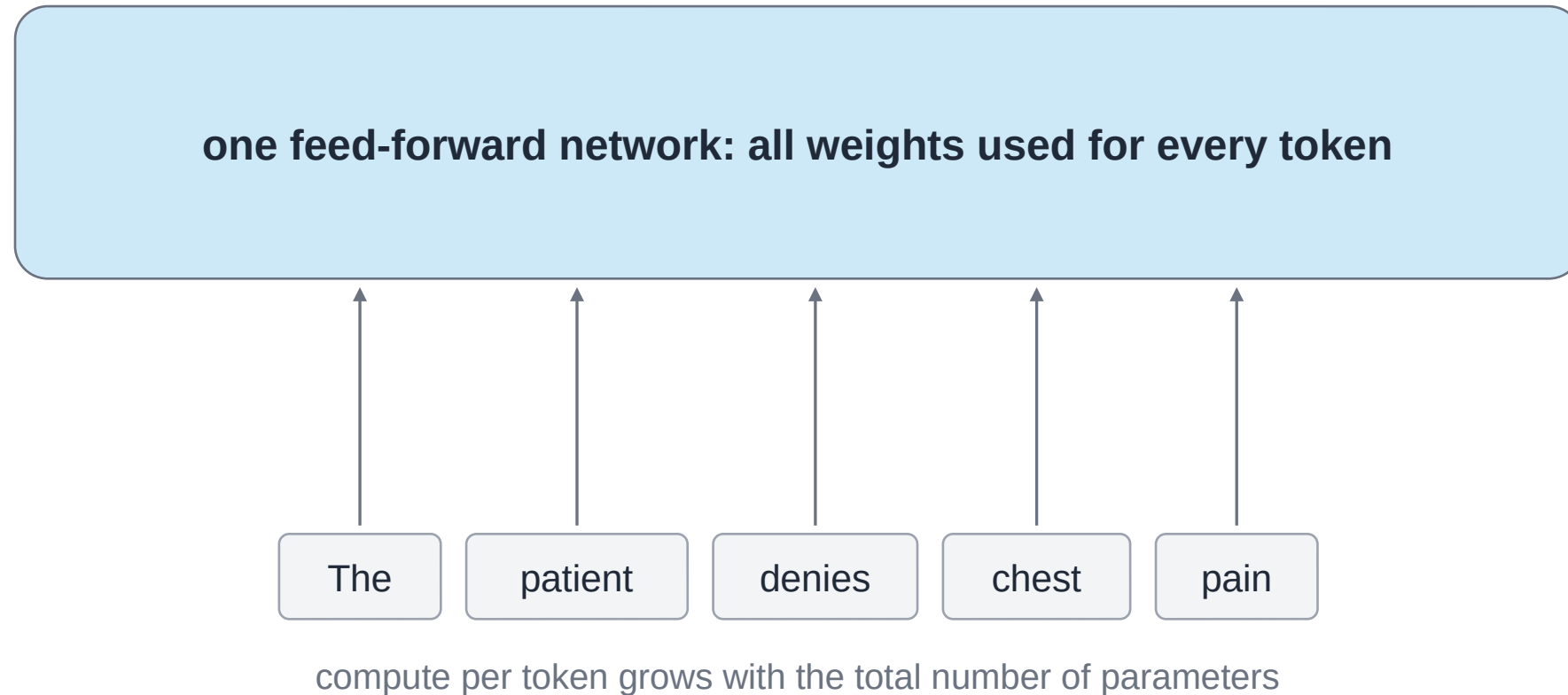
read a whole patient record (or a codebase) without running out of memory



