AI and Testing

EHSAN SAEEDIZADE

MARCH-2025

Goals

UNDERSTAND AI'S ROLE IN SOFTWARE TESTING

DISCUSSION AND EXPLORE AI'S CAPABILITIES AND LIMITATIONS

ENGAGE IN HANDS-ON LEARNING

Why This Is a Hot Topic?

1. Rising Complexity of Software
   - Manual testing is time consuming

2. Growing Use of Large Language Models Engineering
   - Test case generation
   - Bug localization
   - Debugging (LLM-driven testing process)

But AI Is Not Perfect! (YET?)

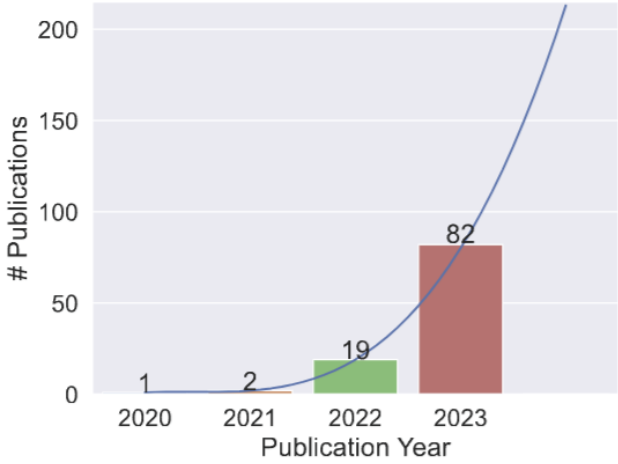


Fig. 3: Trend in the number of papers with year

"Software testing with large language models: Survey, landscape, and vision." *IEEE Transactions on Software Engineering* (2024).
"Evaluating large language models for software testing." Computer Standards & Interfaces 93 (2025): 103942.

Discussion

- How was your experience using AI for test generation?
- What were the strengths and weaknesses of AI-generated tests?
- Did AI identify edge cases correctly?
- Did AI generate any incorrect tests?
- What improvements/points would you suggest to consider for better result?

AI in Software Testing – Capabilities

1. **Higher Readability & Usability**
   - Developers found AI-generated tests easier to understand.
2. **Decent Code Coverage**
   - AI-generated unit tests achieved comparable test coverage to manually written tests.
   - Effectively complement manual testing by detecting additional errors.
3. **Possible Improvements**
   - With iterative refinement (e.g., ChatTester), AI-generated tests improved compilability by 34.3% and assertion correctness by 18.7%.

AI can significantly improve test automation but still needs human verification.

"**No more manual tests? evaluating and improving chatgpt for unit test generation**." arXiv preprint arXiv:2305.04207 (2023).

AI in Software Testing – Limitations

1. **Correctness Issues**
   - 24.8% of AI-generated tests failed execution due to syntax or assertion errors.
   - AI sometimes generated invalid assertions that didn't match program logic.
2. **Security Risks and Mocking Issues**
   - AI fails at generating security tests like SQL injection detection, Mock when needed, unless explicitly trained.
   - Misses edge cases that are critical in penetration testing.

TABLE 3: Performance of unit test case generation

Dataset	Correctness	Coverage	LLM	Paper
5 Java projects from Defects4J	16.21%	5%-13% (line coverage)	BART	[26]
10 Jave projects	40%	89% (line coverage), 90% (branch coverage)	ChatGPT	[36]
CodeSearchNet	41%	N/A	ChatGPT	[7]
HumanEval	78%	87% (line coverage), 92% (branch coverage)	Codex	[39]
SF110	2%	2% (line coverage), 1% (branch coverage)	Codex	[39]

Note that, [39] experiments with Codex, CodeGen, and ChatGPT, and the best performance was achieved by Codex.

