|  |  |  |  |
| --- | --- | --- | --- |
| **ChatGPT** | | **Llama 3** | |
| **Unit Test Fault** | **Input Program Fault** | **Unit Test Fault** | **Input Program Fault** |
| 6 / 8 | 2 / 8 | 10 / 12 | 2 / 12 |

**Distinct Failures for ChatGPT & Llama 3 with Prompt 2**

Distinct failures were counted instead of all failures as this would create an unfair comparison between the two large language models. This is because Llama 3 often had more failures than ChatGPT.