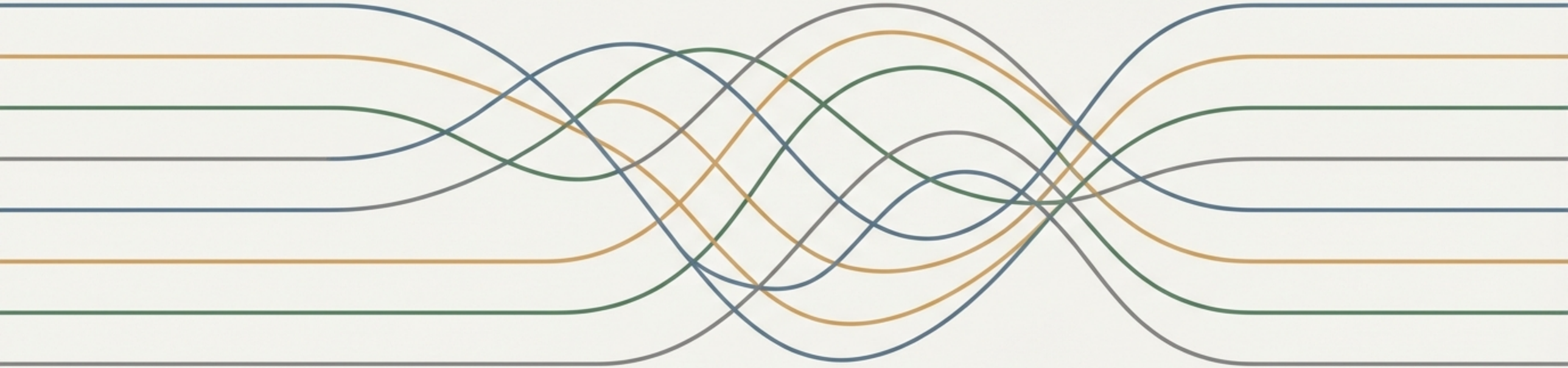
The Breakthrough That Taught AI to Understand Context

Unpacking Multi-Head Attention



The Challenge: Processing Language in Order

For years, models like Recurrent Neural Networks (RNNs) read sentences one word at a time, like a person reading a book.

Inherently Sequential

This creates two major problems:

1. **It's slow.** No parallelization within a sentence.
2. **Long-range memory is hard.** The meaning of the first word can get lost by the time the model reaches the last.

Understanding a sentence requires seeing the whole picture.

The **cat**, which was happily chasing **all** the dogs down the street, was **finally** **tired**.

How does the model remember the subject (‘cat’) after so many intervening words?

A New Approach: What if words could talk to each other?

Self-attention allows every word in a sequence to look at every *_other_* word simultaneously.

Highly parallelizable (`O(1)` sequential operations)

Direct path between any two words (`O(1)` path length)

The Mechanism: A Query System for Words

Think of it like a vector database query. For each word, we create three vectors:

Query (Q)

“What I am looking for.”

Key (K)

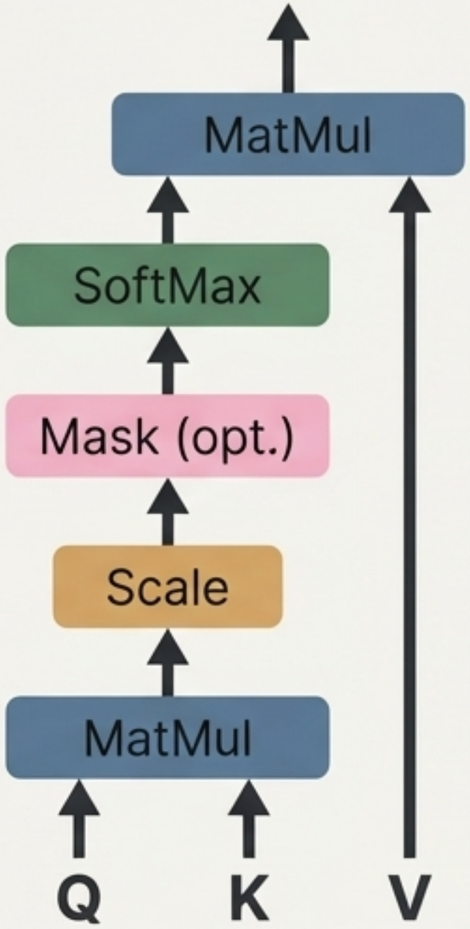
“What I have.”