Mixture of Experts

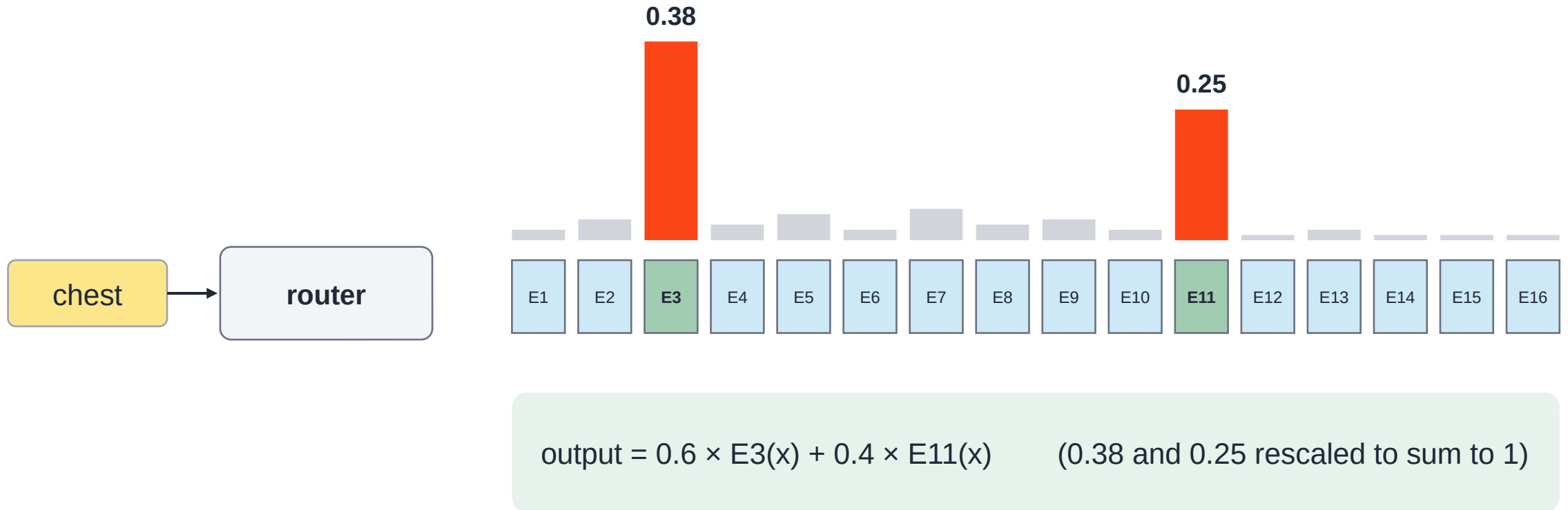
Dense: Every Weight for Every Token

In a dense model, every token passes through the same feed-forward network.



Mixture of Experts: A Router Picks Two

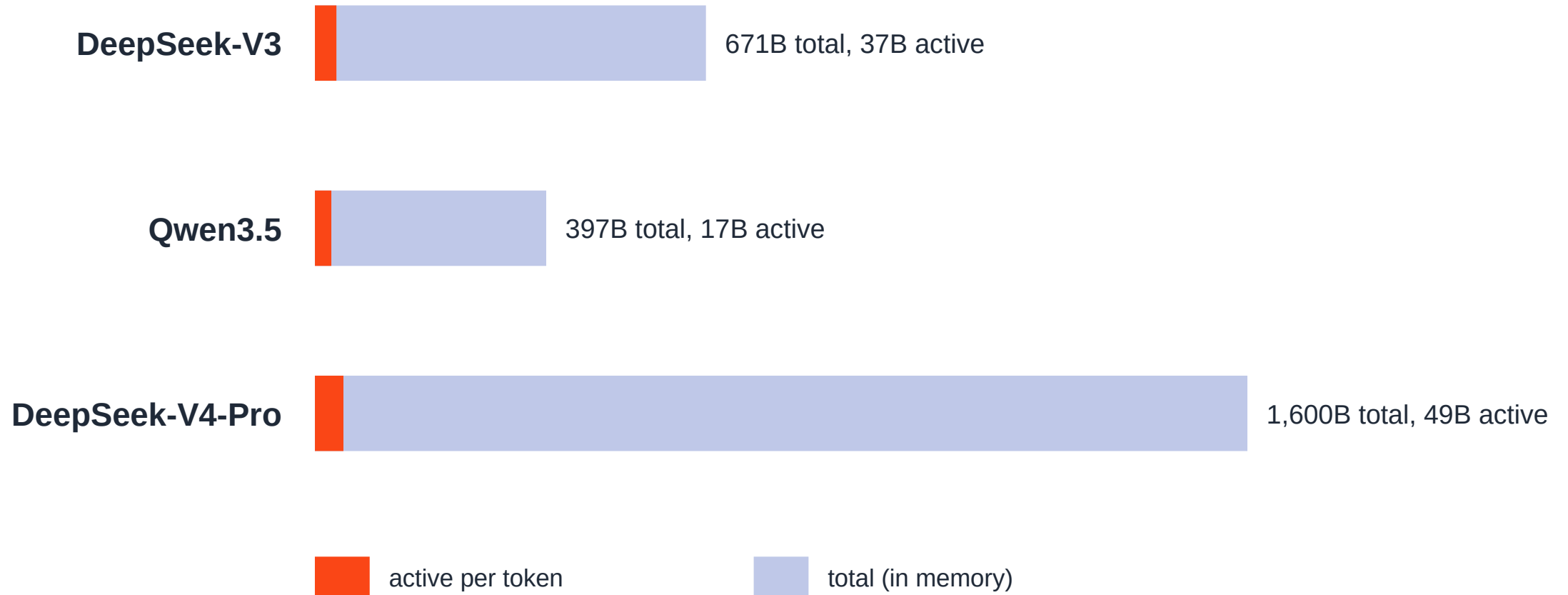
A router scores the experts for each token; only the top two run.



