Thrilled by Your Progress! Large Language Models (GPT-4) No Longer Struggle to Pass Assessments in Higher Education Programming Courses (2023)

Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, Majd Sakr

Selina Cheng

Overview

- **What**:
  - Assess performance on typical assessments in intro and intermediate programming courses
- **Why:**
  - GPT-4 performance in comparison with earlier GPT models reveals significant improvement
  - Now on par with students

Methods

- Sourced data from **3** current Python programming courses
  - **Basics**: Python Essentials - Part 1 (**PE1**)
    - MCQ
  - **Intermediate**: Python Essentials - Part 2 (**PE2**)
    - MCQ
  - **Practical Programming with Python** (**PPP**)
    - MCQ (< 20%)
    - Programming projects with subtasks (80%)

Experimental Design

- **GPT-3**, **GPT-3.5**, and **GPT-4**
- MCQ submitted one by one
- Coding tasks with instructions submitted
  - Auto-grader assigned score and generated feedback
  - Revised solution
  - Repeat until unchanged or full score

Results

Table 3: PE1 results. The graded assignments are colored; green and check mark indicate passing while red means failing.

	Quizzes			Tests		
Module Topic	GPT-3	GPT-3.5	GPT-4	GPT-3	GPT-3.5	GPT-4
Introduction to Python and programming	8/10 (80.0%)	10/10 (100%)	10/10 (100%)	6/10 (60.0%)	✓ 9/10 (90.0%)	✓ 10/10 (100%)
Data types, variables, I/O, operators	6/10 (60.0%)	10/10 (100%)	10/10 (100%)	6/20 (30.0%)	10/20 (50.0%)	✓ 18/20 (90.0%)
Booleans, conditionals, loops, operators	3/10 (30.0%)	7/10 (70.0%)	10/10 (100%)	6/20 (30.0%)	12/20 (60.0%)	✓ 16/20 (80.0%)
Functions, data structures, exceptions	6/12 (50.0%)	9/12 (75.0%)	9/12 (75.0%)	7/22 (31.8%)	12/22 (54.5%)	✓ 20/22 (90.9%)
Completion (Summary Test)	-	-	-	7/35 (20.0%)	17/35 (48.6%)	✓ 27/35 (77.1%)
Course Total	**23/42 (54.8%)**	**36/42 (85.7%)**	**39/42 (92.4%)**	**32/107 (29.9%)**	**60/107 (56.1%)**	**91/107 (85.0%)**

Results

Table 4: PE2 results. The graded assignments are colored; green and check mark indicates passing while red means failing.

	Quizzes			Tests		
Module Topic	GPT-3	GPT-3.5	GPT-4	GPT-3	GPT-3.5	GPT-4
Modules, packages, and PIP	3/10 (30.0%)	6/10 (60.0%)	10/10 (100%)	10/18 (55.6%)	✓ 14/18 (77.8%)	✓ 17/18 (94.4%)
Strings, string list methods, exceptions	7/10 (70.0%)	6/10 (60.0%)	9/10 (90.0%)	4/15 (26.7%)	✓ 11/15 (73.3%)	✓ 13/15 (86.7%)
Object-oriented programming	7/10 (70.0%)	8/10 (80.0%)	9/10 (90.0%)	4/17 (23.5%)	✓ 12/17 (70.6%)	✓ 15/17 (82.4%)
Miscellaneous	8/12 (66.7%)	9/12 (75.0%)	11/12 (91.7%)	4/16 (25.0%)	9/16 (56.2%)	✓ 15/16 (93.8%)
Completion (Summary Test)	-	-	-	11/40 (27.5%)	26/40 (65.0%)	✓ 35/40 (87.5%)
Course Total	**25/42 (59.5%)**	**29/42 (69.0%)**	**40/42 (95.2%)**	**33/106 (31.1%)**	**72/106 (67.9%)**	**95/106 (89.6%)**

Results

Table 5: PPP results. The tests contribute to the grade, typically by no more than 20%. Since in PPP tests themselves do not determine pass or fail no colors are used.

	Quizzes			Tests		
Module Topic	GPT-3	GPT-3.5	GPT-4	GPT-3	GPT-3.5	GPT-4
Python basics and introduction to functions	12/30 (40.0%)	21/30 (70.0%)	27/30 (90.0%)	4/12 (33.3%)	9/12 (75.0%)	10/12 (83.3%)
Control flow, strings, input and output	8/22 (36.4%)	10/22 (45.5%)	16/22 (72.7%)	3/11 (27.3%)	8/11 (72.7%)	10/11 (90.9%)
Python data structures	9/18 (50.0%)	10/18 (55.6%)	14/18 (77.8%)	4/14 (28.6%)	9/14 (64.3%)	10/14 (71.4%)
Object-oriented programming	6/14 (42.9%)	7/14 (50.0%)	11/14 (77.6%)	4/11 (36.4%)	10/11 (90.9%)	11/11 (100%)
Software development	9/19 (47.4%)	12/19 (63.2%)	16/19 (84.2%)	5/10 (50.0%)	7/10 (70.0%)	10/10 (100%)
Data manipulation	6/17 (35.3%)	9/17 (52.9%)	13/17 (76.5%)	5/13 (38.5%)	5/13 (35.5%)	8/13 (61.5%)
Web scraping and office document processing	5/10 (50.0%)	5/10 (50.0%)	5/10 (50.0%)	0/5 (0.0%)	3/5 (60.0%)	3/5 (60.0%)
Data analysis	6/22 (27.3%)	17/22 (77.3%)	18/22 (81.8%)	0/5 (0.0%)	2/5 (40.0%)	2/5 (20.0%)
Course Total	**61/152 (40.1%)**	**91/152 (59.9%)**	**120/152 (78.9%)**	**25/81 (30.9%)**	**53/81 (65.4%)**	**64/81 (79.0%)**

Results

Table 8: Coding tasks results. Max score is the maximum score achievable from a task. First score is the score after first submission. Resubs is the number of re-submissions after the first submission before the full score or no-change impasse were reached. Final score is the score after feedback.

Project Topic	Tasks (skills)	Max	GPT-3.5 1st	GPT-3.5 Resubs	GPT-3.5 Final	GPT-4 1st	GPT-4 Resubs	GPT-4 Final
	Course Total	**760**	**407**	**59**	**505**	**545**	**56**	**634**
			53.6%		**66.4%**	**71.7%**		**83.4%**

Takeaway

GPT-4 now performs reliably on traditional assessments, which has significant ethical implications in education.