Architecture and Scale

The transformer architecture has revolutionized the way machines handle language by enabling the training of sequential data at scale. Thanks to this, today's AI models are massive, with some having billions of parameters (or more) allowing for incredible flexibility across many tasks. The technology is exciting and holds great promise for the future.

Technical Terms:

Sequential data: Information that is arranged in a specific order, such as words in a sentence or events in time.

Self-attention mechanism: The self-attention mechanism in a transformer is a process where each element in a sequence computes its representation by attending to and weighing the importance of all elements in the sequence, allowing the model to capture complex relationships and dependencies.