Embeddings and scaled dot-product attention — review sheet

Use this after attempting the worked examples and practice. Explain each claim in words before calculating.

Key ideas

- Array axes distinguish positions from feature coordinates.
- Queries and keys determine scores; values supply coordinates to mix.
- Row-normalized attention weights sum to one over allowed keys.
- Apply a mask inside normalization and reject rows with no permitted keys.
- A fixed forward pass is distinct from parameter training, generation and scientific validation.

Common errors and repairs

Treating the output vector as token probabilities. Z contains mixed value coordinates; no vocabulary classifier is supplied.

Interpreting invented embeddings as learned meaning. These coordinates and projection coefficients were chosen by hand.

Erasing weights after normalization without rescaling. Normalize only across permitted keys.

Accepting a scientific claim because its explanation sounds plausible. Specify what the calculation establishes and obtain the appropriate independent evidence for claims outside it.

Retrieve without your notes

- What does softmax return for equal scores over two permitted keys?
- Why subtract a row maximum before exponentiation?
- What parameters are learned in this demonstration?

Readiness check

- Trace embedding, query, key, value, score and output array shapes.
- Compute one scaled dot-product attention row by hand.
- Explain stable row normalization and the effect of a causal mask.
- Distinguish a fixed-parameter calculation from training and from text generation.
- Test a numerical claim independently and document what an automated explanation has not established.