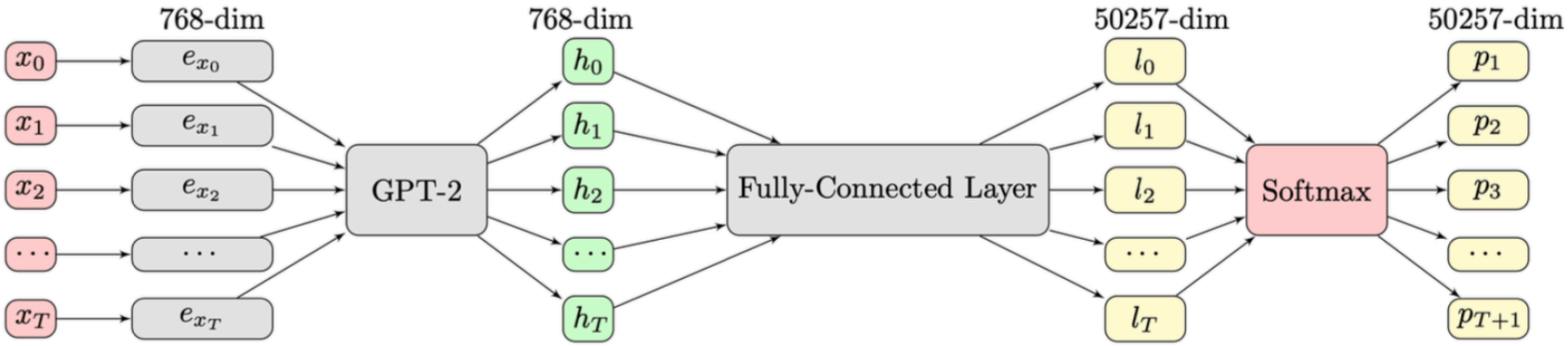

Figure 1: The architecture of our model. T is the sequence length of a given input. x_i is the index token. e_{x_i} is the trainable embedding of token x_i. h_i is the output of GPT-2. l_i is the logit and p_i is the probability. Nodes in gray contain trainable parameters.