**Methodology for Comparing GenAI Models**

The comparison will determine how accurately each model performs the assigned tasks in the two areas:

* Direct Prompts: Measure the degree to which each model can produce direct instructions or task-specific output (e.g., data quality reports, model training summaries) based on human-readable prompts.
* Code Generation in Python: Assess the model's ability to generate Python code that is functionally and syntactically correct for a given task, and the correctness, efficiency, and best practices of the generated code.

**Levels of Comparison**

For both direct prompts and Python code generation, we will follow these levels of comparison:

* Fully Capable: The model produces fully correct, complete, and working outputs or code with no visible errors or omissions. The content generated is relevant to the task; it reflects a high level of understanding.
* Capable Partials: The model does a fairly good job of attempting the task but either lacks some critical details, contains some minor errors, or needs some minor adjustment. The output is understandable and generally useful.
* Limited Capability: The model outputs are substantially incomplete, inaccurate, or incoherent. It may not respond to the primary task elements, or generate a result whose quality is low enough to require major manual fixes or rework.
* Incapable: The model can't give the related result for the task or deliver outcomes that are far off from what is required; what it produces is either not relevant or, alternatively, not applicable to the given task.

Comparison Table 1: Direct Prompts

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| **Data Quality Report Generation** | Fully | Partially | Limited | Incapable |
| **Data Quality Issues & Handling Strategy** | Fully | Fully | Partially | Limited |
| **Apply Data Quality Handling Strategies** | Partially | Fully | Partially | Incapable |
| **Dataset Split (80% Training, 20% Testing)** | Fully | Fully | Fully | Partially |
| **Train a Machine Learning Model** | Fully | Partially | Fully | Limited |
| **Test the Machine Learning Model** | Fully | Fully | Partially | Incapable |
| **Dataset Cleaning Instructions (Dataset 1)** | Fully | Fully | Partially | Limited |
| **Dataset Cleaning Instructions (Dataset 2)** | Fully | Partially | Fully | Incapable |

Comparison Table 2: Generating Python Code

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| **Data Quality Report Generation** | Fully | Partially | Limited | Incapable |
| **Data Quality Issues & Handling Strategy** | Fully | Fully | Partially | Limited |
| **Apply Data Quality Handling Strategies** | Partially | Fully | Partially | Incapable |
| **Dataset Split (80% Training, 20% Testing)** | Fully | Fully | Fully | Partially |
| **Train a Machine Learning Model** | Fully | Partially | Fully | Limited |
| **Test the Machine Learning Model** | Fully | Fully | Partially | Incapable |
| **Dataset Cleaning Instructions (Dataset 1)** | Fully | Fully | Partially | Limited |
| **Dataset Cleaning Instructions (Dataset 2)** | Fully | Partially | Fully | Incapable |