16 experts and top-2 routing are illustrative

Total Parameters versus Active Parameters

Memory scales with the total; compute per token scales with the active part.



Animation: Mixture of Experts

Animation clip: week04-clip3-mixture-of-experts.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

What Mixture of Experts Means for a Hospital

What gets cheaper

compute per token, so faster answers
and more patients per GPU

What does not

memory: every expert must be loaded,
even if most are idle for a given token

Question: a dense 8B model or a 26B mixture of experts with 4B active, for a discharge-summary tool on local GPUs?



More 2026 Changes

Bigger Vocabularies: Fewer Tokens per Note

The same clinical sentence: 21 tokens for GPT-2, 14 for DeepSeek-V3.

“Patient denies chest pain; troponin negative; ECG shows sinus tachycardia.”

GPT-2 (2019): vocabulary 50,257 → 21 tokens

Pat	ient	denies	chest	pain	;	trop	on	in	negative	;	EC	G	shows	sin	us	t	achy
card	ia	.															

DeepSeek-V3 (2024): vocabulary about 129,000 → 14 tokens

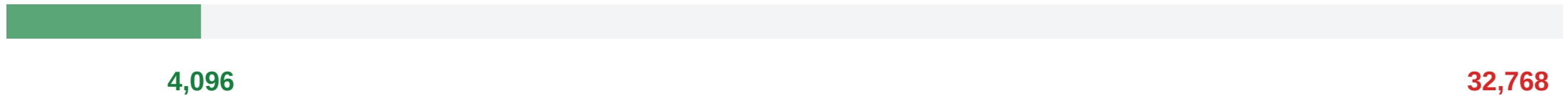
Patient	denies	chest	pain	;	trop	onin	negative	;	ECG	shows	sinus	tachycardia	.
---------	--------	-------	------	---	------	------	----------	---	-----	-------	-------	-------------	---

“tachycardia”: 4 pieces in GPT-2, 1 token in DeepSeek-V3

Longer Context: Stretch the RoPE Angles

Train on short text, then rotate more slowly so long positions look familiar.

