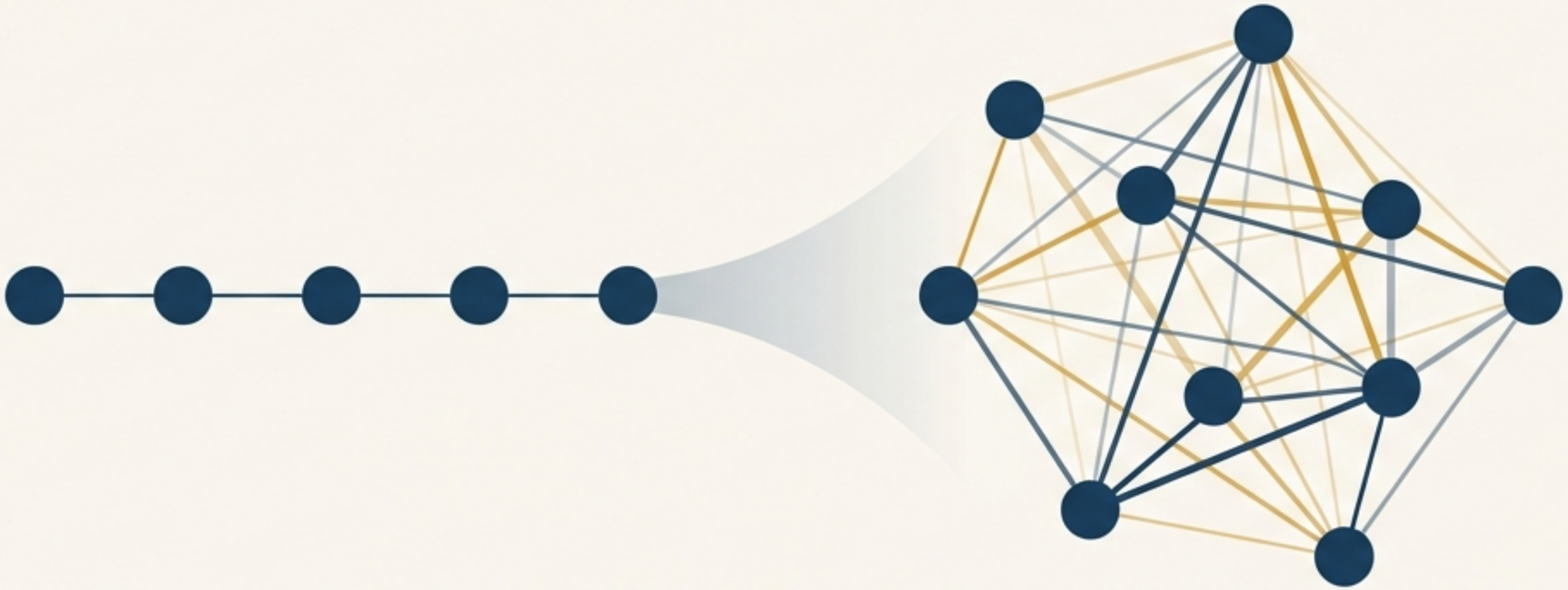
Attention Is All You Need

A New Paradigm for Sequence Transduction Based Solely on Attention

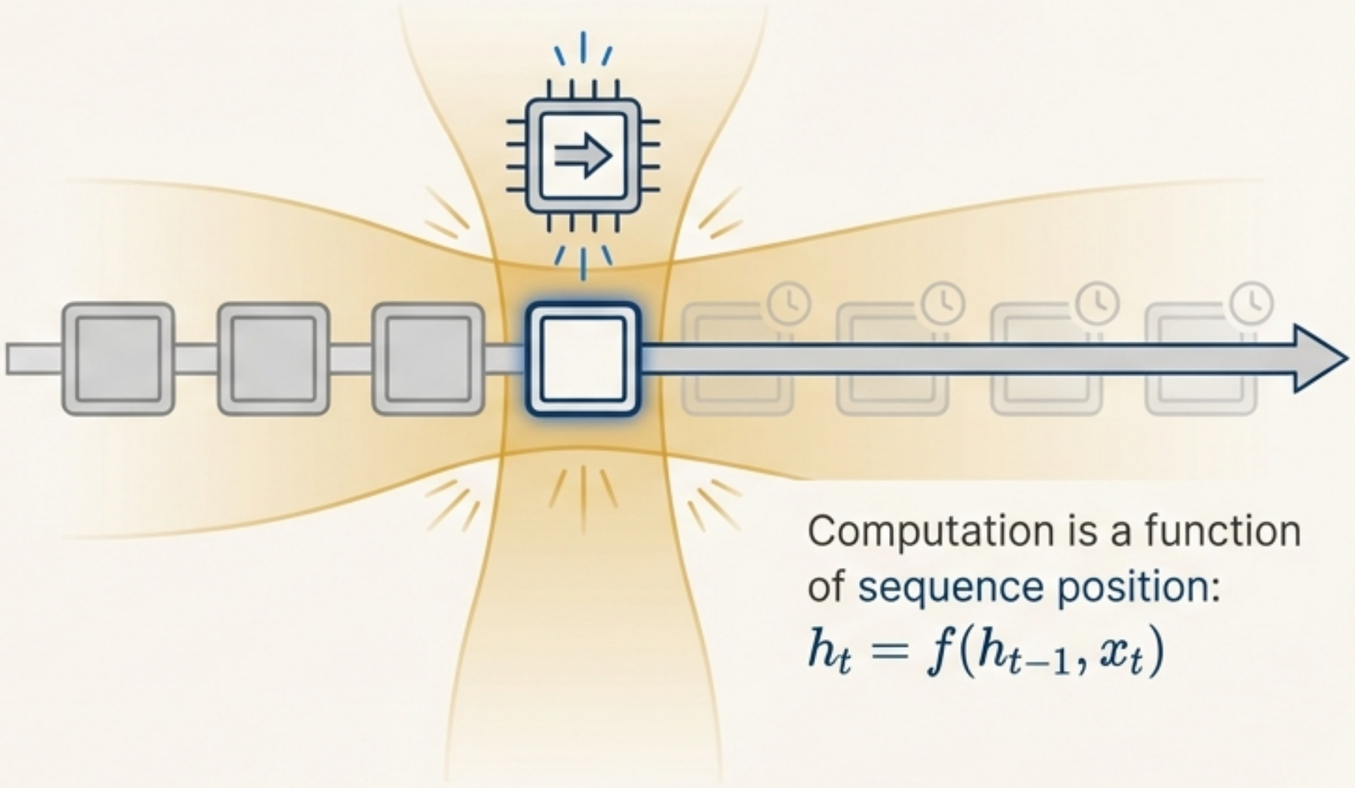


Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,
Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

31st Conference on Neural Information Processing Systems (NIPS 2017)

The Sequential Bottleneck of Recurrent Models

