Supplementary Material: Checklist Hyperparameter Selection ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing

Ryan Liu and Nihar Shah
{ryanliu, nihars}@andrew.cmu.edu
Carnegie Mellon University

1 Choosing GPT-4 hyperparameters: Evaluating NeurIPS checklist

We describe our pilot study to determine the best hyperparameters for GPT-4 to answer peer review checklist questions. To ground this, we use one representative question from each of the five categories within the NeurIPS 2022 checklist. For hyperparameter values, we consider `temperature` values $\{0, 0.1, 0.2, \ldots, 2.0\}$ and `top_p` values $\{0, 0.1, 0.2, \ldots, 1.0\}$. Following OpenAI's recommendation[1], we fix one of the hyperparameter values at 1 at all times, testing the other.

For each set of hyperparameters, we choose the first checklist item in each category, and query three responses from GPT-4. For each response, we look at the full text and provide a rating (1-5) based on its quality and correctness. The criteria for the evaluation is determined before viewing the data:

5. Completely correct, no flaws
4. Correct but minor nitpicks (e.g., small error in explanation)
3. Some degree of error, or unclear, or incapable but mentions that this is the case
2. Clearly incorrect, perhaps hallucinates to justify
1. Nonsensical

The results for the pilot study are shown in Table 1. We find that setting (`temperature`=1.0, `top_p`=1.0) marginally outperforms other choices. Based on these results, we use (1.0, 1.0) as the hyperparameter settings for our GPT-4 checklist experiment. Note that these are also the default parameter values for GPT-4 in both the OpenAI API and playground.

temp	top_p	1(b)			2(a)			3(a)			4(a)			5(a)			Total
0	1.0	5	5	5	5	5	5	5	5	5	3	3	3	5	5	5	69
0.1	1.0	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
0.2	1.0	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
0.3	1.0	5	5	5	5	5	5	5	5	5	3	3	3	5	5	5	69
0.4	1.0	5	5	5	4	4	4	5	4	5	3	3	3	5	5	5	65
0.5	1.0	5	4	5	4	4	4	5	5	4	3	3	3	5	5	5	64
0.6	1.0	5	5	5	4	4	4	5	5	4	3	3	3	5	5	5	65
0.7	1.0	5	5	5	4	4	4	5	4	4	3	3	3	5	5	5	64
0.8	1.0	4	4	5	4	4	4	5	5	5	3	3	3	5	5	5	64

[1]Refer to the `temperature` and `top_p` sections of `https://platform.openai.com/docs/api-reference/chat/create`.

temp	top_p	1(b)			2(a)			3(a)			4(a)			5(a)			Total
0.9	1.0	5	5	5	4	4	4	5	4	5	3	3	3	5	5	5	65
1.0	**1.0**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**4**	**3**	**3**	**5**	**5**	**5**	**70**
1.1	1.0	4	4	5	4	4	4	5	5	4	3	2	2	5	5	5	61
1.2	1.0	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
1.3	1.0	5	3	5	4	4	4	5	5	5	4	3	3	5	5	5	65
1.4	1.0	5	4	5	4	4	4	4	5	5	3	3	3	5	5	5	64
1.5	1.0	5	5	5	4	4	4	5	5	5	3	2	3	5	5	5	65
1.6	1.0	5	4	4	4	4	4	4	4	4	3	2	2	5	5	5	59
1.7	1.0	1	4	5	4	4	4	4	1	1	1	1	1	5	5	4	45
1.8	1.0	4	1	1	1	1	1	1	1	1	1	1	1	1	5	5	26
1.9	1.0	4	1	1	4	1	4	1	1	1	1	1	1	5	1	1	28
2.0	1.0	1	1	1	4	1	4	1	1	1	1	1	1	1	5	1	25
1.0	0	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
1.0	0.1	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
1.0	0.2	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
1.0	0.3	5	5	5	4	4	4	5	5	4	3	3	3	5	5	5	65
1.0	0.4	5	5	5	4	4	4	5	5	5	3	3	3	5	5	5	66
1.0	0.5	5	5	5	4	4	4	4	5	4	3	3	3	5	5	5	64
1.0	0.6	5	5	5	4	4	4	5	4	5	4	4	5	5	5	5	69
1.0	0.7	5	4	4	4	4	4	3	4	5	3	3	3	5	5	5	61
1.0	0.8	5	5	5	4	4	4	5	4	4	3	3	3	5	5	5	64
1.0	0.9	5	4	5	5	5	5	4	5	4	3	3	3	5	5	5	66

Table 1: Results for our pilot to determine hyperparameters. Ratings for individual responses from GPT-4 are out of 5, and the total is out of 75.