Value (V)

“What I will give you.”

A word’s **Query** vector is compared against every other word’s **Key** vector to find a **compatibility score**. This score determines how much of each word’s **Value** vector should be used to update the original word’s representation.

Scaled Dot-Product Attention in Action



$$QK^T$$

Multiply Queries by Keys for similarity scores.

$$QK^T / \frac{QK^T}{\sqrt{d_k}}$$

Scale to stabilize training.

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

Convert scores to weights that sum to 1.

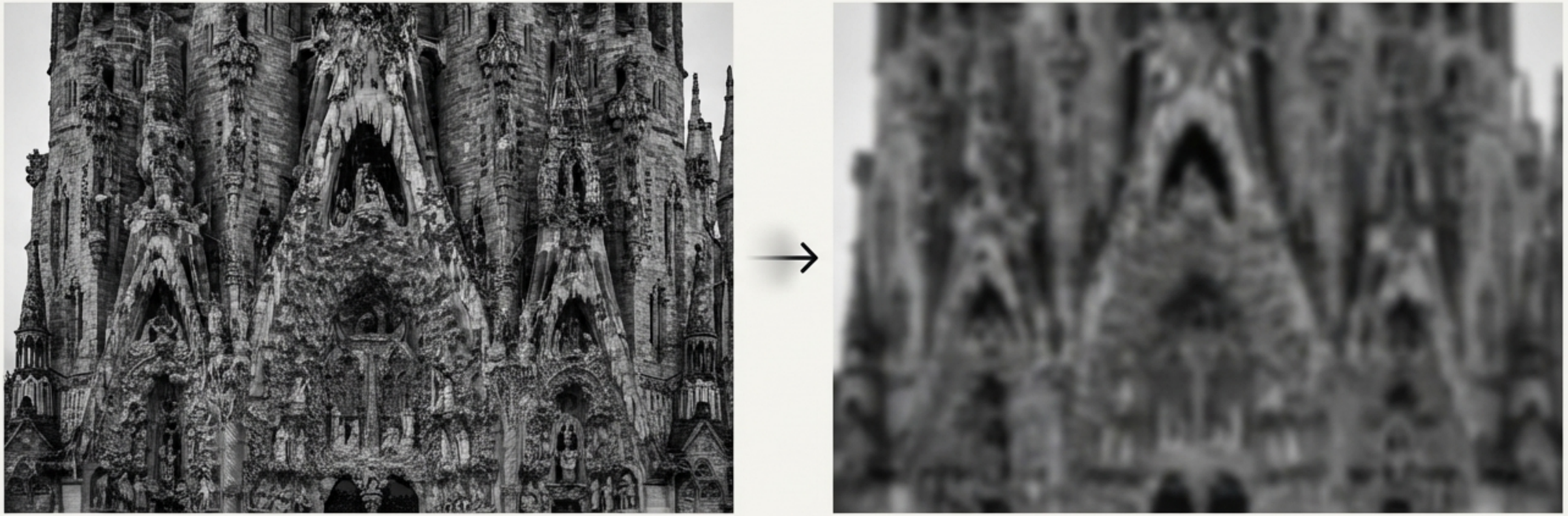
$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multiply weights by **Values** for a new representation.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

But a single perspective can be blurry.

