Q1)

Embedding layer= Vocabulary size * Hidden size

= 40000* 768

= **30720000**

Attention parameters= 3*(Hidden size * Hidden size)

= 3 * (768 * 768)

= 1769472

Attention output parameters= Hidden size * Hidden size

= 768 * 768

= 589824

Total attention parameters per layer = 1769472 + 589824

= 2359296

Feed-forward parameters = (Hidden size*Feed-forward size) + (Feed-forward size * Hidden size)

= (768×3072) + (3072×768)

= 2359296+2359296

= 4718592

Total parameters per transformer layer=2359296 + 4718592

= 7077888

Total transformer layer parameters= 8 * 7077888

= **56623104**

Total number of parameters = 30720000 + 56623104 = **87343104**

The total number of parameters in the BERT model is approximately **87.3 million**

Q2)

Input Embeddings:

Word "flying" has embedding: [0,1,1,1,1,0]

Word "arrows" has embedding: [1,1,0,−1,−1,1]

Query, Key, and Value Vectors:

Query for "flying" (first two dimensions of the "flying" embedding):

Query = [0,1]

Key for "flying" (first two dimensions of the "flying" embedding):

Key (flying) = [0,1]

Key for "arrows" (first two dimensions of the "arrows" embedding):

Key (arrows) = [1,1]

Value for "flying" (first two dimensions of the "flying" embedding):

Value (flying) = [0,1]

Value for "arrows" (first two dimensions of the "arrows" embedding):

Value (arrows) = [1,1]

Scaled Dot-Product Attention

Dot product of query with key (flying):

Query·Key (flying) = (0 * 0) + (1 * 1) = 1

Dot product of query with key (arrows):

Query·Key (arrows) = (0 * 1) + (1 * 1) = 1

Scaling by sqrt 2:

Scaled scores=[1/ sqrt 2 , 1 / sqrt 2]

=[1 / 1.414 , 1 / 1.414]

≈**[0.707,0.707]**

Softmax of Scaled Scores:

Softmax([0.707,0.707]) = [e^0.707 / (e^0.707 + e^0.707) , e^0.707 / (e^0.707 + e^0.707)]

= [0.5,0.5]

Self-attention output: Output=0.5 * [0,1] + 0.5 * [1,1]

Output = [0.5 * 0 + 0.5 * 1, 0.5 * 1 + 0.5 * 1]

=**[0.5,1]**

Q3)

For topic classification with 5 classes, the task-specific linear layer will have:

- **Input size** = 768 (BERT-base hidden state size)
- **Output size** = 5 (number of classes)

The number of task-specific parameters in this linear layer can be calculated as:

Number of parameters=(Input size * Output size) + Output size

=(768 * 5) + 5

=3840+5

=3845

Thus, the **task-specific parameters for topic classification with 5 classes** = **3845 parameters**.

Task-Specific Parameters for Language Identification in a Code-Switched Dataset

In this case, the classification task involves only **2 classes**: one for English and one for Hindi.

The task-specific linear layer will have:

- **Input size** = 768 (BERT-base hidden state size)
- **Output size** = 2 (number of languages)

The number of task-specific parameters in this case will be:

Number of parameters= (Input size * Output size) + Output size

= (768 * 2) + 2

= 1536+2

= 1538

Thus, the **task-specific parameters for language identification** in this two-language code-switched dataset = **1538 parameters**.