GPT-3.5 and GPT-4 successfully **classify student help requests** in programming classes with **no or minimal fine-tuning data**

Efficient Classification of Student Help Requests in Programming Courses Using LLMs

Jaromir Savelka Paul Denny Mark Liffiton Brad Sheese

Motivation

When students seek help from an automated assistant, they may ask a wide range of different types of queries related to their programming assignments. The ability to classify queries into distinct categories can have important educational implications.

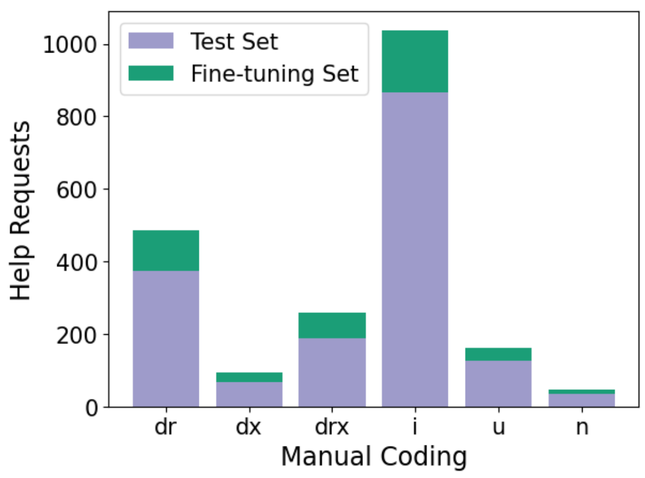
Research Questions

1. How accurately can GPT-3.5 and GPT-4 perform zero-shot classification of student help requests?
2. To what extent can classification performance improve by fine-tuning?

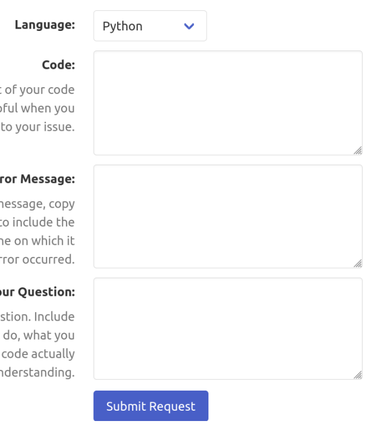
Dataset

The queries were independently coded by two of the authors into the following categories:

1. *Debugging*: Seeking help to resolve errors; sub-categorized into: the error (dr); the desired outcome (dx); or both (drx).
2. *Implementation* (i): Queries about implementing code to solve specific assignment problems.
3. *Understanding* (u): Queries focused on gaining an understanding of programming concepts.
4. *Nothing* (n): Queries that provided no error or meaningful issue.



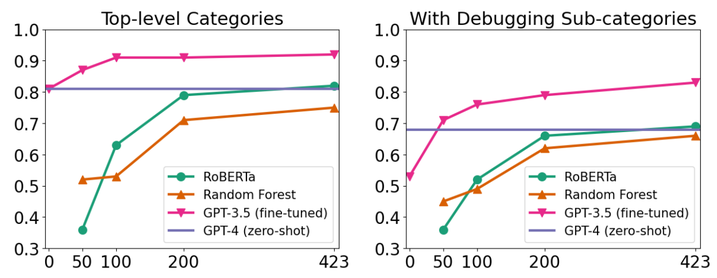
CodeHelp



Results

Query Category	Count	ZERO-SHOT GPT-3.5 P	R	F_1	ZERO-SHOT GPT-4 P	R	F_1	FINE-TUNED GPT-3.5 P	R	F_1
Debugging	630	.84	.91	.87	.90	.77	.83	.94	.92	.93
(error) – dr	374	.64	.02	.04	.69	.44	.54	.76	.90	.82
(outcome) – dx	67	.10	.09	.09	.23	.36	.28	.63	.36	.46
(error & outcome) – drx	189	.23	.75	.35	.50	.51	.50	.62	.46	.53
Implementation – i	867	.82	.89	.85	.78	.93	.85	.94	.93	.93
Understanding – u	127	.82	.24	.38	.74	.48	.58	.77	.85	.81
Nothing – n	35	.33	.06	.10	.50	.11	.19	.70	.89	.78
Overall	**1659**	**.82**	**.83**	**.81**	**.82**	**.82**	**.81**	**.92**	**.92**	**.92**
(debugging types)	1659	.67	.58	.53	.70	.70	.68	.83	.84	.83

Below is the comparison of GPT-4 and GPT-3.5 performance to random forest and RoBERTa base when trained/fine-tuned on progressively harder pool of data points up to 423.



Conclusions

1. GPT-3.5 and GPT-4 models achieved reasonable accuracy in a zero-shot setting.
2. Fine-tuning the GPT-3.5 model on a small amount of labeled data greatly improved its performance, reaching human-level accuracy.

Scan QR code to get the full paper

Carnegie Mellon University

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND