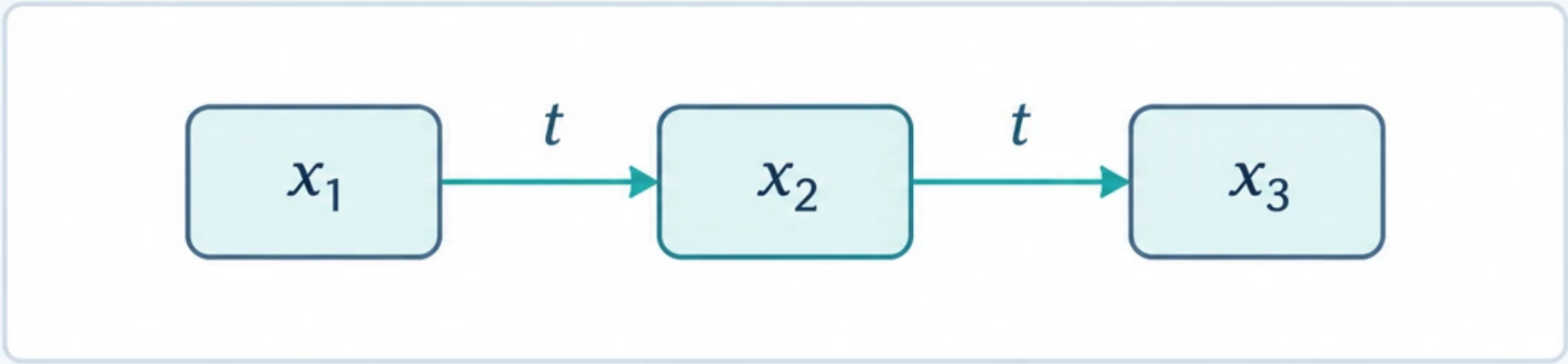
Question 1: Research Problem & Objective

What primary computational limitation of dominant sequence transduction models (RNNs, LSTMs, and CNNs) did Vaswani et al. seek to overcome, and what core architectural hypothesis did the paper formulate?



