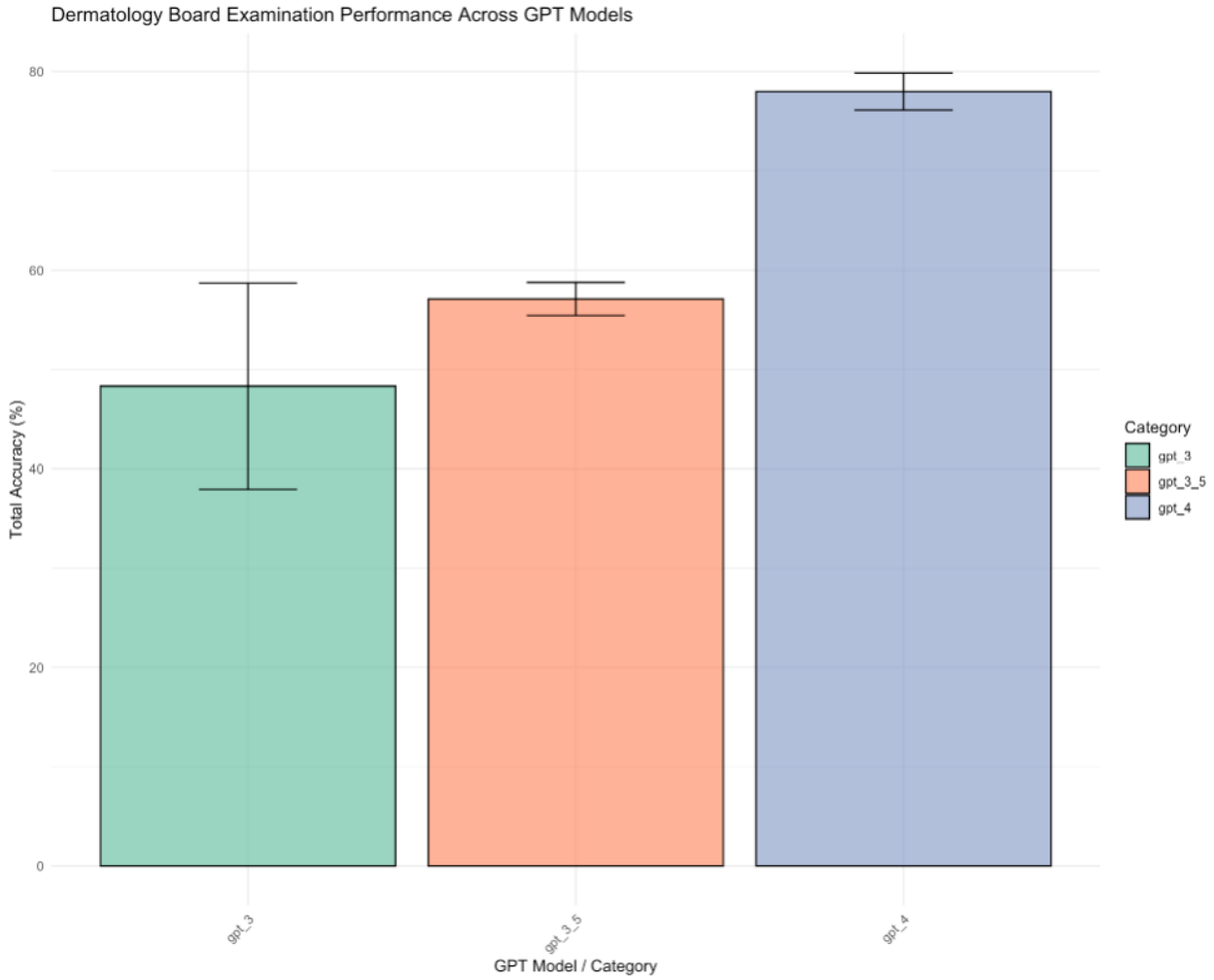

Supplemental Figure II: Total Accuracy for each GPT Model across All Questions Tested

Total score is defined as the total number of correct questions divided by the total number of questions as reported by each study of interest. One study reported on ChatGPT-3 performance. Nine studies reported on ChatGPT-3.5 performance. Ten studies reported on ChatGPT-4 performance. The bars represent 95% confidence intervals.