1. Why do we divide the attention score by √d in the scaled dot-product attention formula?

We divide by √d (where d is the dimension of the key) to prevent the dot-product scores from becoming excessively large, which can push the softmax function into regions with tiny gradients, causing training instability. Scaling by √d keeps gradients stable, improving training performance.

2. How does self-attention help the model understand relationships between words in a sentence?

Self-attention allows each word in a sentence to weigh every other word, capturing dependencies and contextual relationships directly. This helps the model learn meaningful context, identifying the importance of words relative to one another, thereby improving understanding of the sentence's overall meaning.