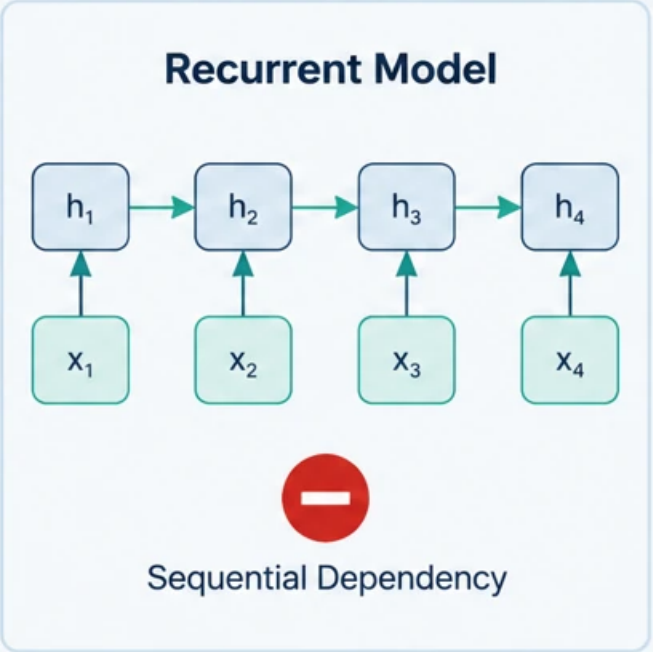
Answer 1: Research Problem & Objective

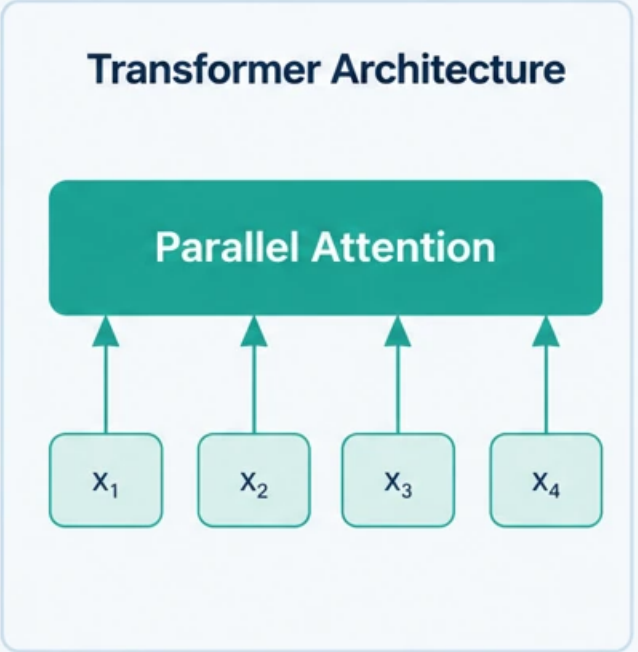
The authors sought to eliminate sequential computation bottlenecks that prevent parallelization during training, hypothesizing that a sequence transduction model could rely entirely on attention mechanisms without any recurrent or convolutional layers.

Section 1 & 2

Section Reference: Section 1 (Introduction) & Section 2 (Background).

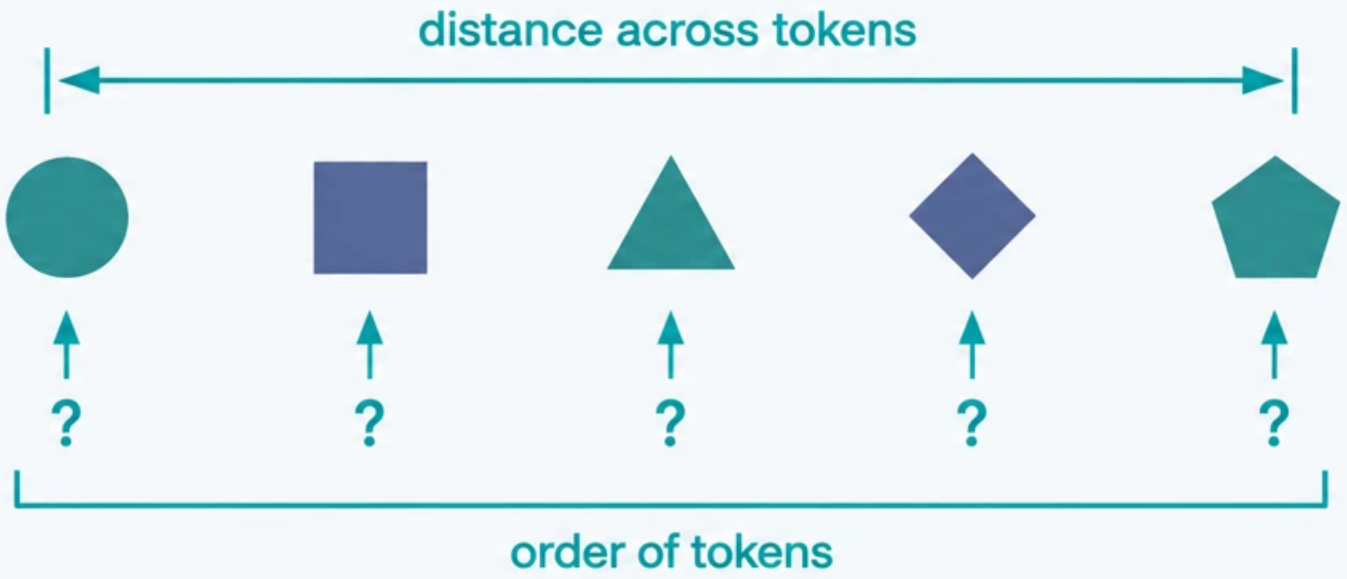
Key Context: Traditional recurrent models compute hidden states h_t sequentially as a function of the previous state h_{t-1} and input for position t. This sequential constraint inherently precludes parallelization within training examples. Vaswani et al. introduced the Transformer to achieve superior translation quality while enabling full sequence parallelization and drastically reducing training time.





