Quiz 4: Attention and Language Models — Answer Key

E190BO

---

1. An embedding maps each token to a vector of numbers that represents it. *(Accept any answer along the lines of "a numerical vector representation of a token/word.")*

2. The attention block lets the surrounding words (the context) update the embedding of "mole," so its vector moves toward the meaning that fits that sentence. *Accept any answer that says the embedding is adjusted using information from the other words.*

3. (a) value (b) query (c) key

4. **later** words cannot influence **earlier** words. *(Also accept "future"/"past" or "next"/"previous.")*

5. Each head can learn a different kind of relationship between words (e.g., adjectives modifying nouns, or which "Harry" is meant). Many heads let the model capture many kinds of context at once.

6. **one-hot encoded** inputs. *(This is why an embedding can be implemented as a simple table lookup.)*

7. **most frequent (most commonly co-occurring) pair** of adjacent tokens/vocabulary items.

8. output = **input** + **residual (a learned change)**. It makes very deep networks much easier to train: if the learned part starts near zero, the layer just passes its input through, and training only needs to learn small adjustments.

9. Any one of:

- It uses momentum (a running average of past gradients), so it doesn't zig-zag back and forth across a narrow "canyon" and makes steadier progress.
- It gives each parameter its own adaptive step size based on how large that parameter's gradients have been.