# Evaluation Report for the Gender Chat-bot of NISR

# *Proposed by SJTU Global Challenge Project*

*“From Shanghai to Rwanda 2.0”*

*August 2025*

# 1 Introduction

This evaluation report provides a comprehensive assessment of the Chat-Bot developed by the National Institute of Statistics of Rwanda (NISR), which is dedicated to systematically organizing gender-related surveys and statistical reports, thereby offering an interactive AI interface for both professionals and the general public.

The assessment seeks to establish benchmarks for enhancing user experience before the Chat-Bot's operational deployment which primarily focuses on two key aspects: first, the evaluation of the model's internal capabilities, including testing its linguistic proficiency, accuracy, and response speed; second, a comparative analysis benchmarking this model against mainstream large-scale models to highlight its distinctive advantages in the context of gender governance in Rwanda.

# 2 Methodology

The Graphical Users Interface (GUI) of the gender chat-bot is shown in Figure.2 where the methodology aligns with conventional approaches employed by mainstream large-scale AI models.

## 2.1 General Methodology

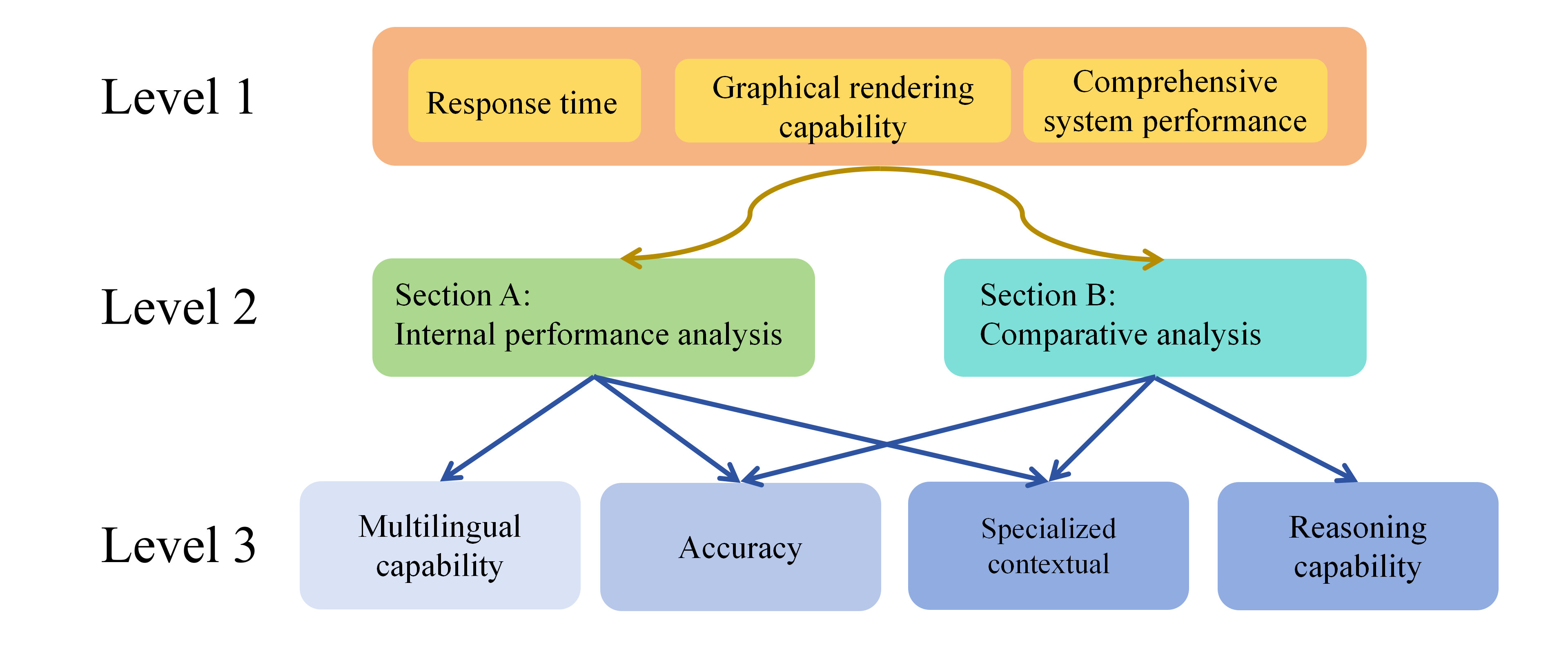


Figure 1. Comparison of the performance with different language input

As shown in Figure.1, A three-tiered evaluation framework is proposed. The top-level indicators encompass response time, graphical rendering capability, and comprehensive system performance evaluation. The second tier comprises an internal performance analysis of the Chat-Bot system, along with a comparative analysis between the Chat-Bot and internationally prevalent generalized AI Bot systems. The third tier delves into granular operational metrics, including accuracy assessment, multilingual proficiency analysis, associative reasoning capability evaluation, and specialized contextual scenario analysis.

