A Supplemental Materials

A.1 Complete Training Procedure

The complete training procedure is shown in Algorithm 2 in the next page.

A.2 Random Policy and Oracle Policy Performances

The random policy randomly selects K comments at each step. The oracle policy knows the true karma score of each comment in the datasets and it always chooses the top-K comments with highest true karma scores at each step. Table 7 shows the performance of the random policy and the oracle policy on different datasets when $N = 10$ and $K = 3$. Table 8 shows the performance of the random policy and the oracle policy across various action sizes with $K = 2, 3, 4, 5$ and fix $N = 10$ on the askscience dataset.

Subreddit	Random Policy	Oracle Policy
askscience	392.0±10.1	1695.4±48.4
askmen	188.0±4.4	524.5±26.3
todayilearned	528.3±29.0	1994.5±65.2
worldnews	351.4±11.5	1328.0±40.1
nfl	328.3±14.6	1032.5±5.7

Table 7: Performance of the random policy and the oracle policy on different datasets

K	Random Policy	Oracle Policy
2	254.1±20.9	1571.2±27.8
3	392.0±10.1	1695.4±48.4
4	589.3±22.0	1697.8±35.3
5	745.9±26.4	1808.5±47.2

Table 8: Performance of the random policy and the oracle policy with different action size on askscience dataset

Algorithm 2 Q-learning

Initialize the experience memory D
Initialize θ randomly
Set $\theta^{-} = \theta$
for $episode = 1 \rightarrow H$ **do**
  Randomly pick a discussion tree
  Read the initial state s_1, and a set of possible sub-actions, $c_1 = \{c_{1,1}, \ldots, c_{1,K}\}$
  for $t = 1 \rightarrow +\infty$ **do**
    if $rand() < \epsilon$ **then**
      Select action $a_t = \text{Greedy}(s_t, c_t, Q(\cdot, \cdot; \theta), K)$
    else
      Select action a_t uniformly at random
    Observe reward r_{t+1}
    Read the next state s_{t+1}, and next set of possible sub-actions, $c_{t+1} = \{c_{0,1}, \ldots, c_{0,K}\}$
    Store a transition tuple, $(s_t, a_t, r_{t+1}, s_{t+1}, c_{t+1})$ in D
    Sample random mini batch of transition tuples $(s_j, a_j, r_{j+1}, s_{j+1}, c_{j+1})$ from D
    Set

$$y_j = \begin{cases} r_{j+1} & \text{if } s_{j+1} \text{ is terminal} \\ r_{j+1} + \gamma \max_{a'} Q(s_{j+1}, a'; \theta^{-}) & \text{otherwise} \end{cases}$$

    Perform a step of gradient descent on the loss $L(\theta) = (y_j - Q(s_j, a_j; \theta))$ with respect to θ
    Set $\theta^{-} = \theta$ for every F steps
    if c_{t+1} is empty **then** break