- Dominant architectures (RNNs, LSTMs, GRUs) process data sequeriially, token by token.
- This inherent sequential nature **precludes parallelization** within training examples.
- Learning **long-range dependencies** is difficult due to long signal paths.
- The fundamental constraint of sequential computation remains, limiting performance and scalability.

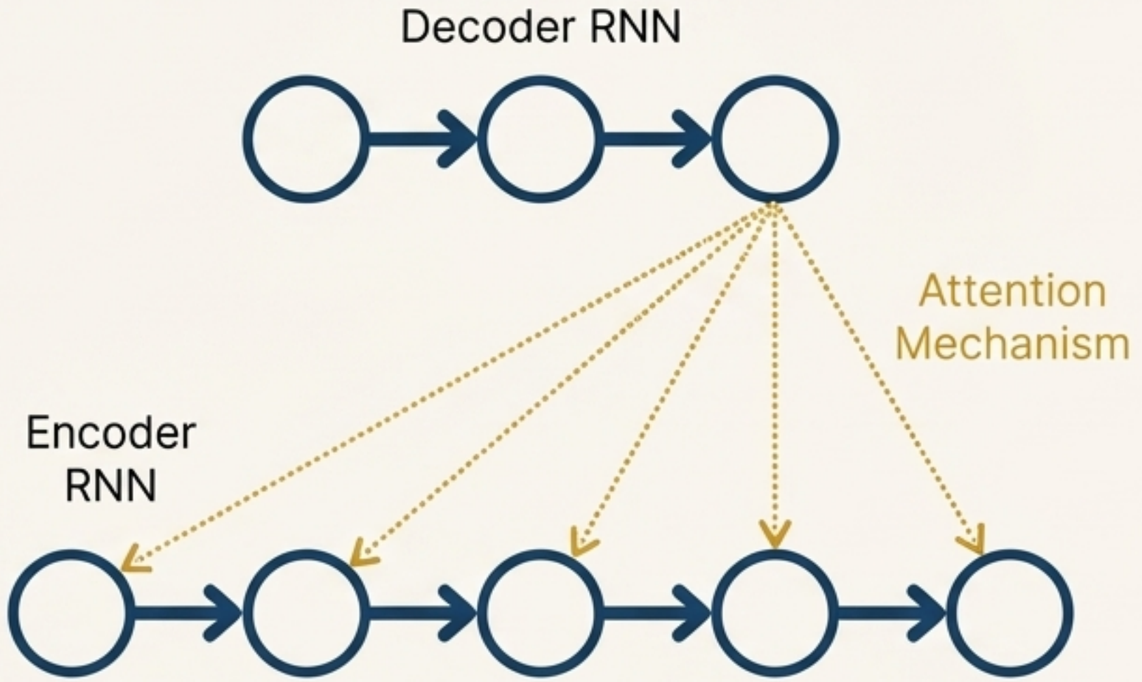


Computation is a function of sequence position:

$$h_t = f(h_{t-1}, x_t)$$

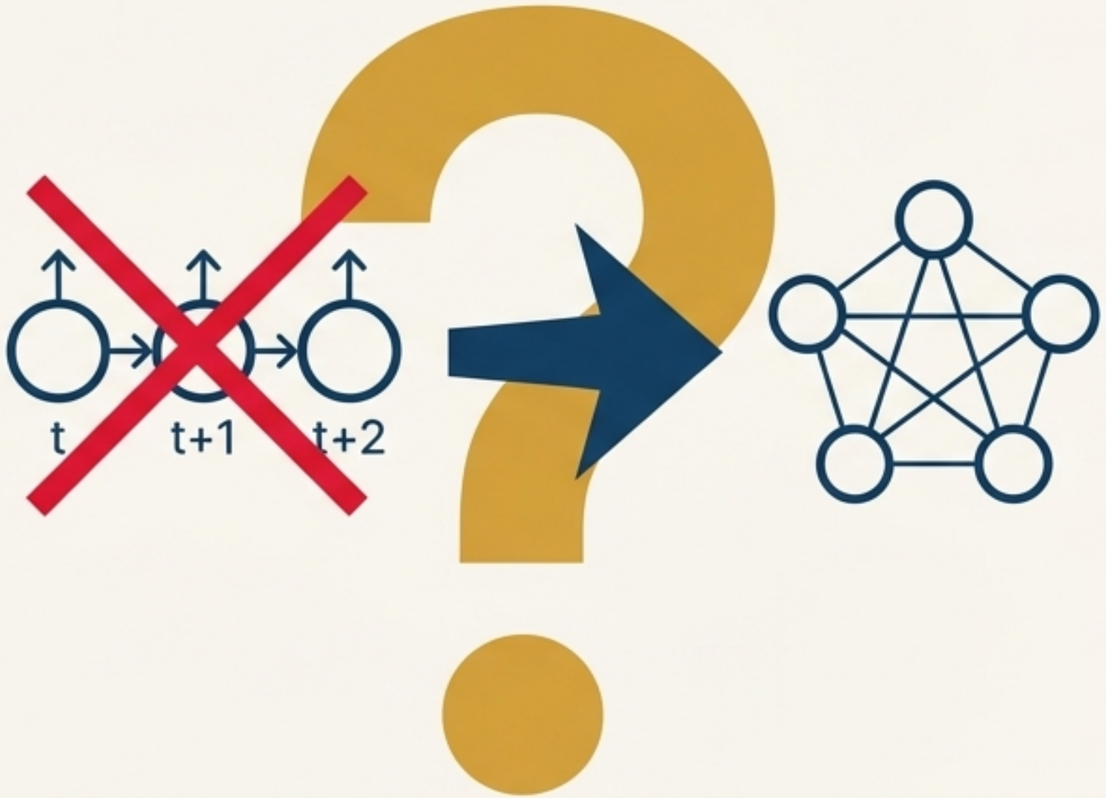
Attention Was an Add-on, Not the Main Engine

- Attention mechanisms were created to help RNNs model dependencies regardless of distance.
- They allowed a model to selectively focus on relevant parts of the input sequence.
- However, attention was almost always used **in conjunction with a recurrent network**.
- The core constraint of sequential computation was never resolved.



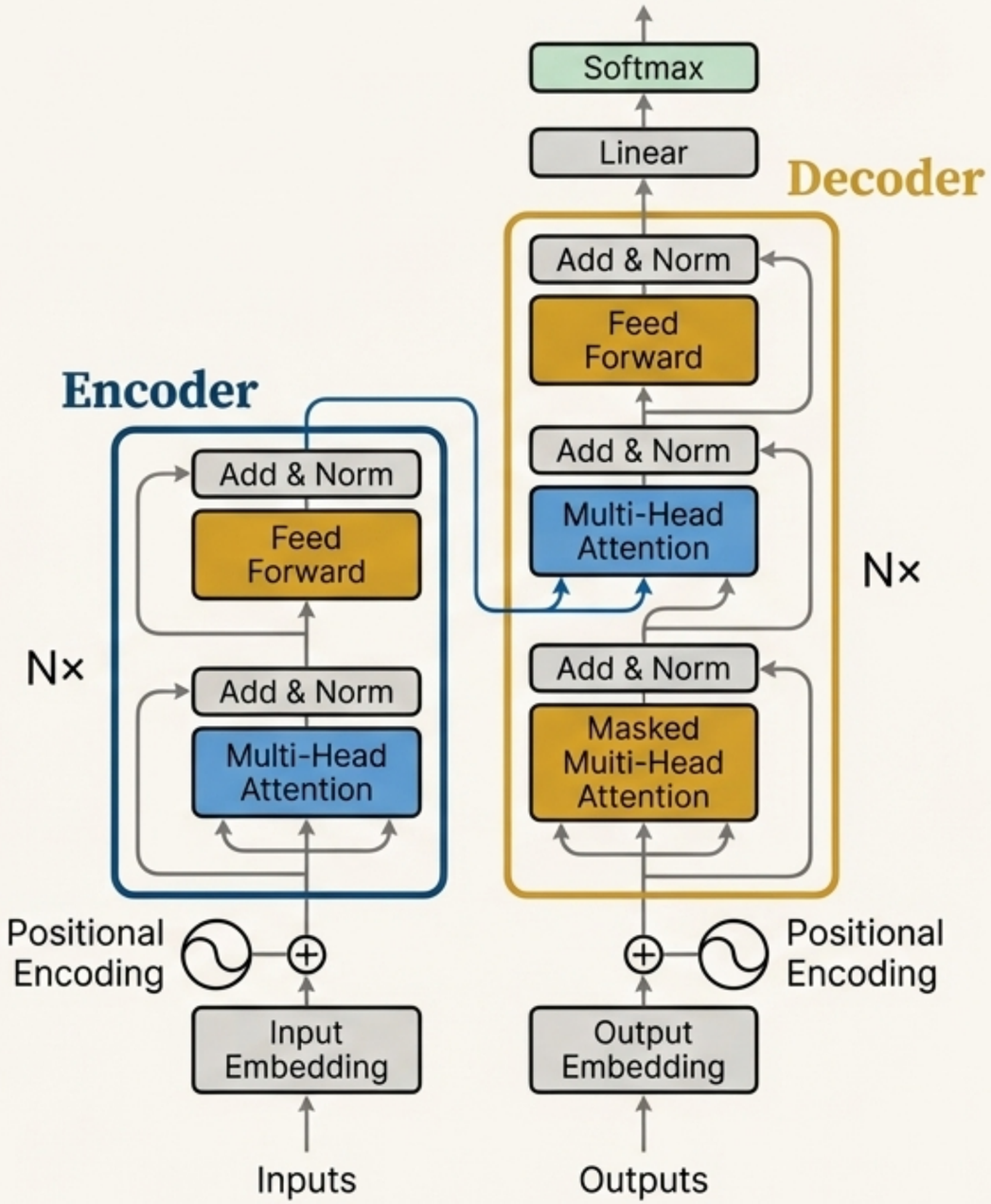
Can We Build a High-Performance Model Using *Only* Attention?