Question 2: Method & Architectural Design

1. How does the Transformer handle two fundamental sequence-processing challenges without recurrence: (1) capturing relational dependencies across distant positions in parallel, and (2) tracking the relative or absolute order of tokens?



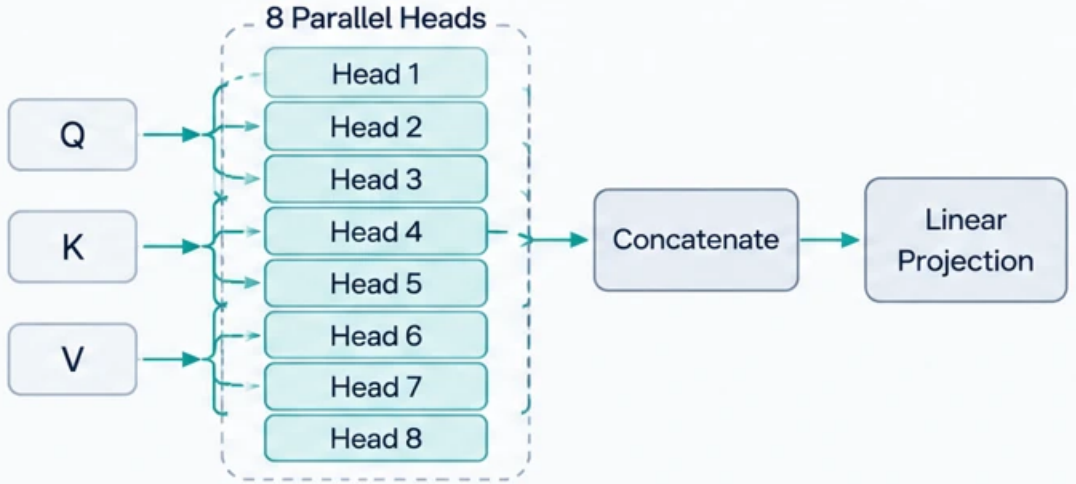
Answer 2: Method & Architectural Design

- (1) Multi-Head Scaled Dot-Product Attention captures relational dependencies across different representation subspaces simultaneously, while (2) Positional Encodings inject token order directly into input embeddings.

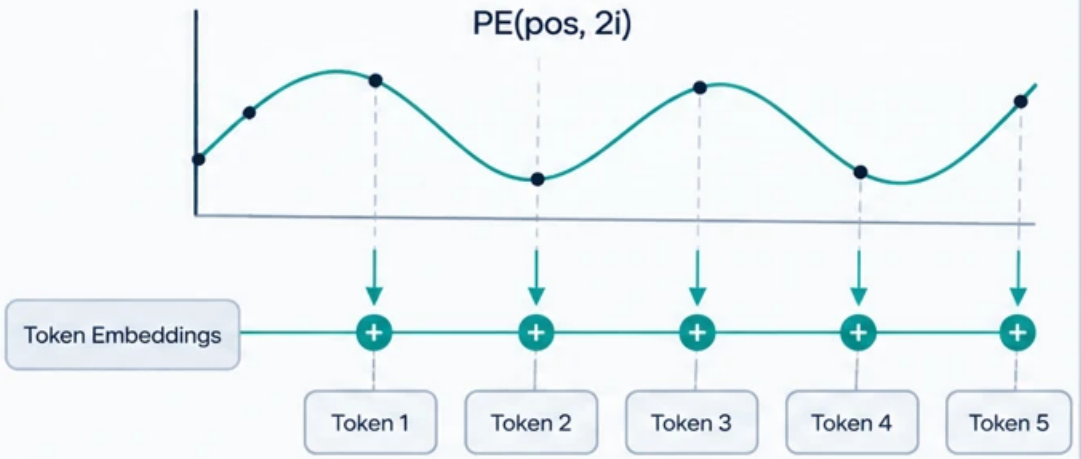
Section Reference: Section 3.2 (Attention) & Section 3.5 (Positional Encoding).

