1.a)i)

α is kind of a categorical probability distribution because:

n number of categories, α1, . . . , αn event probabilities

and αi > 0, $\sum \alpha\mathrm{i} = 1$ which are the prerequisites for a categorical probability distribution.

ii)

copying in attention $k_j^T q \gg k_i^T q, i \neq j$

iii)

we know that αi > 0, $\sum \alpha\mathrm{i} = 1$ ans also from αj = $\sum_{i!=j}$ αi we understand that

αj = 1 so $\sum_{i=1}^{n} vi\ ai$ = vj aj = vj .

iv)

When one of the key vectors is almost identical to the given query, the attention weight assigned to it will be significantly higher. As a result, the output will be strongly influenced by that specific key's value. In this way, we can say that the model has copied the value.

b)i)

$M(V_a + V_b) = V_a$

$M V_a = M(c_1 a_1 + c_2 a_2 + \dots + c_m a_m) = MA_c = V_a$ and $a_j^T a_i = 0$ if i==j $a_j^T a_i = 1$

- ⇨ $M = A^T$
- ⇨ $A^T A c = c_1 a_1^T a_1 + \cdots + c_m a_m^T a_m = c$

$M V_b = M(d_1 b_1 + d_2 b_2 + \dots + d_p b_p) = MB_d = V_b$ and $a_j^T b_k = 0$

- ⇨ $M = A^T$
- ⇨ $A^T B_d = d_1 a_1^T b_1 + \cdots + d_p a_p^T b_m = 0$

$M_{va} + M_{vb} = va$ => $A^T A_c + A^T B_d = c + 0 = c$

So $M = A^T$ and vector of weights is c.

ii)

c = (1/2)($V_a + V_b$) => $x_a = x_b = 0.5$ and part a

⇨ $q^T k_a = q^T k_b \gg q^T k_i \quad \forall i$!= a, b

⇨ $q^T k_a = q^T k_b = \beta$ => $\frac{\exp(\beta)}{\sum_{j=1}^{n} \exp(\beta)} = \frac{\exp(\beta)}{n-2+2\exp(\beta)}$ for $\beta \gg 0$

⇨ $\exp(\beta) = \infty$ => $\frac{\exp(\beta)}{2\exp(\beta)} = 0.5$

⇨ $q = \beta(k_a + k_b) \quad \beta \gg 0$

c)i)

Considering that α tends to approach zero, the covariance matrices' diagonal elements also become extremely small. so, when sampling k_i with mean μ_i and covariance Σi, the sampled value k_i will have a value close to μ_i.

Furthermore, since the means μi are all perpendicular it leads to the same expression for q, which is $\beta(\mu_a + \mu_b)$.

ii)

we know μ_i T μ_i = 1 and α is small and $\sum a =$ αI + (1/2) (μ_aTμ_a)

⇨ ka ∈ [$0.5\mu_a$, $1.5\mu_a$] i != a
⇨ ka = Xμa and X = N (1, 0.5)
⇨ ki = μi ∀I != a
⇨ $k_a{}^T q = X \mu_a{}^T \beta(\mu_a + \mu_b) = X\beta \quad \beta \gg 0$
⇨ $K_b{}^T q \approx \mu_b{}^T \beta(\mu_a + \mu_b) = \beta \quad \beta \gg 0$
⇨ $K_i{}^T q \approx \mu_i{}^T \beta(\mu_a + \mu_b) = \beta(\mu_i{}^T \mu_a + \mu_i{}^T \mu_b) = \beta(0 + 0) = 0 \quad \beta \gg 0$
⇨ $\frac{\exp(\text{Ka T q})}{\sum_{j=1}^{n} \exp(\text{Ki T q})} = \frac{\exp(\text{X q})}{\exp(\text{X q}) + \exp(\beta)} = \frac{1}{1 + \exp(\beta(1-\text{X}))}$

⇨ $\frac{\exp(\mathrm{Kb\ T\ q})}{\sum_{j=1}^{n}\exp(\mathrm{Ki\ T\ q})} = \frac{\exp(\beta)}{\exp(\mathrm{X}\,\beta) + \exp(\beta)} = \frac{1}{1+\exp(\beta(1-\mathrm{X}))}$

⇨ X is a shifted softmax to right by 1. So for the start and end of X period which is [0.5 , 1.5]

⇨ For X = 0.5 and $\beta \gg 0$ $\frac{1}{1+\exp(\beta(1-0.5)} = \frac{1}{1+\infty}$ = 0 and $\frac{1}{1+\exp(\beta(0.5-1)} = \frac{1}{1+0}$ = 1

⇨ $\frac{1}{1+\exp(\beta(1-1.5)} = \frac{1}{1+0}$ = 1 and $\frac{1}{1+\exp(\beta(1.5-1)} = \frac{1}{1+\infty}$ = 0

So c = v_a if X is 1.5 and c = v_b if X is 0.5

It differs from part i because in part i c is balanced combination of both v_a and v_b but here c swings between these.

d)i)

c1 = 1/2 ($v_a + v_b$) and $q_1 = \beta(\mu_a + \mu_b)$

c2 = 1/2 ($v_a + v_b$) and $q_2 = \beta(\mu_a + \mu_b)$

c = 1/2 (c1 + c2) = 1/2 (1/2 ($v_a + v_b$) + 1/2 ($v_a + v_b$)) = 1/4 ($v_a + v_b$) + 1/4 ($v_a + v_b$) = 1/2 ($v_a + v_b$)

ii)

c is like part (c)ii). If we add more attention the swings of c between va and vb will be less. If X is 1 then we have :

c = $\frac{1}{1+\exp(\beta(1-1)}v_a + \frac{1}{1+\exp(\beta(1-1)}v_b = \frac{1}{2}(v_a + v_b)$

2.

d) accuracy on the dev set: 10.0 out of 500.0: 2.0%

accuracy “London” : 25.0 out of 500.0: 5.0%

f) accuracy on the dev set: 72.0 out of 500.0: 14.399999999999999%

g)i) accuracy : 43.0 out of 500.0: 8.6%

ii) the complexity of attention operation is reduced to $O(d \times m)$.

complexity of self attentions in the latent transformer blocks will reduce to $O(m2)$.

multi-headed attention has a time complexity of $O(\ell 2d + \ell d2)$

the time complexity of the perceiver model is $O(dm + Lm2)$,

3.

a) it had basic knowledge. Because it was trained on a large dataset.

b) the information made, will be wrong and this can lead to insecure events.

c) same as part b