- **Hypothesis:** Recurrence and convolution are not essential for state-of-the-art sequence transduction.
- **Research Question:** Can a model based *solely* on attention mechanisms match or exceed the performance of the best recurrent models?
- **The Goal:** Eliminate the sequential bottleneck to unlock massive parallelization and directly model global dependencies with constant path length.



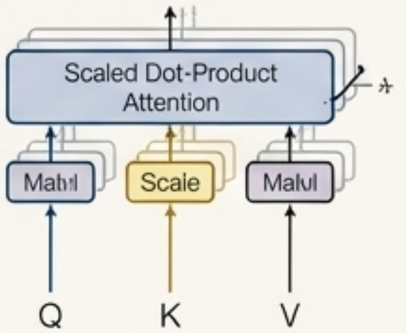
The Transformer: An Architecture Built on Self-Attention

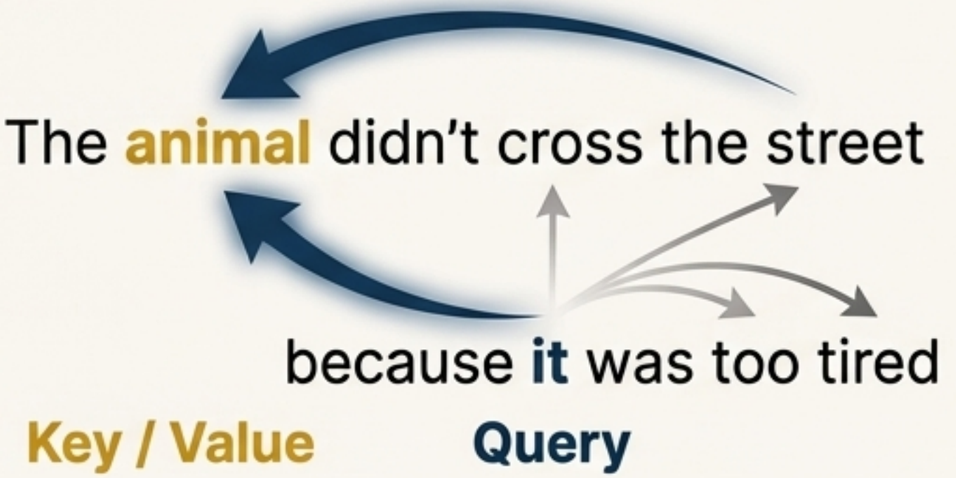
- A standard **Encoder-Decoder** structure for transduction tasks.
- The encoder and decoder are built from stacks of identical layers (N=6 in the base model).
- Each layer contains two primary sub-layers:
  1. **Multi-Head Self-Attention**
  2. **Position-wise Feed-Forward Network**
- Crucially, the model contains **no recurrent or convolutional layers**.



The Core Engine: Calculating Relevance with Queries, Keys, and Values

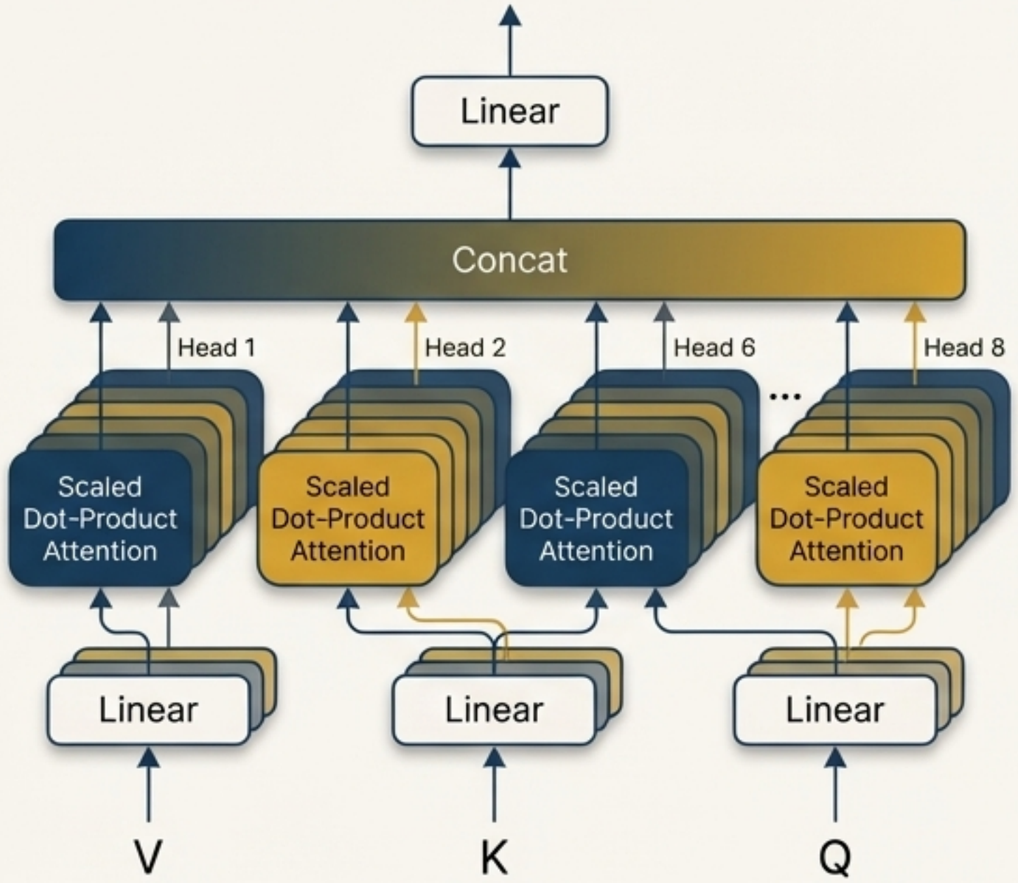
- Self-attention allows the model to weigh the importance of all words in a sequence relative to a single word.
- For each word's input vector, the model creates three new vectors:
  - **Query (Q):** What am I looking for?
  - **Key (K):** What information do I contain?
  - **Value (V):** What information should I provide?
- The model calculates a score between a word's **Query** and every other word's **Key**. These scores become the weights for the **Values**.