Key Context: Instead of performing a single attention function, Multi-Head Attention projects Queries (Q), Keys (K), and Values (V) h=8 times into lower-dimensional spaces ($d_k = d_v = 64$), computing attention in parallel across subspaces. Because the model contains no recurrence, fixed sinusoidal positional encodings using sine and cosine functions of varying frequencies are added directly to the input embeddings to provide positional awareness.

Multi-Head Attention



Sinusoidal Positional Encoding



Question 3: Principal Results

On the standard WMT 2014 English-to-German and English-to-French translation benchmarks, how did the Transformer (big) compare to previous state-of-the-art models in translation quality (BLEU score) and computational training cost?

Model	EN-DE BLEU	EN-FR BLEU	Training Cost (FLOPs)

Answer 3: Principal Results

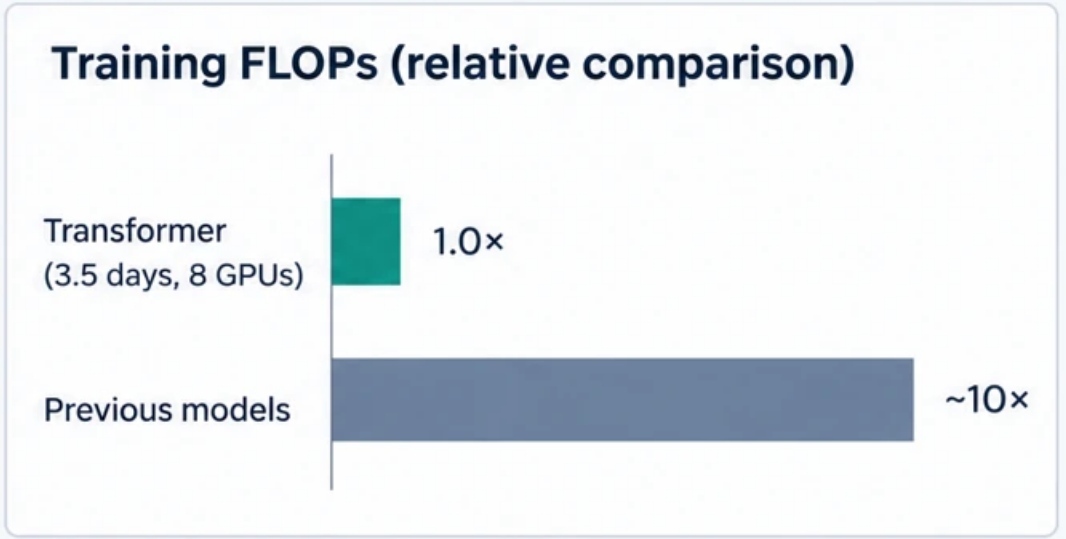
The Transformer (big) achieved a new state of the art on both benchmarks—28.4 BLEU on English-to-German (+2.0 BLEU over the previous best) and 41.8 BLEU on English-to-French—while requiring only 3.5 days of training on 8 P100 GPUs, a fraction of previous models' computational cost.

Section Reference: Section 5.1 (Machine Translation) & Table 2.

Key Context: Previous top models (including ByteNet, MoE, and GNMT+RL ensembles) required tens of thousands of GPU-days or large model ensembles. The Transformer established a new benchmark quality level while training at 2.3 x 10^19 FLOPs for English-to-German, outperforming models that used an order of magnitude more compute.

