Following the reviewer's suggestions, we have performed additional experiments to further demonstrate the effectiveness of the proposed TRIPLE framework. Preliminary results have been obtained mainly on two tasks due to the limited rebuttal time, while the full results will be reported in the revised paper.

Additional comparisons with NeuralUCB. First, we compare the NeuralUCB method used by INSTINCT with the proposed TRIPLE framework in Fig. 1. The result shows that on with limited budgets, the proposed TRIPLE framework using BAI-FB designs enjoys a steady gain over NeuralUCB.

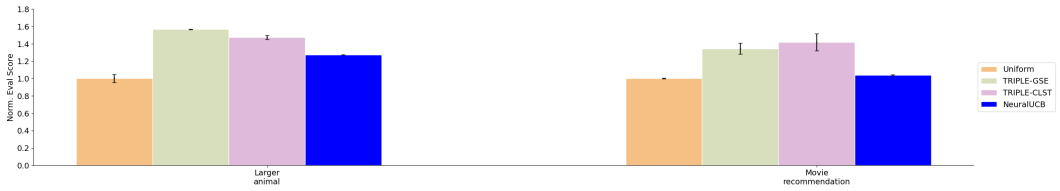

Figure 1: Performance comparisons between TRIPLE and NeuralUCB on GPT3.5 with 30 prompts and budgets of 150.

Additional results on Gemma and Mistral. We also evaluated the proposed TRIPLE framework across additional LLMs, in particular, Gemma and Mistral. The preliminary results are shown in Figs 2. It can be observed that similar to original results collected on GPT3.5 and Llama2, TRIPLE has shown remarkable performance gains compared to the baselines of Uniform or UCB, demonstrating its adaptability.

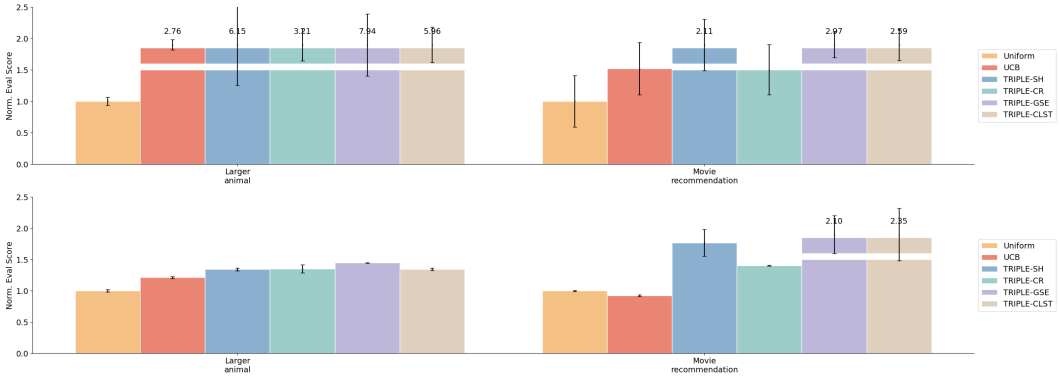

Figure 2: Performance comparisons on Gemma-7b (top) and Mistral-7b (bottom) with 30 prompts and budgets of 150.

Performance on large prompt pools. We have also checked the capability of TRIPLE on a larger prompt pool containing 100 candidates. Preliminary results are shown in Fig. 3. It can be observed that the superiority of TRIPLE is more pronounced compared with baselines of Uniform and NeuralUCB when handling this large prompt pool, demonstrating its scalability.

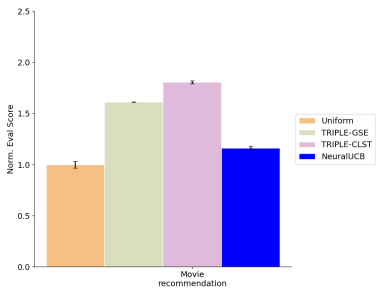

Figure 3: Performance comparisons on GPT3.5 with 100 prompts and budgets of 500.