angles seen in training



Before: position 32,768 → an angle never seen in training

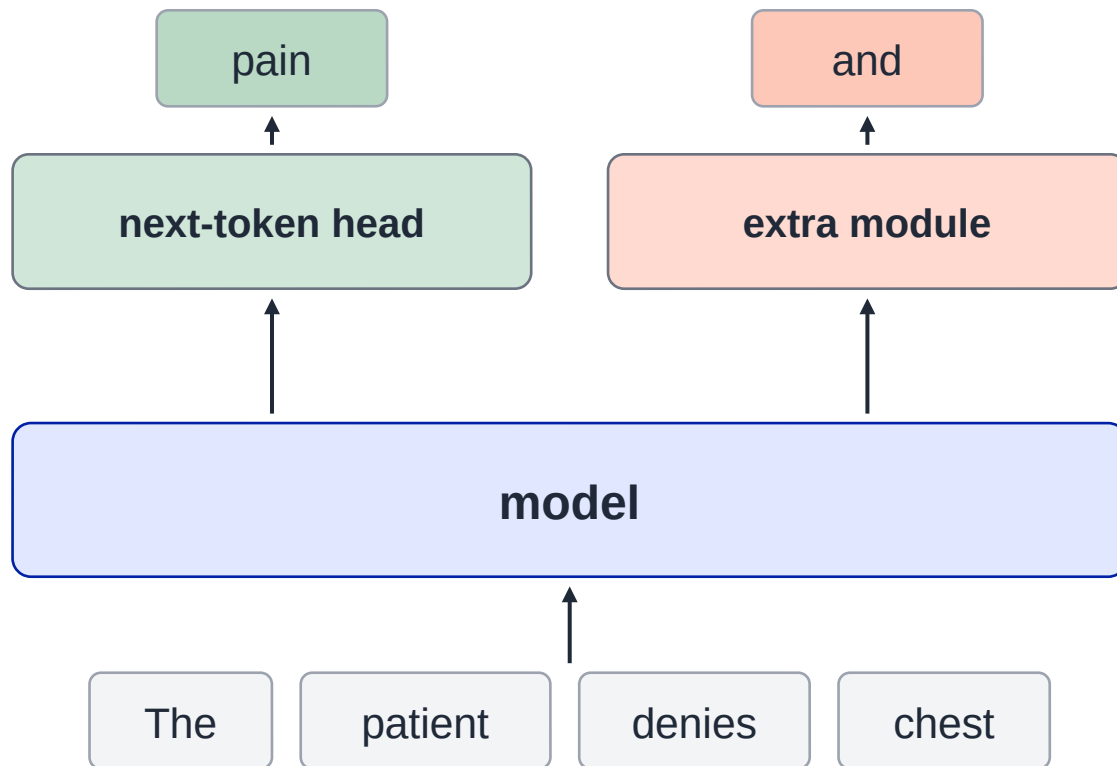
Stretch 8×: position 32,768 → the angle that position 4,096 had



DeepSeek-V3 context extension (YaRN)

Multi-Token Prediction

During training, an extra module also predicts the token after next.



Why

two training signals per position
in DeepSeek-V3: one extra module
at generation: draft a token
ahead, then verify it

Animation: More 2026 Changes

Animation clip: week04-clip4-more-changes.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

Putting It Together: MedGemma

Thursday's paper: an open medical model built from today's pieces.

Base	Gemma 3
Block	RMSNorm · gated FFN · RoPE · grouped-query attention
Long input	5 sliding-window layers (1,024 tokens) per global layer
Vocabulary	about 262,000 tokens
Images	MedSigLIP encoder (4B multimodal version)
Added training	medical text and medical images

Recap: Inside Modern LLMs

The block

RMSNorm before each sublayer, SwiGLU, RoPE inside attention, grouped-query attention.

The cache

Memory grows with every token; grouped-query, latent, and sliding-window attention shrink it.

Experts

Mixture of experts: compute follows the active parameters, memory the total.

Reach

Bigger vocabularies and stretched RoPE angles let a model read a whole record.

Before Thursday

Read

DeepSeek-AI, DeepSeek-V3 Technical Report (2024) and Sellergren et al., MedGemma Technical Report (2025).

Write

Reading responses on Canvas before 1:00 PM Thursday.

Explore

Week 4 reference notebook (ungraded): opens Qwen3-0.6B and a small mixture of experts and reproduces every number in the clips.

Next week

Post-training and reasoning: from pretrained model to clinical assistant.