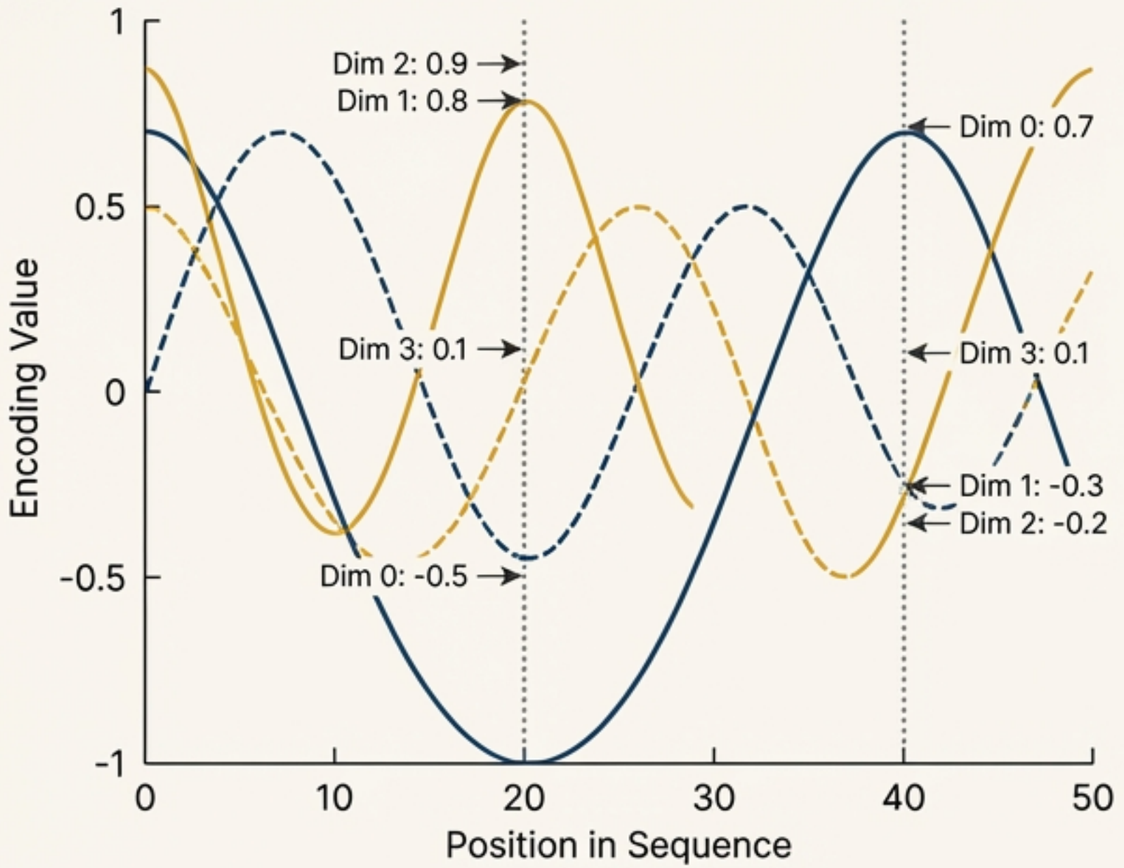
Multi-Head Attention: Analyzing from Multiple Perspectives

- Instead of one attention calculation, Multi-Head Attention runs the process in parallel multiple times (h=8 heads in the paper).
- Each 'head' is a separate projection of the Queries, Keys, and Values.
- This allows the model to jointly attend to information from different representation subspaces at different positions.
- For example, one head might learn syntactic links while another learns semantic relationships.



