Abhishek Varghese (he/him)
Author
Posted Oct 30 3:43pm

Discussion Topic: In-Class Exercise 2 clarificationsIn-Class Exercise 2 clarifications
Hi Everyone,

The assignment is live on Gradescope. I'll also add it to the Files tab.

There were a few confusions today on what the assignment is asking to do, so I've added a workflow summary (These are not instructions; please refer to the exercise document first, and then check this for an updated overview on what to do):

Part 1:
Takeaway: One metric to test the strength of the test suite on a given code is to check coverage. Bugs can hide in lines that were never run during the tests.
Method: You have 40 LLM generated solutions and their tests. Run the coverage test on each of those and create a table. This part does not involve any LLM calls, you just need to setup the coverage library for your tests.
Choose 2 among those 40 for Part2 and Part3 using the criterion given in the exercise document. Ideally you want to select 2 that has the highest percentage of test cases passed and lowest coverage. These are the two problems that you will be using for Parts 2 and 3 of this assignment.

Part 2 :
Question : Can LLMs be used to generate test cases to improve coverage?
Method : For each of those 2 (Problem statements, LLM generated code, benchmark test suite), try and improve the coverage with the instructions given in the document. Iterate until you get a high coverage. You now have 2 (Problem statement, LLM generated solution, LLM generated test suite) for part 3.

Part 3 :
Question : Even though the LLM generated tests have high coverage do they actually capture bugs?
Method : For each of these 2 (Problem statement, LLM generated solution, LLM generated test suite), check how well the tests capture bugs.

Please reach out if this is still unclear.

All the best,
Abhishek

Abhishek Varghese (he/him)
Author
Posted Nov 6 3:26pm | Last edited Nov 6 3:28pm

Discussion Topic: In-Class Exercise 2 FAQ : What to do if almost all my solutions have 100% test case passed and 100% coverage.
In-Class Exercise 2 FAQ : What to do if almost all my solutions have 100% test case passed and 100% coverage
Hi all,

If you are having issues with your LLMs being too good, or the problems you worked on in assignment 1 not being complex enough to have iterative improvements for test cases, you can think about using a more complex problem from a different test-bench than what you originally used (or a different LLM).

If you choose to take the route of using new (problem statement x test-bench x LLM solution) :
1. You will still include everything asked for in part 1 (that helps you understand how to setup coverage tests in python).
2. Then you can choose 2 problems of your choice from outside the set of the 40 you used in part 1.
3. Make sure you have a test-bench for these two new (and more complex) problems.
4. Include in your report details about the source of the new problem statements, a summary (a line or two) of the task asked in the problem statement, LLM family used, LLM prompting technique used, %test passed, %coverage.

I'll keep the comments open on this one for any further questions.

Best
Abhishek

2 Replies

XiaoFan Lu (he/him)
Nov 6 6:07pm | Last reply Nov 10 11:41am

Reply from XiaoFan Lu
"choose 2 problems of your choice from outside the set of the 40 you used in part 1."

For exercise 1 we only have 10 problems, but for each problem we need solution from 2 different LLMs and 2 different prompt technique, (that doesn't include our own innovation)? is that how 40 comes from? 10 * 2*2?

if that is true can we just do one to one mapping? for each problem pick one solution generated in exercise 1? I don't understand the purpose of having test cases test against multiple similar solutions.

Reply to post from XiaoFan Lu Reply
Abhishek Varghese (he/him)
Author
Nov 10 11:41am

Reply from Abhishek Varghese
Hi XiaoFan,
Thank you for contacting me via piazza. Just to give more clarification to what I mentioned there, for each problem you have 4 different solutions. Hence the tests will give you 4 different coverage results, which you need to report. As per what's been asked in the exercise document, if you give just one report per problem, you will lose points. Hope this brings more clarity to what Part1 is asking you to do.
Best
Abhishek

Piazza:

1) **Question about recent exercise 2 announcement**
Updated 3 days ago by XiaoFan Lu
Clarification needed on test case scope:
We have 10 problems, each with 4 solutions (2 LLMs × 2 prompts = 40 total solutions).
Question 1: Should we write:

- A) 10 test suites (one per problem), where each suite tests all 4 solutions for that problem?
- B) 40 separate test suites (one per solution)?

Question 2: If Option A, can we simplify by selecting just 1 solution per problem (10 solutions total) instead of testing all 40?
I understand testing multiple implementations teaches test robustness, but I want to confirm if that's a required learning objective for this assignment or if we can focus on test coverage/quality with fewer solutions.

Instructors' Answer
Updated 3 days ago by Abhishek Varghese
You don't have to build 40 separate test suites. You can build one for each (10 separate). But you have to report the coverage results for all 40 solutions (one per LLM solution)

2) **Question on convergence criterion in case of small code size**
Updated 3 days ago by Anonymous Calc
Hi, I’m getting large coverage jumps because my code is small (e.g., 88% → 94% → 100%). The assignment says convergence is when the increase is <3%, but in my case, even small test additions exceed that. Should I still follow the <3% rule, or is it fine to stop once I reach 100% coverage?

Instructors' Answer

Updated 2 days ago by Abhishek Varghese
100% coverage is still a convergence, as there is no scope of improvement after that. There is no need for more iterations after 100% coverage.

3) **HTML to git?**
Updated 3 hours ago by Anonymous Beaker
I noticed that for part 1, the size of the html generated (using coverage html) is very large. Is it alright to generate a text summary along with the xml report? Or are we required to push the html reports as well?

Instructors' Answer
Updated 3 hours ago by Abhishek Varghese
You can add a text summary and screenshots to your report. But make sure the code you're pushing to GitHub has all the components so that we can generate the reports on our end.

4) **Exercise 2 outline missing in gradescope**
Updated 1 week ago by Anonymous Gear
Hello,

The outline for Exercise 2 isn't set up in Gradescope for submission. Is that expected, or should we select all pages manually and submit?

Instructors' Answer
Updated 1 week ago by Abhishek Varghese
Thank you for bringing this to my attention. The outline has now been added to the assignment.

5) **Question on convergence criterion in case of small code size**
Updated 3 days ago by Anonymous Calc
Hi, I'm getting large coverage jumps because my code is small (e.g., 88% → 94% → 100%). The assignment says convergence is when the increase is <3%, but in my case, even small test additions exceed that. Should I still follow the <3% rule, or is it fine to stop once I reach 100% coverage?

Instructors' Answer
Updated 2 days ago by Abhishek Varghese
100% coverage is still a convergence, as there is no scope of improvement after that. There is no need for more iterations after 100% coverage.

Anonymous Comp 2 hours ago
I believe we are also concerned about the 6% jump rather than 3 % jump as asked in the assignment, shouldn't convergence be fine no matter the percentage as long as coverage improves after each iteration

0
Abhishek Varghese 27 minutes ago
Hi, I see the confusion.
The convergence criteria is not the same as preferred improvement per iteration.

When you train a machine learning model, with CrossEntropyLoss as your metric, it may start fast 2 -> 1.73 -> 1.5 -> 1.10 -> 0.09 ......

Eventually it converges and the gradient updates no longer make significant change to the loss function 0.0510 -> 0.0500 -> 0.0499 -> 0.0498 -> .....

As a result any more time spent training is met with minimal improvement in the model performance, and its usually best to terminate the training. Something similar is going on here, and the exercise is asking you to use Line/Branch coverage as your metric until convergence and the convergence criteria is when the metric does not increase by more than 3%. Does that help make more sense?