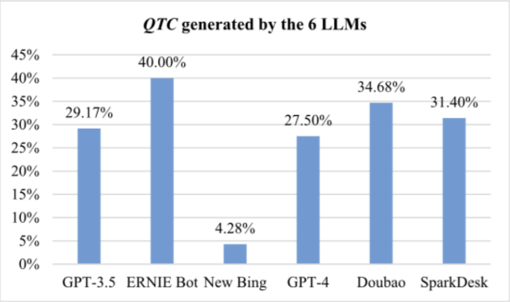
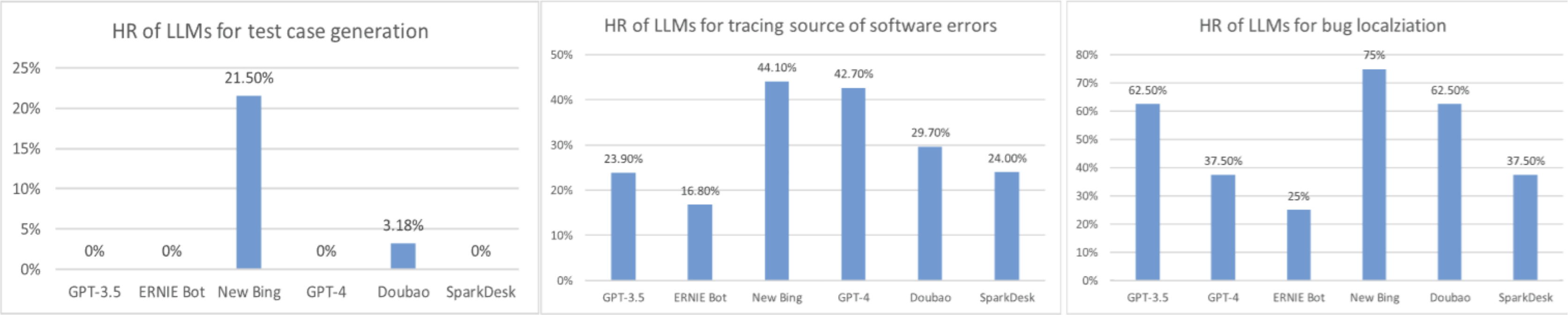


Fig. 2. Quality of test cases (*QTC*) generated by the six large language models

"Software testing with large language models: Survey, landscape, and vision." *IEEE Transactions on Software Engineering* (2024).
"Evaluating large language models for software testing." Computer Standards & Interfaces 93 (2025): 103942.

AI in Software Testing – Limitations

1. **Contextual Understanding is Limited**
   - AI often misinterprets business logic, leading to functionally useless test cases.
   - AI is prone to hallucinations!



(a) (b) (c)

AI-generated tests are not always reliable—human oversight is needed to correct and refine them.

"Software testing with large language models: Survey, landscape, and vision." *IEEE Transactions on Software Engineering* (2024).
"Evaluating large language models for software testing." Computer Standards & Interfaces 93 (2025): 103942.

Class Activity- AI-Based Testing for Authentication Service

Similar to HW3, maybe a bit complex code to practice mocking ☺

You will test a simplified Authentication Service that includes:

- AuthService: Handles login, signup, and session management.
- User: Represents individual user accounts.
- UserStorage: Handles database queries. (Which needs to be mocked)

Steps for the Activity

1. Review and Understand Code
2. Generate AI-Based Test Cases
3. Run & Evaluate the Tests

Comparison of AIs

ChatGPT-4o

8 test cases, 5 failed

Name	Stmts	Miss	Cover	Missing
auth_service.py	49	32	35%	13-14, 18-20, 24-26, 30-39, 42-57, 60-61
user.py	62	25	60%	35-42, 50-53, 56-61, 64-70, 73-76
TOTAL	170	79	54%	

