The result for question 5 can be perceived as least bad to worst. In that case, we can look at the detailed result as:

1. Codestral-Mamba

Execution 1:

Best result out all. Corrects all the bugs in the code.

Does not explain all the bugs but correct all of them.

Execution 2:

Left a bug at the end of the code

Execution 3:

Fixed all the bugs in the code, but the explanation which is provided is unrelated to the bugs fixed, i.e., the bug which was fixed was different, and the explanation provided was different.

1. Granite-code:8b-instruct

Execution 1-3:

Same result. Could only debug the first class, does not provide output of the second class and main function.

1. Granite-code:8b-dense

Execution 1:

Does not read the full code either but reads a few more line and fixes one more bug than 8b-instruct. In one bug, it removes the buggy line instead of fixing it in the code.

Execution 2:

Reads a few more lines and fixes a couple more bugs but is still not able to read and finish the full code.

1. Llama3.1

Execution 1:

Fixes all the bugs. Also adds some exception handling to the code.

Execution 2:

Fixes all the bugs, also adds some documentation to the code.

Execution 3:

Fixes all the bugs, also does some unnecessary code changes like changing the name of the function. Does not do documentation and error handling in this case.

NOTE: In all the three scenarios of Llama, it did not provide a satisfactory explanation. It gave details of what it added to the code (like error handling) but does not provide good explanation of the bug fixes.

1. Starcoder2:  
     
   Simply returns the error from the execution, instead of fixing the bugs.

The following table contains a summary of how the models performed. Each cell contains the total number of bugs fixed by the model out of the total 10 bugs.

**Summary Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name-> | Granite-code:8b-instruct | Granite-dense:8b | Llama3.1:8b | Codestral-mamba:7b | Starcoder2:7b |
| Execution 1 | 3/10 | 4/10# | 10/10 | 10/10 | 0/10 |
| Execution 2 | 3/10 | 6/10 | 10/10 | 9/10 |  |
| Execution 3 | 3/10 | 6/10 | 10/10 | 10/10\* |  |

# It removed one line containing bug instead of fixing it

\*- Bugs were removed in the code but incorrect explanation/ explanation for irrelevant sections was given