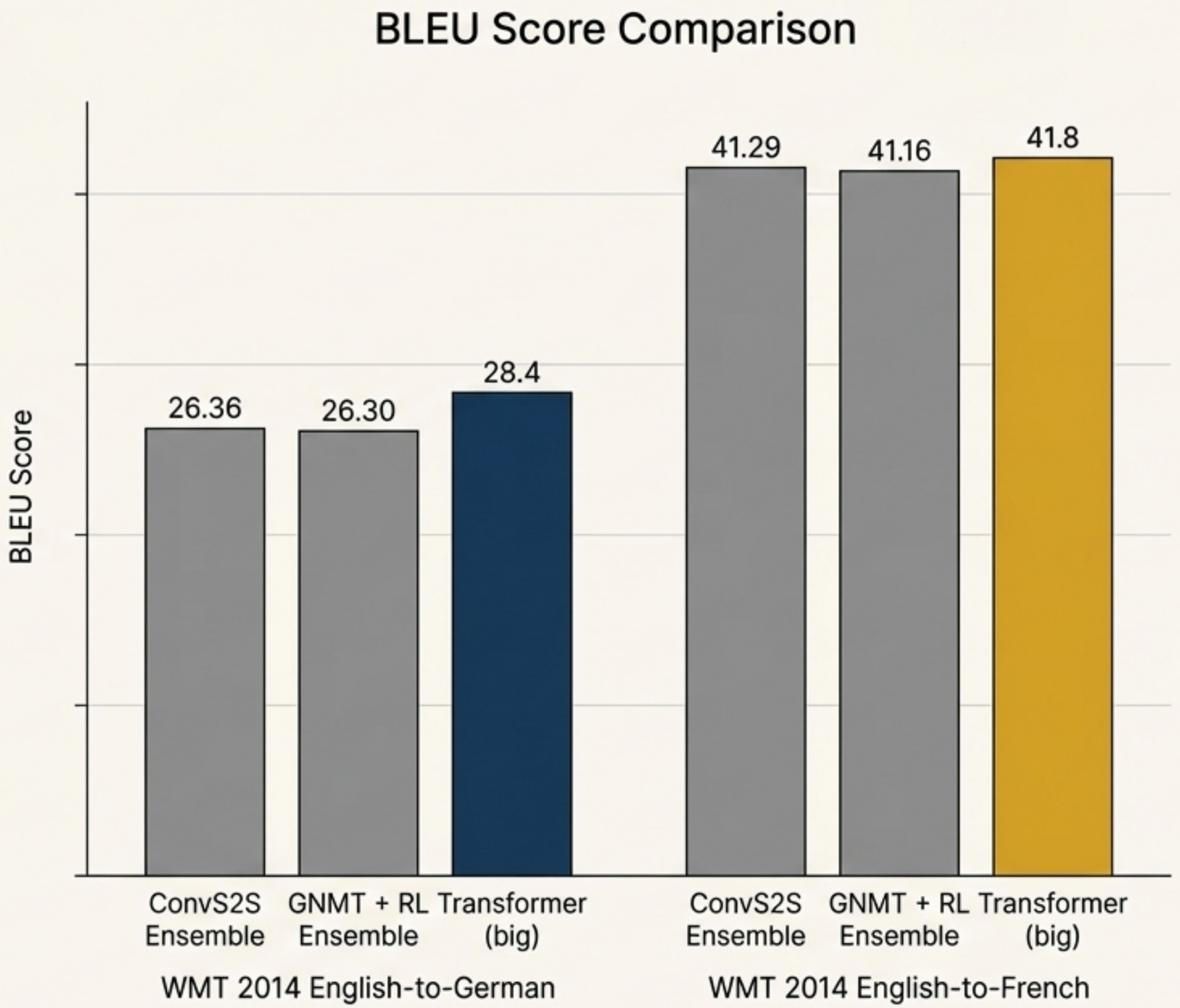
Injecting Word Order with Positional Encodings

- A pure self-attention model is a “bag of words”—it has no inherent sense of word order.
- **Solution**: A “positional encoding” vector is added to each input embedding.
- The paper uses sine and cosine functions of different frequencies to create a unique positional signature for each token.
- This allows the model to learn relative and absolute positions from the input signal itself.