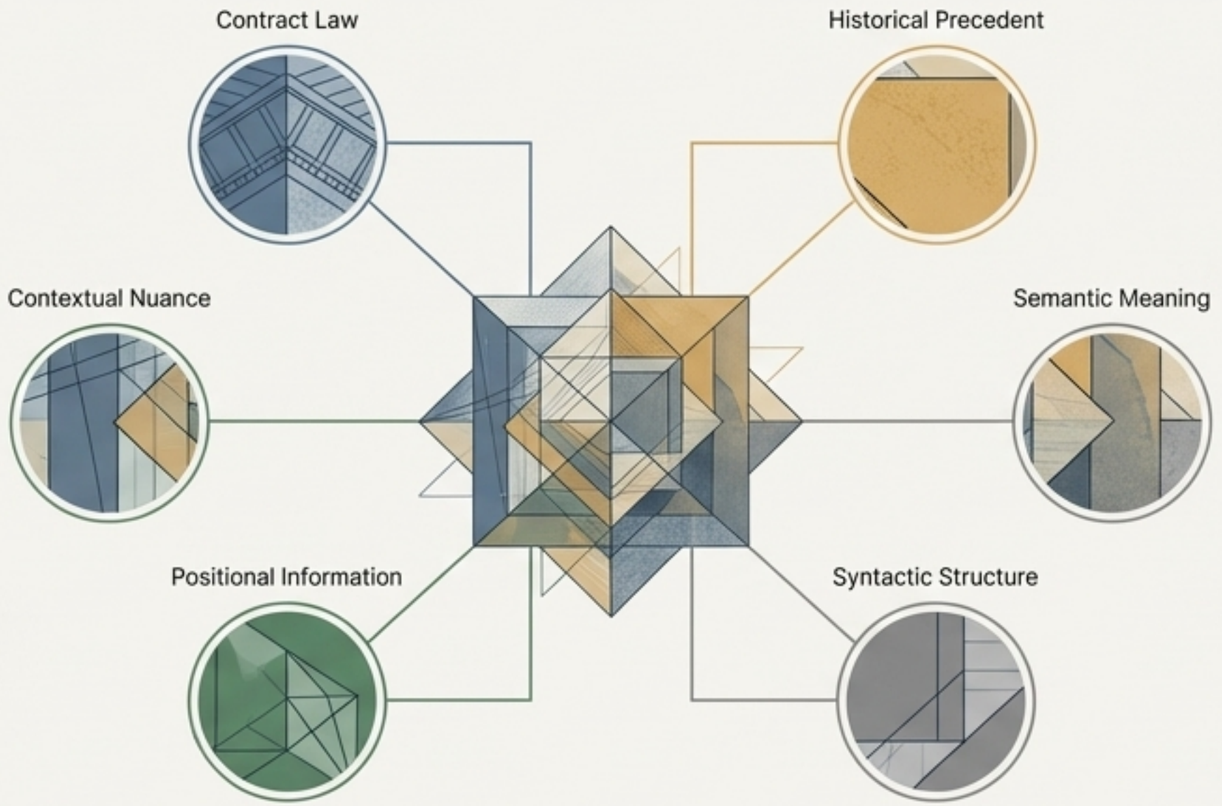
“Imagine trying to understand a complex legal document by asking a single lawyer for a one-sentence summary. You get the main idea, but you lose all the nuance.”



