1 Attention Exploration

1.1

(a)

Let us note that due to the formula that defines each α_i the sum of all α_i is 1. Hence, by looking at each value as category we easily get that this is a categorical distribution.

(b)

This would happen if

1. there exists some k_j where $k_j^T q$ is large and
2. for all other k_i where $i \neq j$ $k_i^T q$ is small.

Let us remember that $a^T b = |a||b|\cos(\theta)$ where θ is the angle between a and b. Hence for the conditions above to hold we need that k_j is parallel to q and all other k_i are in the opposite direction to q.
Plus, we need that either q is very large (i.e large norm) or that k_j is very large or that the other k_i vectors norm is very large.

(c)

Assuming that the conditions above are correct, we get that $c \approx v_j$

(d)

This means that the output c will depend almost only on a single v_j that corresponds to the k_j that is parallel to q. This is, obviously, not desirable since we want the output to be contextualized, i.e depend on the full sentence. Without this property the model will not be expressive enough.

1.2

(a)

Let us define the following matrix M:

$$M = c_1 \frac{a_1 a_1^T}{a_1^T a_1} + c_2 \frac{a_2 a_2^T}{a_2^T a_2} + \cdots + c_m \frac{a_m a_m^T}{a_m^T a_m}$$

If we multiply M by a_i we get that

$$Ma_i = c_i a_i$$

i.e, a_i is an eigenvector of M with eigenvalue c_i. That means that

$$Mv_a = M(a_1 + \cdots + a_m) = Ma_1 + \cdots + Ma_m = c_1a_1 + \cdots + c_ma_m = v_a$$

Plus, when multiplying M by v_b we get that

$$Mv_b = c_1\frac{a_1a_1^Tv_b}{a_1^Ta_1} + c_2\frac{a_2a_2^Tv_b}{a_2^Ta_2} + \cdots + c_m\frac{a_ma_m^Tv_b}{a_m^Ta_m}$$

we know that $a_i^Tv_b = 0$ for all i since

$$a_i^Tv_b = a_i^Tb_1 + ... + a_i^Tb_m = 0$$

which is true because the two basis are supposed to be orthogonal. Hence

$$Mv_b = 0$$

Hence we finally get that

$$Ms = Mv_a + Mv_b = v_a$$

(b)

we define q as follows:

$$q = (k_a + k_b) * H$$

where H is some Huge number. It holds that

$$k_aq = k_bq = H$$

Plus, it holds that

$$k_iq = 0$$

if $i \neq a, b$. We get that $\alpha_a = \alpha_b = \frac{e^H}{n-2+2e^H}$ and $\alpha_i = \frac{1}{n-2+2e^H}$ for $i \neq a, b$.
For large enough H we get that $\alpha_a \approx \alpha_b \approx \frac{1}{2}$ and $\alpha_i \approx 0$ for $i \neq a, b$. Hence $c \approx \frac{1}{2}(v_a + v_b)$

1.3

(a)

Let us note that for $k_i \sim N(\mu_i, \alpha_iI)$ we can rewrite k_i as $k_i = \mu_i + \sigma_i$ where $\sigma_i \sim N(0, \alpha_iI)$. Similarly to the clause before we define

$$q = (\mu_a + \mu_b)H$$

we get that for $k_a = \mu_a + \sigma_a$

$$(\mu_a + sigma_a)(\mu_a + \mu_b)H = H(1 + \sigma_a(\mu_a + \mu_b)) \approx H$$

and similarly for k_b. For k_i where $i \neq a, b$ we get that

$$(\mu_i + \sigma_i)(\mu_a + \mu_b)H = \sigma_i(\mu_a + \mu_b)H \approx 0$$

Hence we get similarly to the previous clause that $c \approx \frac{1}{2}(v_a + v_b)$

(b)

We get that

$$k_a q = (\mu_a + \sigma_a)(\mu_a + \mu_b)H \approx H(1 \pm ||\sigma_a||)$$

Let us note that $||\sigma_a||$ flactuates from 0 to $\frac{1}{2}||\mu_a|| = \frac{1}{2}$.
Hence, for different samples, we can see that the value of $k_a q$ flactuates from $1.5H$ to $\frac{H}{2}$. Hence we get that α_a flactuates from being equal to $\alpha_b * e^{0.5H}$ to being equal to $\alpha_b * e^{-0.5H}$.
Let us remind our selves that $\alpha_a, \alpha_b < 1$ and that H is a huge number. We that if $\alpha_a = \alpha_b * e^{0.5H}$ then $\alpha_a \approx 1$ and $\alpha_b \approx 0$. Similarly, if $\alpha_a = \alpha_b * e^{-0.5H}$ then $\alpha_a \approx 0$ and $\alpha_b \approx 1$.
Hence, the expectation is $E[c] \approx \frac{1}{2}(v_a + v_b)$ but c flactuates from $c \approx v_a$ to $c \approx v_b$.