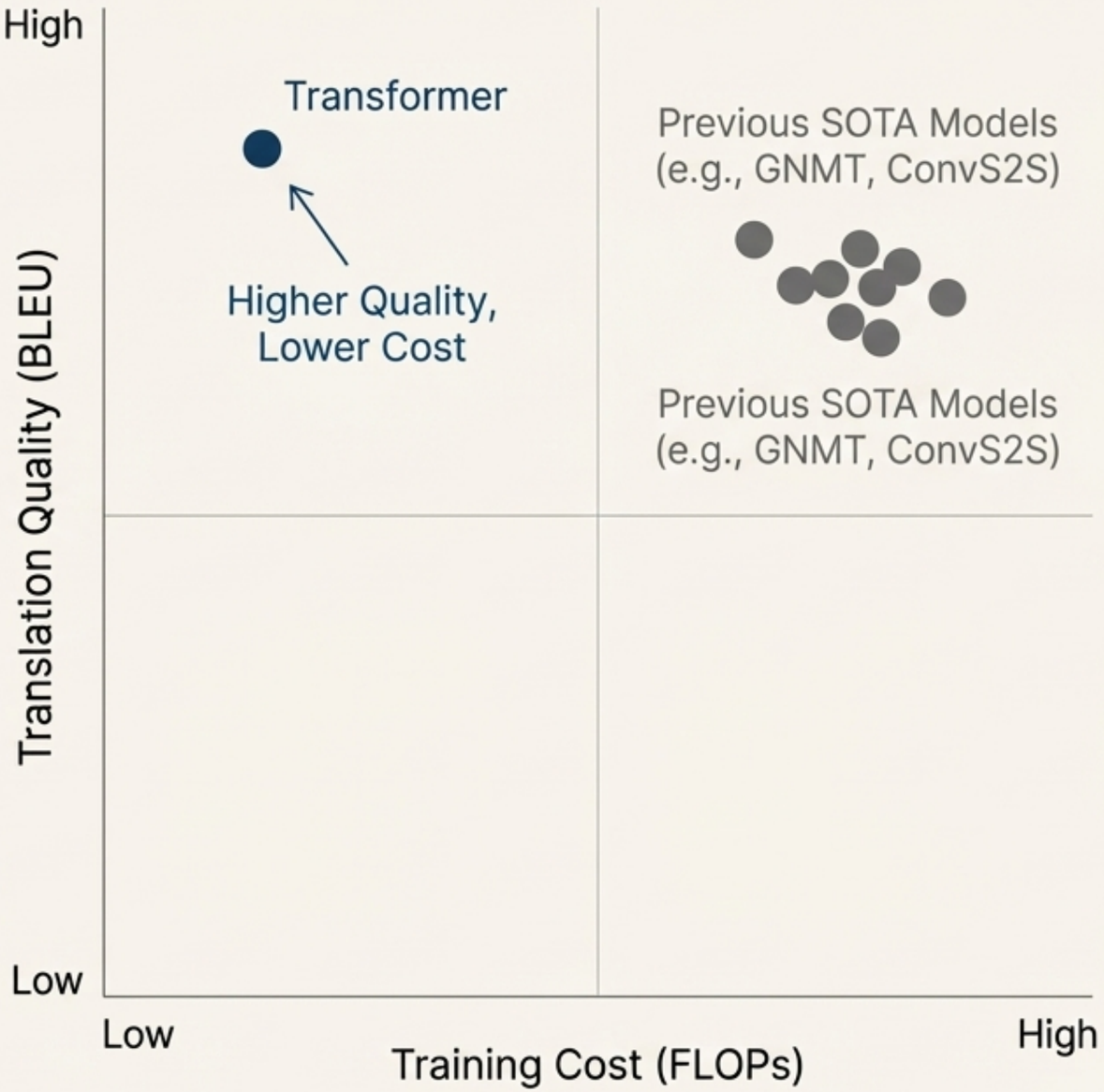
A New State of the Art in Machine Translation

- The Transformer outperformed all previous models on the WMT 2014 English-to-German task.
- Achieved a **28.4 BLEU** score, an improvement of over 2.0 BLEU against the previous best results, including ensembles.
- Established a new single-model record on the English-to-French task with a **41.8 BLEU** score.
- Even the base model surpassed all previous models and ensembles.



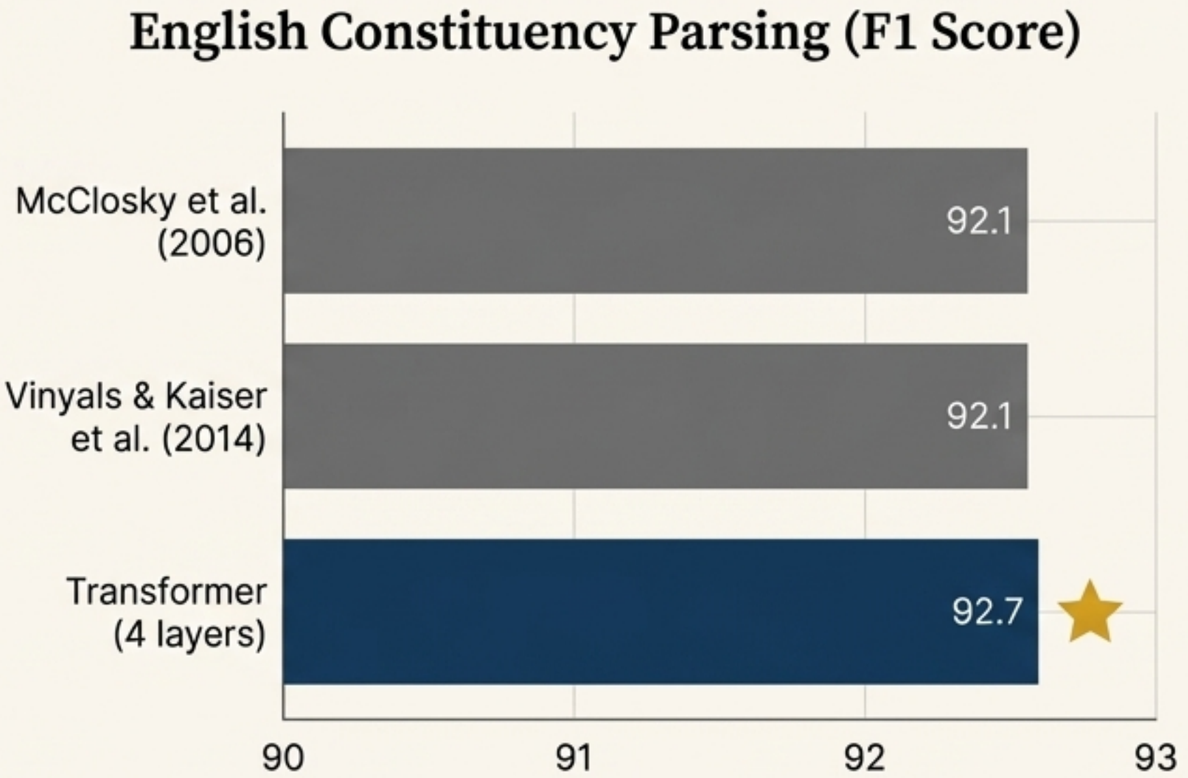
SOTA Performance at a Fraction of the Training Cost

- The parallel architecture is highly efficient on modern hardware (GPUs/TPUs).
- The large model was trained in **just 3.5 days on 8 P100 GPUs**.
- This was a “small fraction of the training costs of the best models from the literature.”
- The training cost (FLOPs) was an order of magnitude lower than top competing models like GNMT.