With a single attention head, the model has to average all the different types of relationships (syntactic, semantic, positional) into a single representation.

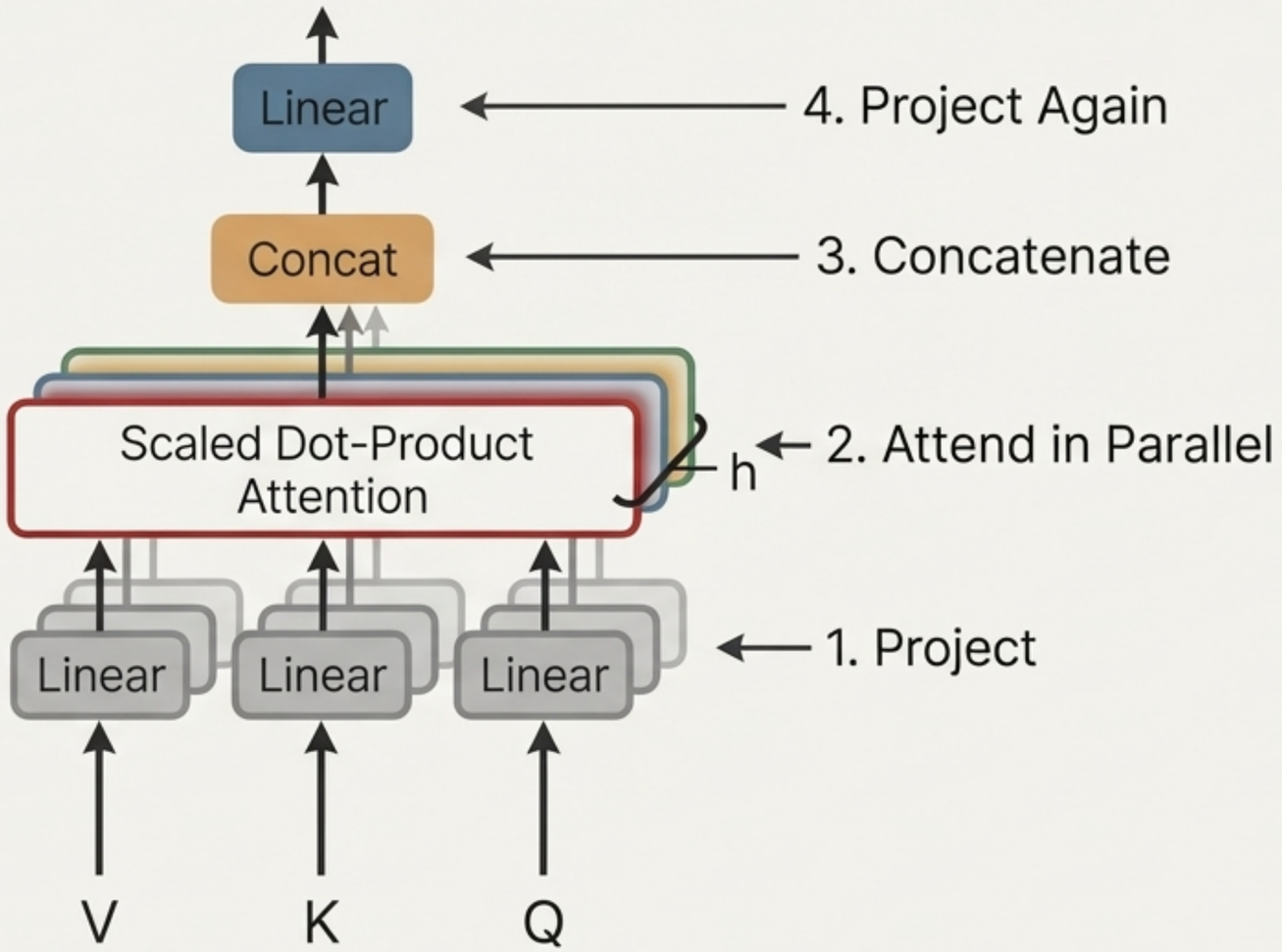
The Breakthrough: A Committee of Specialists

