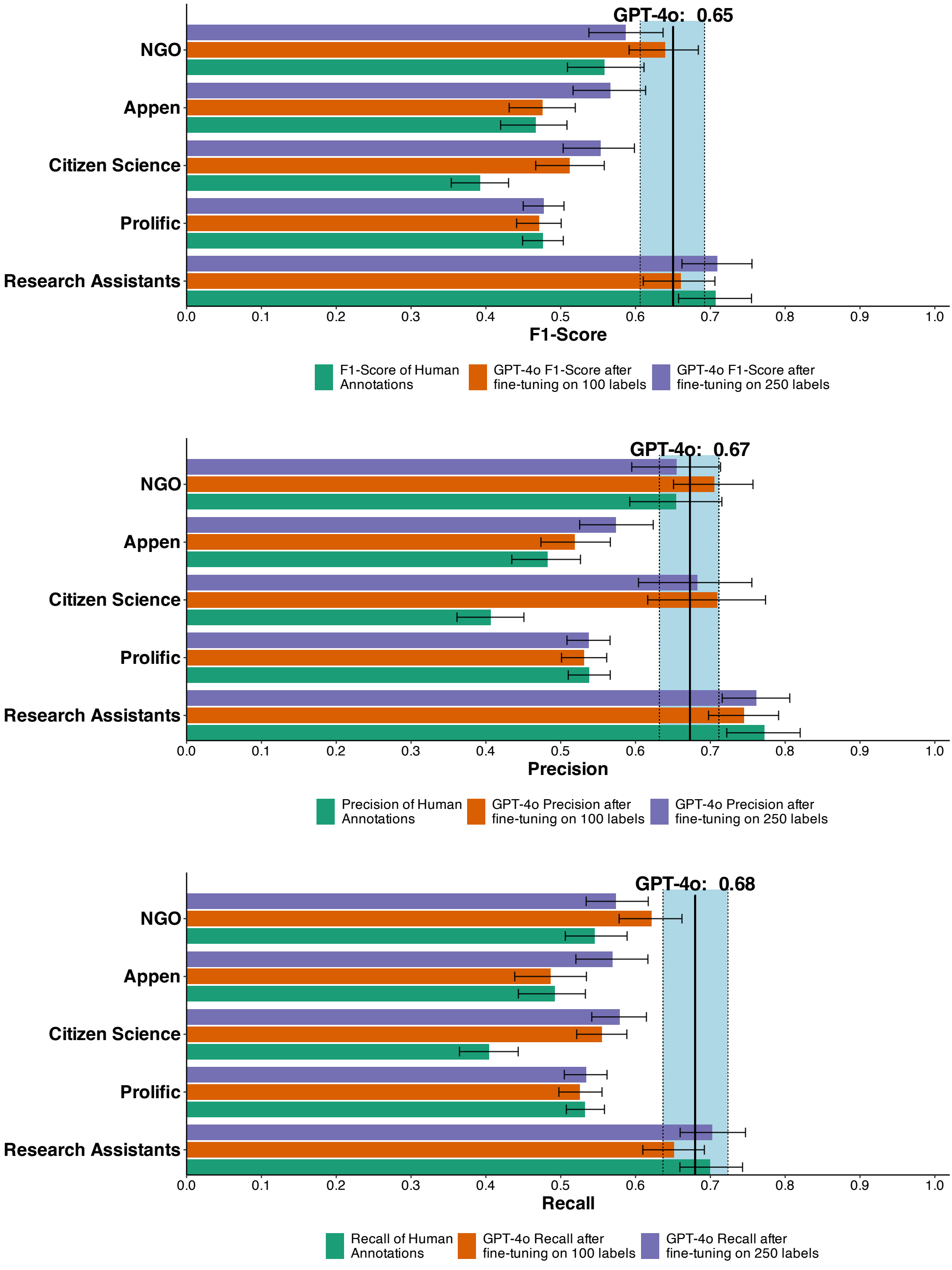
GPT-4o: 0.65
NGO
Appen
Citizen Science
Prolific
Research Assistants
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
F1-Score
F1-Score of Human Annotations
GPT-4o F1-Score after fine-tuning on 100 labels
GPT-4o F1-Score after fine-tuning on 250 labels
GPT-4o: 0.67
Precision
Precision of Human Annotations
GPT-4o Precision after fine-tuning on 100 labels
GPT-4o Precision after fine-tuning on 250 labels
GPT-4o: 0.68
Recall
Recall of Human Annotations
GPT-4o Recall after fine-tuning on 100 labels
GPT-4o Recall after fine-tuning on 250 labels