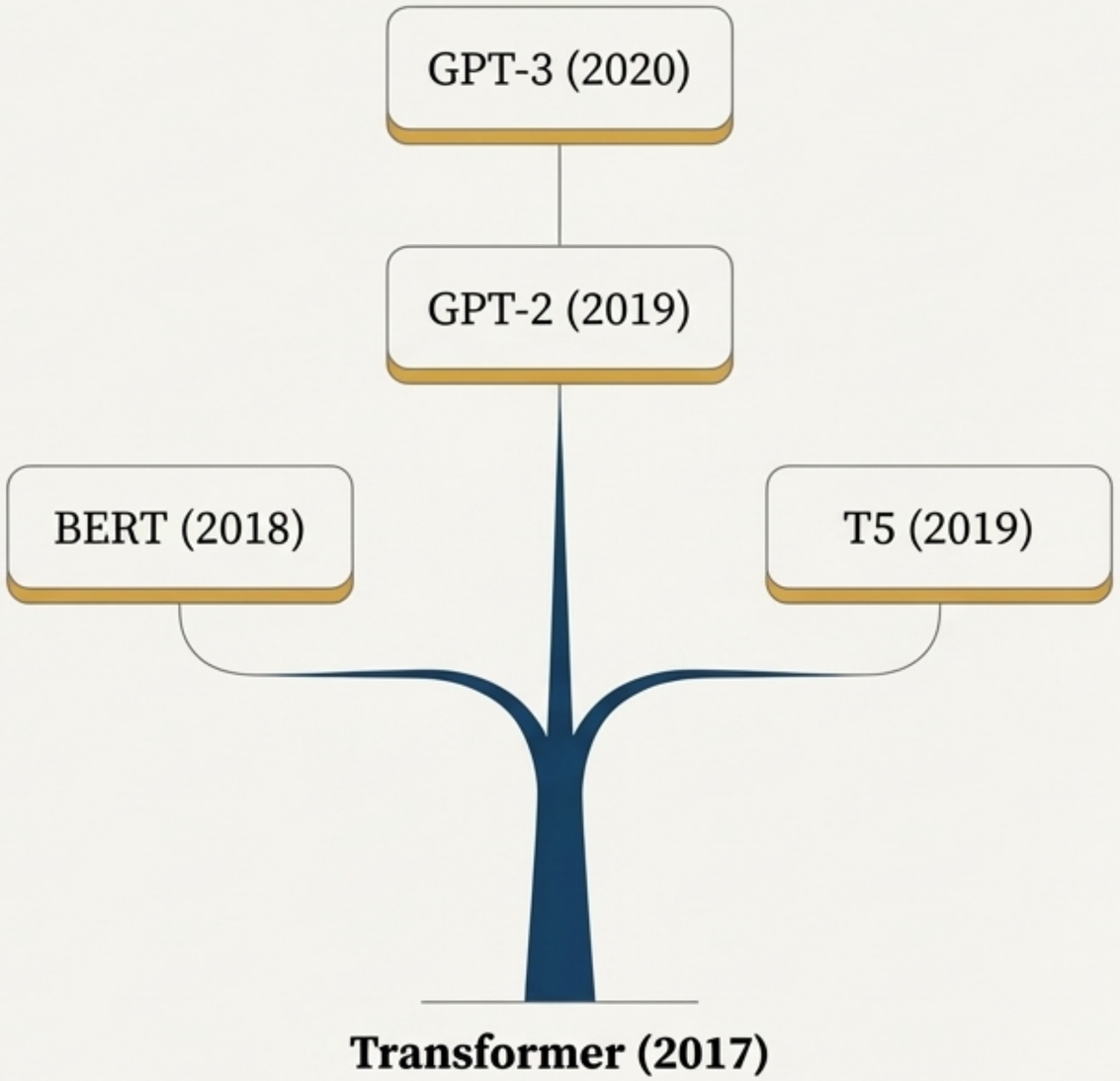
The Architecture Generalizes to Other Complex Tasks

- The model was tested on **English constituency parsing**, a structurally difficult task.
- Despite minimal task-specific tuning, the Transformer performed surprisingly well.
- Achieved an F1 score of **92.7** in a semi-supervised setting, outperforming previous models like the BerkeleyParser.
- It proved its effectiveness even when trained on a small dataset of only 40K sentences.



The Foundation for Modern Large Language Models

- **A Break from Recurrence:** Proved that recurrence is not necessary for top-tier sequence modeling.
- **Unlocking Scale:** The parallelizable architecture enabled the training of massive models that were previously infeasible.
- **The Blueprint for Modern AI:** The Transformer is the foundational building block for nearly all modern LLMs, including BERT, GPT, T5, and beyond.



The Future is Attention-Based

- The paper concluded by proposing several avenues for future research:
  - Applying Transformers to new modalities like **images**, **audio**, and **video**.
  - Developing **restricted attention mechanisms** to handle very long sequences more efficiently.
  - Making the generation process (decoding) **less sequential**.
- **Conclusion**: By replacing recurrence with self-attention, the Transformer provided a more powerful, parallelizable, and scalable foundation for sequence modeling, setting the agenda for years of AI research to come.

Images & Video

Audio

Efficient Attention for Long Sequences