(N) A Token of Your Attention (1/2) [Solution]

N1. **(a)** .60 **(b)** .04 **(c)** .48 **(d)** .16

N2. **(e)** .40 **(f)** .20 **(g)** .30 **(h)** .60 **(i)** .20

N3. 1. C 2. D 3. E 4. A 5. B 6. C

N4. *The cat eats the rat on the mat.*

N5. **(j)** 𐑖 **(k)** 𐑚𐑖 **(l)** 𐑒𐑖𐑯 and 𐑒𐑚𐑡 **(m)** 𐑯𐑚𐑤 **(n)** 𐑖𐑓

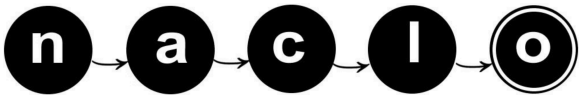
N6.

- Begin by distributing the attention of the verb across the sequence.
- A verb considers the following from most to least important:
  - Itself
  - Its inflection
  - Its subject
  - Its objects or obliques
  - Anything else (e.g., determiners, prepositions, adjectives)
- Only the categories present in the sequence are considered (e.g., a sentence without any objects or obliques has the hierarchy ELSE $<$ SUBJ $<$ INFL $<$ VERB), and each category is worth an incrementally larger value (e.g., INFL is worth three times as much as ELSE in the hierarchy exemplified above).
- Since the attention weights in a row must add up to 1, it's easiest to describe the calculations using algebra (i.e., the lowest category in the hierarchy has value x, the sum of the values in a row are set equal to 1, and x is calculated accordingly).
- Next, we calculate how much attention each token pays to the verb: each token pays twice the amount of attention to the verb as the verb pays to it (e.g., if the verb gives an attention weight of 0.25 to a token, then that token must give a weight of 0.50 to the verb). However, a token in the lowest category does not double this attention weight; rather, the attention between the two are equal.
- Next, we calculate the attention weight for the remaining non-diagonal cells. We have already calculated the attention weights for all edges involving the verb, so the remaining non-diagonal cells represent the edges between two tokens that are not the verb. Each of these cells is assigned the lowest attention weight previously used (i.e., x in our algebraic analysis).
- Finally, we complete the attention weight for the remaining diagonal cells. Since the attention weights in each row must add up to 1, we fill the remaining diagonal cells with the corresponding values.

Worked example: *The cat meowed.*

We start with the empty attention matrix:

	the	*cat*	*meow*	*ed*
the	?	?	?	?
cat	?	?	?	?
meow	?	?	?	?
ed	?	?	?	?



(N) A Token of Your Attention (2/2) [Solution]

Since there is no object or oblique here, the hierarchy is ELSE $<$ SUBJ $<$ INFL $<$ VERB. Let these have values x, $2x$, $3x$, and $4x$ respectively. Since the row sum must be 1, we get $x + 2x + 3x + 4x = 1$, so $10x = 1$ and therefore $x = .10$.

Doing this, we first fill in the verb row:

	the	*cat*	*meow*	*ed*
the	?	?	?	?
cat	?	?	?	?
meow	.10	.20	.40	.30
ed	?	?	?	?

Next, each token pays twice as much attention to the verb as the verb pays to it, except for the lowest category, where the attention is equal. Doing this, we get:

	the	*cat*	*meow*	*ed*
the	?	?	(.10)	?
cat	?	?	.40	?
meow	.10	.20	.40	.30
ed	?	?	.60	?

Next, we assign the lowest value, namely $x = .10$, to the remaining non-diagonal cells. Doing this, we get:

	the	*cat*	*meow*	*ed*
the	?	.10	.10	.10
cat	.10	?	.40	.10
meow	.10	.20	.40	.30
ed	.10	.10	.60	?

Finally, we fill the remaining diagonal cells so that each row sums to 1. Doing this, we get:

	the	*cat*	*meow*	*ed*
the	.70	.10	.10	.10
cat	.10	.40	.40	.10
meow	.10	.20	.40	.30
ed	.10	.10	.60	.20