

Q1:

1. BERT use fine-tuning to train a model to fit different task. This means that in the process of fine-tune, we use language data with answer and question to train the gradient descend. We expect the model to output two integers for answer. One means the start of the answer in the context. The other means the end. We need to random initialize two vector and use them to do inner product with output of BERT pretrained model. After softmax the position with highest inner product become answer.
2. GPT uses the provided prompt as a guide to generate a relevant response. The prompt provides direction for the model to follow in order to generate an answer that is consistent with the question. Based on the prompt and the context, GPT generates a candidate answer by selecting the most relevant passage or sentence(s) from the context that contains the information needed to answer the question. At last, GPT then uses additional language processing techniques to refine the candidate answer, such as identifying any synonyms or related terms that may be relevant and ensuring that the answer is grammatically correct and makes sense in the context of the passage. Due to the way of generation, GPT's answer may not be the same sentence in the provided context, but the combination of different part of context

第一組使用「請從最後一篇的文章中找出最後一個問題的答案」(左)

「從最後一篇文章擷取問題的正確答案」(右)

0, 2007年1月16日, 2007年1月16日	0, 2007年1月16日, 2007年1月16日
1, 200公里, 200公里。	1, 200公里, 200公里
2, 廣州, 廣州	2, 廣州, 廣州
3, 2015, 1990年	3, 2015, 1990年
4, 文化大革命, 武昌起義	4, 文化大革命, 文化大革命
5, 廣州, 廣州	5, 廣州, 廣州
6, 廣州, 廣州	6, 廣州, 廣州
7, 武昌起義, 1911年10月10日	7, 武昌起義, 廣州文章: 清朝末期, 廣州爆發了幾次武裝起義
8, 香港, 廣州	8, 香港, 廣州
9, 1921, 廣州城址在廣州縣的西南部, 廣州縣的西北部	9, 1921, 公元前214年, 秦始皇出兵徵兵南, 南海縣任
10, 7月, 2010	10, 7月, 2010年8月1日
11, 2008年, 從化市	11, 2008年, 廣州获得這個獎項是從1990年开始的。
12, 從化市, 廣州的人口密度是100倍分之一以下。	12, 從化市, 廣州的人口密度為100倍分之一以下。
13, 80, 90%	13, 80, 80%
14, 非洲黑人, 非洲黑人	14, 非洲黑人, 非洲黑人
15, 普通話, 廣州國民政府成立后, 廣州的國民政府和國民政府的國	15, 普通話, 廣州的社會運動和政治活動主要以廣東語為主。在廣州, 社會

可以見得右邊的正确率較高, 有些錯誤的答案也離正確解答更近, 我認為是因為「擷取」比「找出」更精確一些, 畢竟模型是從文本中截出一段是答案的文字

第二組使用「請從最後一篇的文章中找出最後一個問題的答案」(左)

「在 P 中, Q 的答案是 A」(右)

0, 2007年1月16日, 2007年1月16日	0, 2007年1月16日, 從化市
1, 200公里, 200公里。	1, 200公里, 約200公里
2, 廣州, 廣州	2, 廣州, 廣州
3, 2015, 1990年	3, 2015, 1990
4, 文化大革命, 武昌起義	4, 文化大革命, 因為政局不穩定, 政府大力支持以日用品為主的輕工業, 廣州
5, 廣州, 廣州	5, 廣州, 廣州
6, 廣州, 廣州	6, 廣州, 廣州
7, 武昌起義, 1911年10月10日	7, 武昌起義, 廣州文章: 清朝末期, 廣州爆發了幾次武裝起義
8, 香港, 廣州	8, 香港, 香港
9, 1921, 廣州城址在廣州縣的西南部, 廣州縣的西北部	9, 1921, 2012年6月
10, 7月, 2010	10, 7月, 2010年7月
11, 2008年, 從化市	11, 2008年, 廣州获得了「創建國家健康城市」的獎項。文章: 廣州是
12, 從化市, 廣州的人口密度是100倍分之一以下。	12, 從化市, 越秀區
13, 80, 90%	13, 80, 廣州人口
14, 非洲黑人, 非洲黑人	14, 非洲黑人, 非洲黑人
15, 普通話, 廣州國民政府成立后, 廣州的國民政府和國民政府的國	15, 普通話, 廣州的語言交換主要講普通話, 但當地人仍然使用廣東話

在有些答案中右邊是比左邊更加接近正解的, 但同時也產出一些匪夷所思的錯誤。錯誤可能來自於 prompt 的表達比較不直觀, 也不完全是中文, 導致模型預測錯誤。不過這樣的隨機性也導致有些結果是更接近正解的。詳細的比例可能需要更多測資才能顯現。

第三組使用「請從最後一篇的文章中找出最後一個問題的答案」(左)

「According to the passage, print the answer of the question」(右)

0, 2007年1月16日, 2007年1月16日	0, 2007年1月16日, 2007年1月16日
1, 200公里, 200公里。	1, 200公里, 平均每小時可以走約200公里。
2, 廣州, 廣州	2, 廣州, 廣州
3, 2015, 1990年	3, 2015, 1990年
4, 文化大革命, 武昌起義	4, 文化大革命, 武昌起義
5, 廣州, 廣州	5, 廣州, 廣州
6, 廣州, 廣州	6, 廣州, 廣州皇城
7, 武昌起義, 1911年10月10日	7, 武昌起義, 清朝末期
8, 香港, 廣州	8, 香港, 從化市
9, 1921, 廣州城址在廣州縣的西南部, 廣州縣的西北部	9, 1921, 1911年10月10日
10, 7月, 2010	10, 7月, 2010年7月
11, 2008年, 從化市	11, 2008年, 廣州在1990年以前, 廣州市政府已經获得了「創建國家健康城
12, 從化市, 廣州的人口密度是100倍分之一以下。	12, 從化市, 越秀區
13, 80, 90%	13, 80, 廣州
14, 非洲黑人, 非洲黑人	14, 非洲黑人, 非洲黑人
15, 普通話, 廣州國民政府成立后, 廣州的國民政府和國民政府的國	15, 普通話, 廣州地區的社會生活與日常生活, 包括飲食, 娛樂, 交通, 教育,

可以看到右邊的表现不太好，可能因為 prompt 為英文，讓模型無法確切接收任務的資訊