

1 RESPONSE TO REVIEWER 8SRK

Dear Reviewer 8sRK, we sincerely thank you for your valuable feedback on our submission. Below is our responses to the concerns you raised. We have incorporated the following contents into the updated version of our paper, which we believe will help enhance the quality of our submission.

Q1: The presentation can be improved, especially lack in explaining why the method is better than others in an intuitive and easy to follow way. The authors argue that to avoid bias, the budget should be allocated uniformly, if so how could this method be more sample efficient than random? I guess the reason is if model A is much better than B and model B is much better than C, then it's not necessary to compare A and C a lot. But if thats the reason why this method is more sample-efficient, it would be contradictory to the uniform assumption. Could the authors provide more insight into this?

A1: Your question raises important points for discussion, and we will respond from two perspectives:

First, you mentioned leveraging the transitivity of model performance to reduce the preference budget. This approach is feasible when the optimization goal is to determine the relative ranking of models or to identify the best model (as demonstrated by the UCB algorithm (Zhou et al., 2024)). However, our goal is to precisely evaluate the true capability values of each model rather than their order or the selection of the optimal model. For example, for models A, B, and C, we aim to determine whether their capability values are $[0.9, 0.41, 0.4]$ or $[0.9, 0.89, 0.4]$, rather than just concluding that $A > B > C$. This distinction is crucial in practical applications because model deployment decisions often involve balancing performance and cost. If A's API price is significantly higher than B's and we know with precision that the performance gap between A and B is minimal (e.g., 0.9 vs. 0.89), we might prefer model B. Without precise capability values, such decisions become difficult when only ranking information is available.

Therefore, when our optimization goal is to determine the exact capability values of models, uniform sampling becomes intuitive. First, since tasks vary in difficulty, the same model may perform differently across tasks. As mentioned on line 213 in the original manuscript, the transitivity of model performance may not hold in some cases ($A > B, B > C, C > A$). Uniform sampling across candidates helps mitigate these biases and improves accuracy. Second, since we need the capability values for all models, uniformly sampling across them ensures balanced data collection, reducing the uncertainty in the estimated capability values for any particular model.

REFERENCES

Jin Peng Zhou, Christian K. Belardi, Ruihan Wu, Travis Zhang, Carla P. Gomes, Wen Sun, and Kilian Q. Weinberger. On speeding up language model evaluation. *CoRR*, abs/2407.06172, 2024. doi: 10.48550/ARXIV.2407.06172. URL <https://doi.org/10.48550/arXiv.2407.06172>.