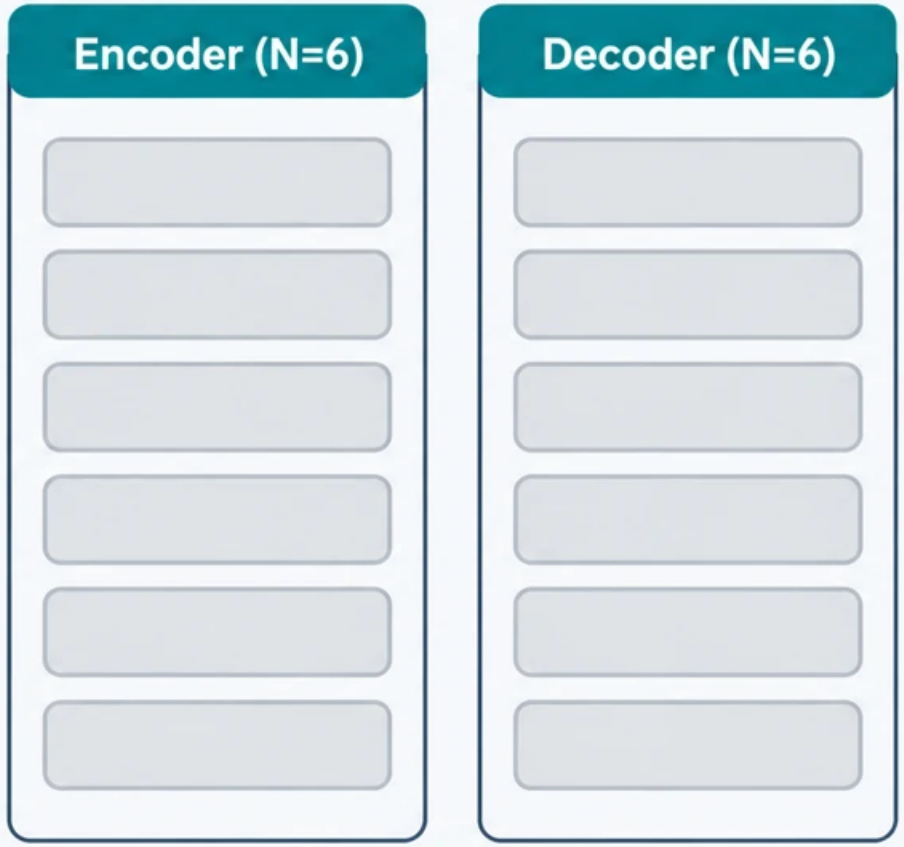
28.4
BLEU (EN-DE)

41.8
BLEU (EN-FR)



Question 4: Figure Interpretation

In Figure 1 of the paper (the Transformer model architecture), what critical modification is applied to the Multi-Head Attention layer in the Decoder stack that differentiates it from the Encoder stack, and what is its operational purpose?

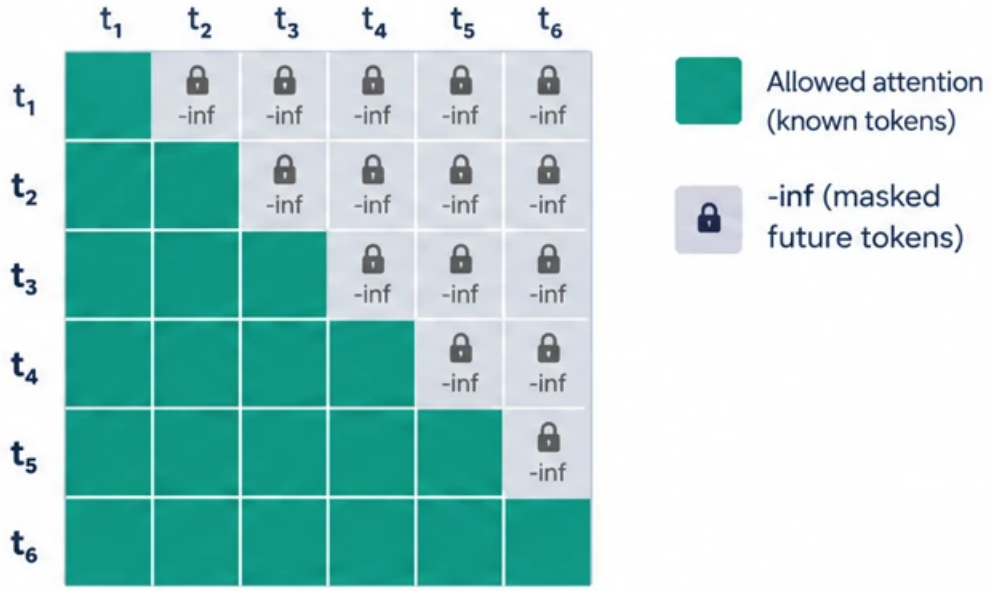


The Decoder uses Masked Multi-Head Attention to prevent positions from attending to subsequent future positions, preserving the auto-regressive property during generation.