ChatGPT

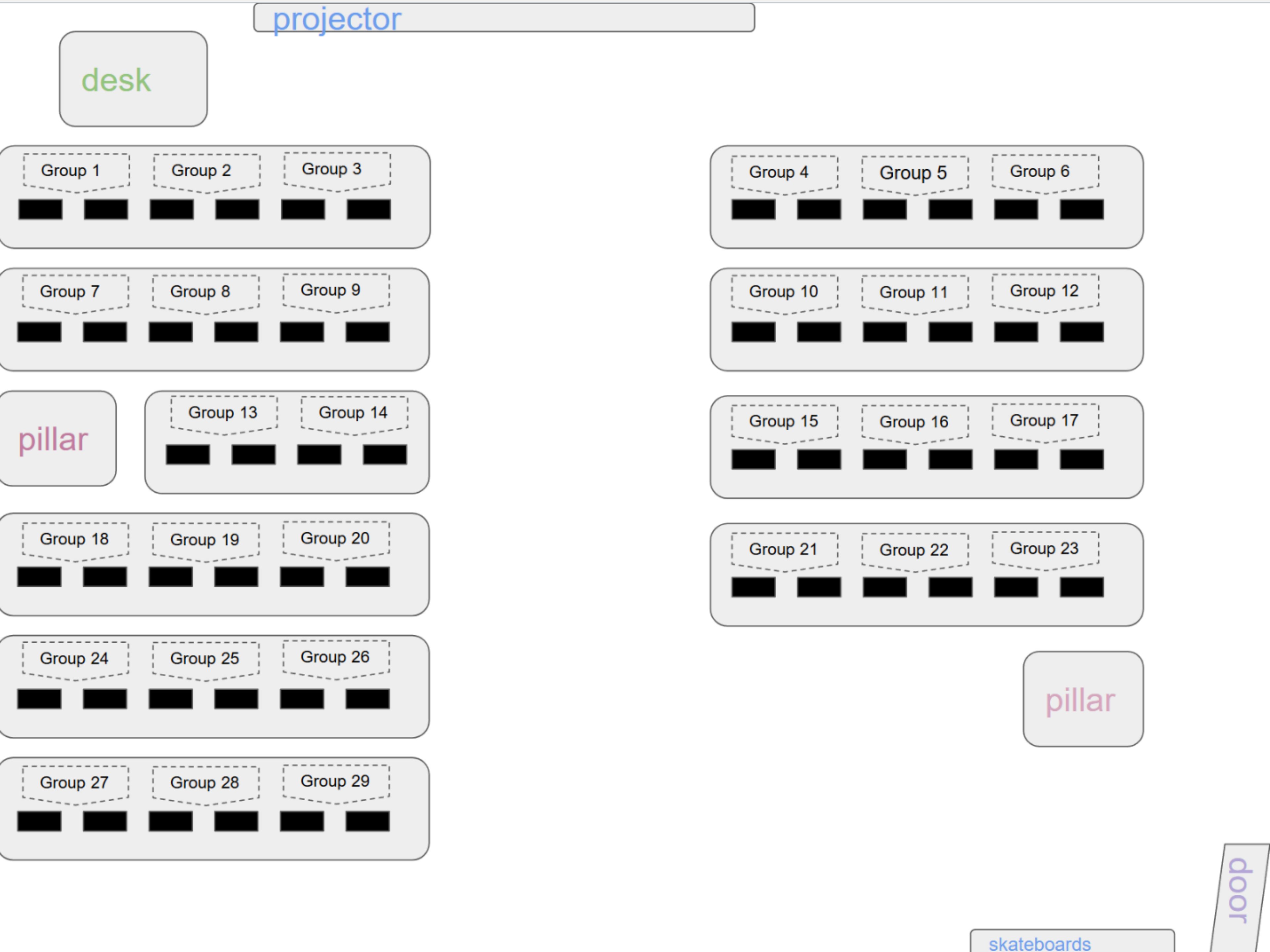
16 test cases, 3 failed

Name	Stmts	Miss	Cover	Missing
auth_service.py	49	9	82%	24-26, 32, 34, 36, 50-52
user.py	62	25	60%	35-42, 50-53, 56-61, 64-70, 73-76
TOTAL	215	53	75%	

Copilot

8 test cases, all pass

Name	Stmts	Miss	Cover	Missing
auth_service.py	49	8	84%	32, 34, 36, 44, 47, 50-52
user.py	62	35	44%	10-11, 14-32, 35-42, 45-53, 56-61, 64-70, 73-76
TOTAL	172	44	74%	


projector
desk
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
Group 7
Group 8
Group 9
Group 10
Group 11
Group 12
pillar
Group 13
Group 14
Group 15
Group 16
Group 17
Group 18
Group 19
Group 20
Group 21
Group 22
Group 23
Group 24
Group 25
Group 26
pillar
Group 27
Group 28
Group 29
door
skateboards