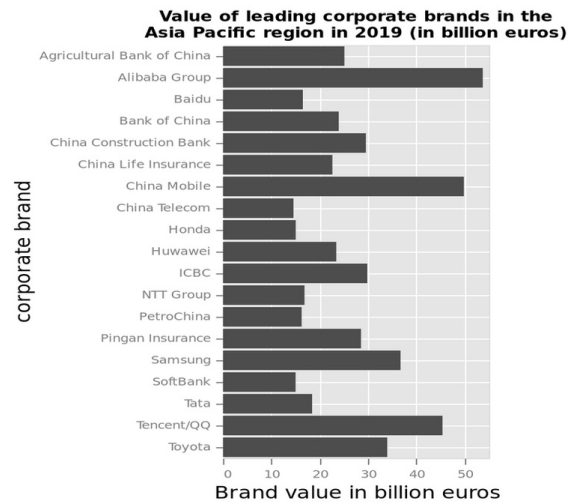


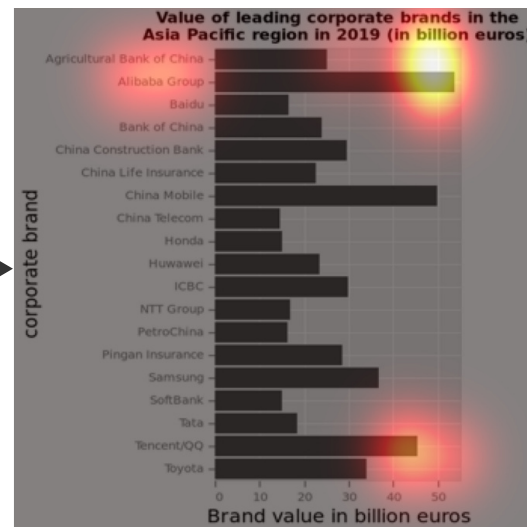
## Input Plot



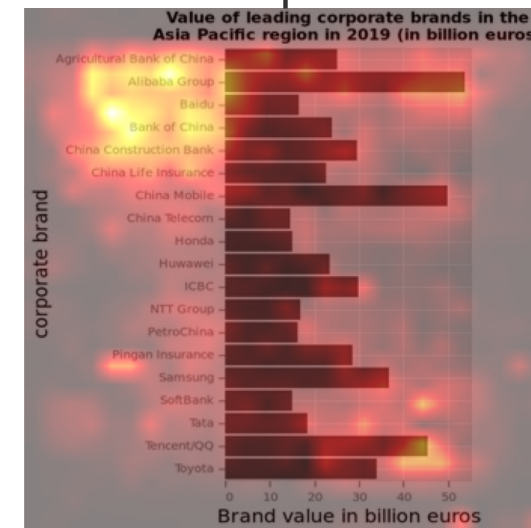
## Human Annotation and Recording Eye-gaze Data



## Human Eye-gaze Attention Matrix (G)



$$\mathcal{L}_{W-MSE} = \frac{1}{N} \sum_{i=1}^N w_i \cdot (G_i - A_i)^2$$



$$\mathcal{L}_{total} = \mathcal{L}_{LM} + \mathcal{L}_{W-MSE}$$

## Question and Answers Generated Based on the Chart Summaries From VisText

Was Alibaba Group the most valuable brand at over €50 billion?

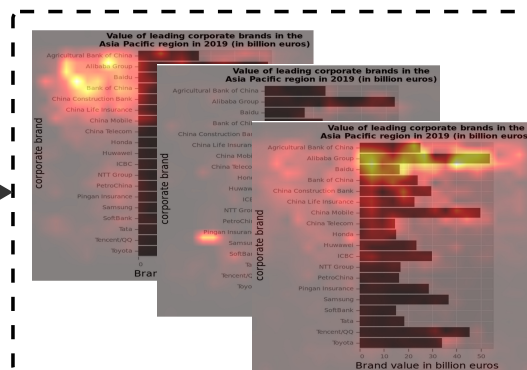
Answer: **True**



## LARGE VISION LANGUAGE MODEL



## Attention Between Text and Vision Tokens Averaged over All Heads and Text Tokens



**Model Output Tokens:**  
True </s>

**Average Over the First M Layers**

$$\mathcal{L}_{LM} = - \sum_{t=1}^T \log P(y_t | x_{\leq t})$$