"Now, imagine you give that document to a committee of eight specialists. One is an expert in contract law, another in historical precedent... By combining their reports, you gain a deep, multi-faceted understanding."



This is **Multi-Head Attention**. Instead of one attention calculation, we run multiple in parallel, each focusing on a different aspect of the language.

The Anatomy of Multi-Head Attention



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

Seeing the Specialists at Work

The authors of “Attention Is All You Need” visualized the attention patterns of different heads and found they learn to perform distinct, interpretable tasks.

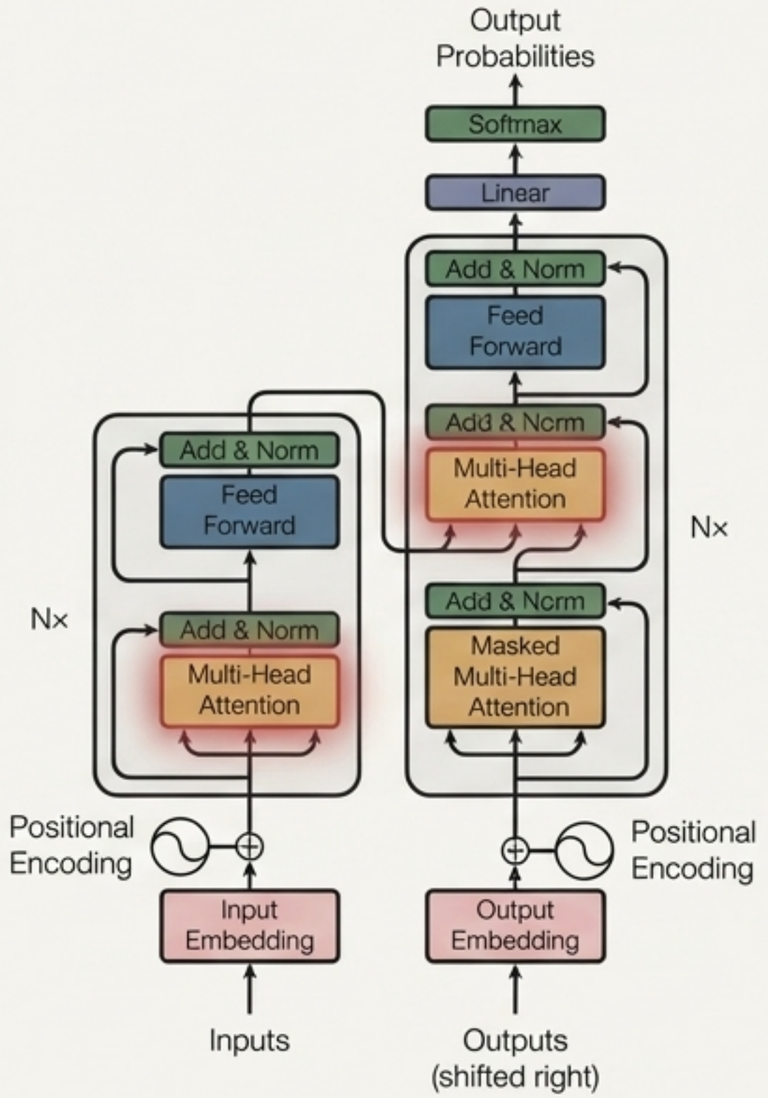
Tracking long-distance dependencies

...**making** the registration or voting process **more difficult.**

Resolving pronouns

The Law... but **its** application should be just...

The Heart of the Transformer



The Transformer... [relies] entirely on an attention mechanism to draw global dependencies between input and output, dispensing with recurrence and convolutions entirely.

A New State of the Art

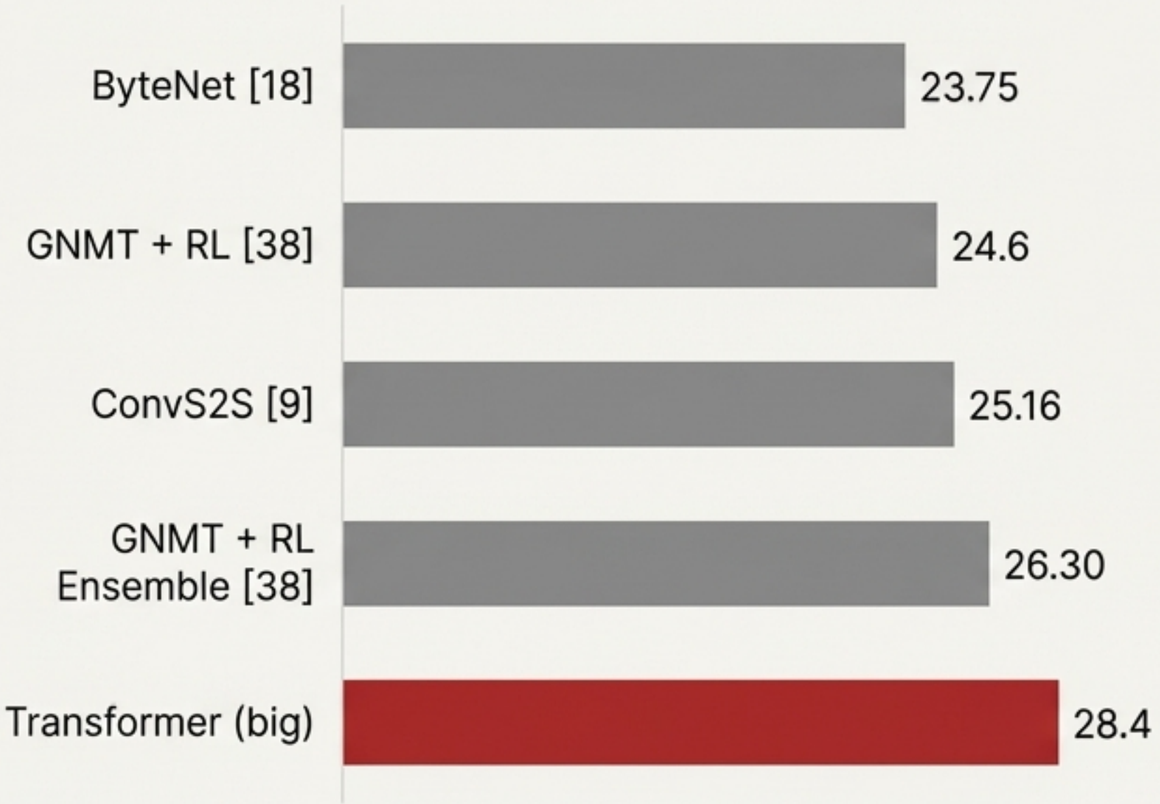
Quality

Achieved a new state-of-the-art score of **28.4 BLEU** on the WMT 2014 English-to-German task, over 2.0 BLEU higher than any previous model.

Efficiency

More parallelizable and required **significantly less time to train** (3.5 days on 8 GPUs for the big model).

WMT 2014 English-to-German (BLEU Score)



More Than a Model, A New Foundation

The principles of multi-headed self-attention did not just create a better translation model. They provided the architectural foundation for the large language models—like GPT and BERT—that define artificial intelligence today.