|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric No** | **Metric Name** | **Description** | **Measurement** | **Formula** |
| 1 | **Prompts Required for Complete Step Definitions** | Count the number of prompts needed to generate all step definitions with as much code code as possible. | Integer value | - |
| 2 | **DAO Utilization and Accuracy** | Evaluate the usage, accuracy, and reminders for DAOs by the AI. | Binary (1/0), Integer values | DAO Utilization Score = DAO Usage × (Accuracy of DAOs Guessed - Reminders Required) |
| 3 | **Service Utilization and Accuracy** | Evaluate the usage, accuracy, and reminders for services by the AI. | Binary (1/0), Integer values | Service Utilization Score = Service Usage × (Accuracy of Services Guessed + new acceptable services - Reminders Required) |
| 4 | **Accuracy of Domain Classes Guessed/Used** | Measure the percentage of domain classes guessed or used correctly by the AI | Percentage value | Accuracy of Domain classes Percentage = (Number of correct guessed Domain classes / Total Number of Domain classes) × 100 |
| 5 | **Acceptable Scenario Solutions** | Measure the percentage of scenarios that are acceptable based on completeness, accuracy, functionality, and integration. | Percentage value | Acceptable Scenario Percentage = (Number of Acceptable Scenarios / Total Number of Scenarios) × 100 |
| 6 | **Better Than Acceptable scenario Solutions** | Count the instances where the LLM provided better than acceptable solutions for a scenario | Integer value | - |
| 7 | **Accuracy of Personas and Hard-Coded Variables** | Measure the percentage of step definitions that the AI used the provided personas and hard-coded variables in the generated code(georgeRed, harryPotter….) | Percentage Value | Accuracy of Personas = (Number of step definitions personas or hard-coded variables were used/Total Number of Step definitions) × 100 |
| 8 | **Accuracy of Attributes in Correctly Guessed Domain Classes** | Measure how accurately the AI guesses or uses attributes in the correct domain classes (phase 1,2) | Integer values | Attributes Accuracy Score = (Correct Attributes / (Correct Attributes + Missing Attributes + Extra Attributes)) × 100 |
| 9 | **Accuracy of Methods (Excluding Getters and Setters) in Correctly Guessed Domain Classes** | Measure how accurately the AI guesses or uses methods (excluding getters and setters) in domain classes(phase 1,2,3) | Integer and binary values | Methods Accuracy Score = (Correct Methods) / (Correct Methods + Missing Methods + Extra Methods) × 100 |
| 10 | **Logical Consistency of Generated Code** | Evaluate if the generated code makes logical sense in the context of the application, even if it compiles. | Binary value (1/0) | - |
| 11 | **Repetition/additional explanation** | Count the instances where the LLM needed addiotional explanation or repetition of prompts | Integer Value | - |

# Evaluation Table