Section Reference: Figure 1 (The Transformer - model architecture) & Section 3.1.

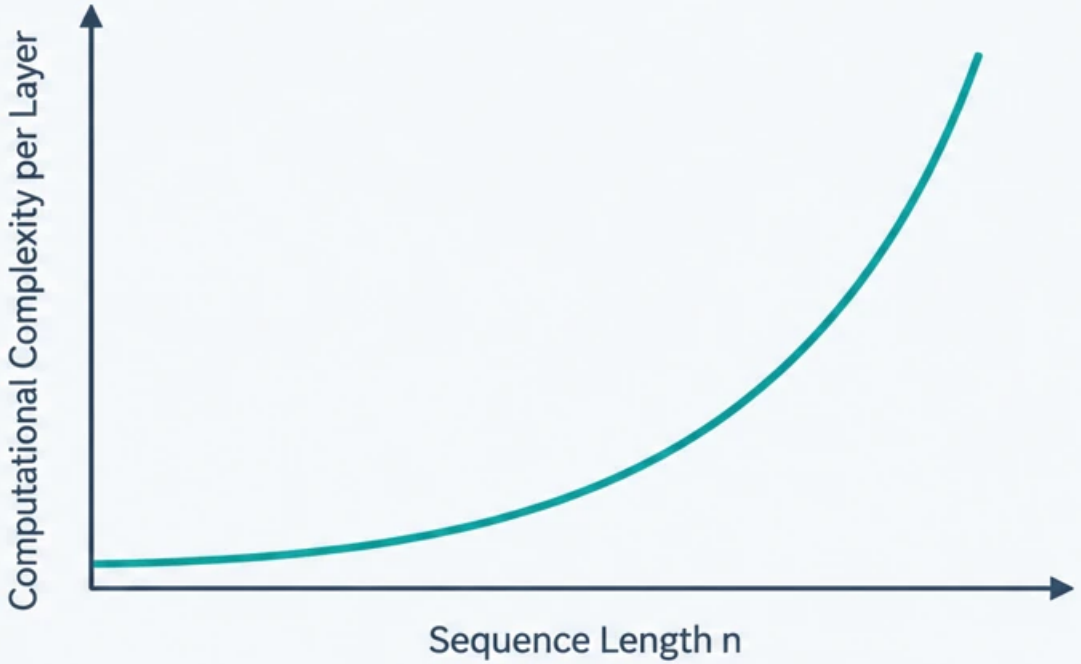
Key Context: Figure 1 illustrates an Encoder stack of N=6 identical layers and a Decoder stack of N=6 layers. While the encoder's self-attention allows all positions to attend to all other positions, the decoder's first sub-layer applies an attention mask by setting future dot products to negative infinity (-inf). This ensures predictions for position i depend strictly on known output tokens at positions less than i.

Figure 1 Detail: Masked Self-Attention



Question 5: Limitations & Future Directions

What primary computational scaling limitation of self-attention does the paper identify for very long sequences, and what future research directions do the authors explicitly propose in their conclusion?



Answer 5: Limitations & Future Directions

Self-attention exhibits quadratic computational complexity $O(n^2)$ with respect to sequence length n, making very long inputs costly; the authors proposed extending attention to modalities beyond text (images, audio, video) and investigating restricted or local attention spans for long sequences.

Section Reference:
Section 3.2.2, Table 1 (Maximum Path Length), & Section 6 (Conclusion).

Key Context:
Table 1 compares layer types: Self-Attention has complexity O(n^2 * d) per layer. When sequence length n exceeds the representation dimension d, self-attention becomes the computational bottleneck. In Section 6, the authors state plans to apply Transformers to problems with inputs and outputs in different modalities and explore restricted attention mechanisms connecting each output position to a neighborhood of size r.

Table 1: Complexity Comparison

Self-Attention	Restricted Self-Attention
$O(n^2 * d)$	$O(r * n * d)$

Future Multimodal Roadmap

