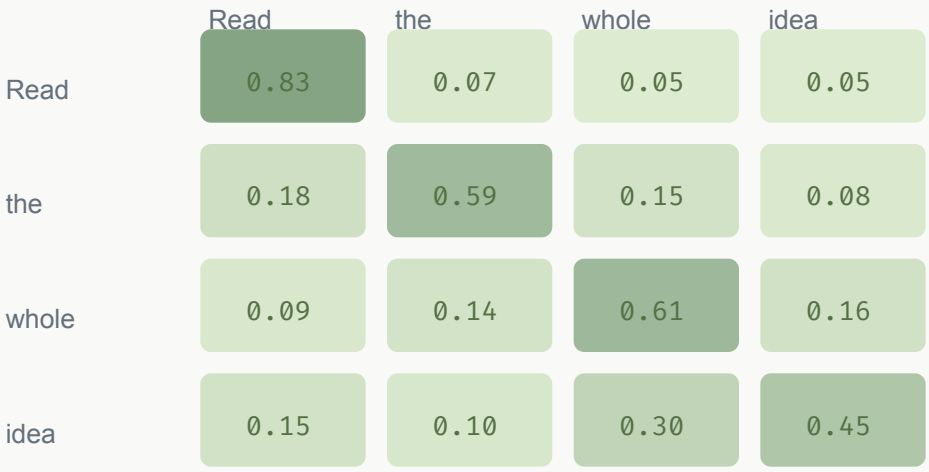
Attention, explained

How tokens gather useful context from other tokens.

01 ASK. MATCH. COMBINE.

Each token produces a query, a key, and a value. A query is compared with keys; the resulting scores become weights used to combine the values.



Illustrative weights; each row sums to 1.

02 CONTEXT CHANGES THE REPRESENTATION

The output for one token is a weighted sum of value vectors. Different heads can learn different relationships, giving the model several ways to connect information.

From scores to context

The scaled dot-product attention operation.

```
Attention(Q, K, V)
= softmax(QK^T / sqrt(d_k)) V
```

Compare

QK^T computes query-key dot products. Scaling by the square root of the key dimension keeps score magnitudes more manageable.

Normalize

Softmax is applied across keys for each query. The resulting nonnegative weights sum to one.

Combine

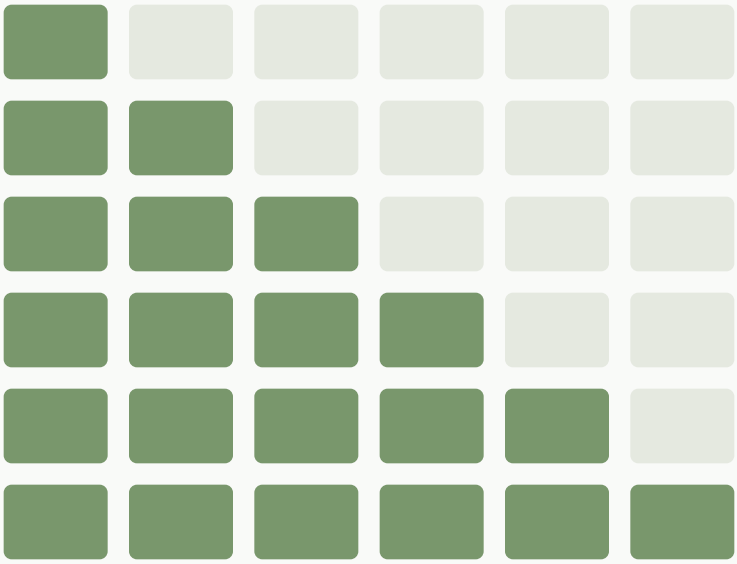
Multiply by V to form a weighted sum of values for every query.

What can a token see?

Masks define which positions are available as context.

CAUSAL ATTENTION

In a causal language model, a token can attend to its own position and earlier positions. Future positions are masked before softmax.



Visible context grows one position at a time.

Further reading: Vaswani et al. (2017), Attention Is All You Need.
arxiv.org/abs/1706.03762