1.4

(a)

We define $q_1 = q_2 = (\mu_a + \mu_b)H$. Due to the analysis in previous clauses we know that both c_1, c_2 are equal to $\frac{1}{2}(v_a + v_b)$ under this definition of q_1, q_2.
Hence, their average is also equal to $\frac{1}{2}(v_a + v_b)$.

(b)

Let us denote the value c that we've gotten in clause 1.3b by c^*. We've seen that $E[c^*] = \frac{1}{2}(v_a + v_b)$ with some variance we denote by σ^*. Let us note that c_1, c_2 are equal to c^* (under the same sampling, of course, as they are random variables). Hence, we get that $E[c_1] = E[c_2] = \frac{1}{2}(v_a + v_b)$ and that $Var[c_1] = Var[c_2] = \sigma^*$.
Our c is equal to $\frac{1}{2}(c_1 + c_2)$. Hence, we get that $E[c] = \frac{1}{2}(E[c_1] + E[c_2]) = \frac{1}{2}(v_a + v_b)$ and that $Var[c] = \frac{1}{4}(Var[c_1] + Var[c_2]) = \frac{1}{2}\sigma^*$.
We've been able to keep the same expectation and reduce the variance by a factor of 2, using multi-head attention.

2 Pretrained Transformers

2.d

The result of london baselinr is -
Correct: 25.0 out of 500.0: 5.0%
The result of the finetuned model is -
Correct: 10.0 out of 500.0: 2.0%

2.f

Correct: 115.0 out of 500.0: 23.0%

2.g

Let us note that the pretraining data contains the birthplace information about the individuals which are in the birth_dev dataset. It seems that the model is able to "remember" the birthplace knowledge that it acquires during pretraining, and it is able to infer that knowledge during evaluation. Let us also note that the finetuning process does not make the model forget the knowledge it acquired during pretraining.

3 In-Context Learning

3.1.b

Results for beam search were better for all prompt sizes. This is probably due to the trade-off between precision and creativity when considering "beam search" or "sampling decoding".

3.2.b

Validation question:

What are Jimmy Reed, Skip James, Elmore James and Big Bill Broonzy best known as?

Prompt:

Answer: Eleonore Roosevelt Question: On which Caribbean island did Princess Diana spend he first Christmas after her divorce was announced? Answer: Barbuda Dependency, Antigua and Barbuda Question: Who wrote the song Momma Told Me Not To Come? Answer: Randy Newman Question: What is the distance between bases on a little league baseball field? Answer: sixty distance Question: Which NASA space probe was launched to Venus in 1989? Answer: List of things named for Ferdinand Magellan Question: Which member of the Monkees came from Washington DC? Answer: Peter Tork Question: Who was the woman sentenced to six years in jail after the murder of Stompei Seipi? Answer: Winnifred Mandela Question: Which Italian fashion designer was murdered on the orders of his ex-wife? Answer: House of Gucci

Validation question:

For which movie did Katharine Hepburn win her second Oscar?

Prompt:

Question: Which African country is sandwiched between Ghana and Benin? Answer: Religion in Togo Question: Which African country is bordered by Benin, Ghana, Ivory Coast, Niger, and Mali? Answer: Burkina Faso Question: Which country mainly makes up the Horn of Africa? Answer: Somolia Question: In The Banana Splits what sort of animal was Snorky? Answer: Elephantineness Question: What are the two main arms of the River Nile called? Answer: Blue Nile and White Nile Question: Name the East African country which lies on the equator. Answer: Prehistory of Kenya Question: What are the international registration letters of a vehicle from Algeria? Answer: DZ (disambiguation) Question: Banting and Best pioneered the use of what? Answer: Insullin

By selecting the closest vectors to the test question, we are effectively providing the model with contextual

information that closely matches the specific question being asked. This approach helps the model focus on relevant training examples that are more likely to contain useful information for generating accurate answers.

3.3

In zero-shot learning, the model has not been explicitly trained on the specific task of question answering. Instead, it relies on its general language understanding and knowledge acquired during pre-training. Without task-specific training, the model may lack the fine-tuned knowledge and expertise necessary to generate accurate and relevant answers.

Some generation examples:

I'm not sure, but I'm pretty sure it's not me. I don't 'Run!‖ †Run for your life! ‡‡Run Gummo: "Chico" and "Harpo" are my last names. It depends on the type of plant. For example, if the plant is a shrub

References