## 2.2 Internal test

The Graphical Users Interface (GUI) of the gender chat-bot is provided for professional and public users where the methodology aligns with conventional approaches employed by mainstream large-scale AI models.

The testing of Chat-Bot consists of two components: language testing and misleading testing. The first part examines the bot's response performance in Chinese, English, Kinyarwanda, and multilingual mixed scenarios. The second part employs misleading and ambiguous queries to test the bot, evaluating its robustness when handling vague inputs.

## 2.3 Comparison test

We compared three models: Chat-Bot, DeepSeek-V3, and ChatGPT-4o. To ensure fairness in testing, we first input the required PDF documents into the latter two models before using them.

In the multi-model comparative analysis, beyond accuracy, we focused on examining each model's capability in processing in-depth information. For instance, we evaluated content extraction involving statistical definitions and conducted queries regarding data trend analysis within tabular content.

# 3 Results and Discussions

## 3.1 Strengths

### 3.1.1 Correctness in Fact-Checking

The model demonstrates a strong ability to correct inaccurate information provided by the user. It will insist on the correct information and cite its sources to validate its response, even when the user insists on an incorrect assertion.

### 3.1.2 Adherence to Training Data

The model appears to stay strictly within the confines of its training data. It will indicate when it does not know an answer rather than generating a speculative response. (Note: More testing is needed to confirm this behavior is consistent).

In the extreme test process, when asked questions completely unrelated to the article, the chatbot gave clear responses: “Out of scope.”, which generalized model fails to do that.

### 3.1.3 High Accuracy in Textual Data

the model performs well in accuracy when the required information is present directly within the sentences of its source documents.

As shown in the Table 1,the accuracy indicates the chatbot's high reliability when the required information is present directly within the sentences of its source documents.

**Table 1: The text and table accuracy comparison between different the chatbot and other LLMs**

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Chatbot | ChatGPT | Deepseek V3 |
| Text and Table | 21/21 | 21/21 | 21/21 |

### 3.1.4 Neat Answers in Kinyarwanda

Compared to other languages, queries in Kinyarwanda tended to yield more straightforward and less redundant answers. In terms of discussing only the accuracy of the results, Kinyarwanda showed the same results as English.

## 3.2 Issues

### 3.2.1 Varied Response Time in Different Language Environment

The model's response time is inconsistent. It is particularly slow when handling questions that are not direct. There is also a significant performance gap between languages, with Kinyarwanda and Chinese queries sometimes taking over a minute, whereas English is much faster. This is likely impacted by the local server environment, and a higher bandwidth server is recommended. A

Table 2. Comparison of the performance with different language input

|  |  |  |  |
| --- | --- | --- | --- |
| Language | Time | Accuracy | Issues |
| Chinese | 45s | √ | Redundant info |
| English | 28s | √ | / |
| Kinyarwanda | 55s | √ | Redundant info |

### 3.2.2 Lack of Precision in multimodality

**Low Accuracy in Image Data:** The model struggles to accurately extract information presented in graphs and charts. Some of the data in the PDF image can be recognized. Other non-vector data is basically difficult to recognize. We believe that the model does not currently involve image recognition functions.

As shown in the Table 3, the chatbot demonstrates excellent performance in text and table data extraction. While its proficiency in these areas is a key strength, performance in chart-based data extraction requires further development to achieve the same high standard. The figure accuracy of the chatbot may be caused by the accidental repetition of the figure data in the text. This has been identified as a priority area for future enhancement.

**Table 3: The figure accuracy comparison between different the chatbot and other LLMs**

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Chatbot | ChatGPT | Deepseek V3 |
| Figure | 4/16 | 6/16 | 7/16 |

**A slightly off in reference accuracy:** At times, the model cites incorrect page numbers as the source for its information, which requires correction. Seven demos are required to provide the information source. All references are correct, but there are still some issues with page positioning.

**Table 3: The referenced accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Demo | Correct PDF | Correct Page |
| References | 7 | 7/7 | 5/7 |

### 3.2.3 Unnecessary omissions and additions

**Unnecessary Rounding:** The model was observed rounding numbers when it should not (e.g., answering 4 instead of 3.9). In more than 200 questions related to data accuracy, there were four instances of omitted decimal points in the data. It is recommended to explicitly instruct the model to avoid rounding.

**Unnecessary additions under Cross-Lingual Responses:** The nature of the model's responses changes depending on the language used. The same question asked in different languages produces answers that, while similar in content, are organized differently and vary in their emphasis.

In non-English contexts for certain issues, Chat-Bot may provide excessive explanations for the causes of data, which are not explained in the data provided by NISR.

## 4 Suggestions for the future work

**Improve Multimodality and Data Extraction:** The most significant area for improvement is the chatbot's ability to interpret graphical data. The current model struggles to accurately extract information from charts and non-vector images. Future work should focus on integrating or developing a robust image recognition function to improve accuracy in this domain. Additionally, the model should be instructed to avoid unnecessary rounding of numerical data to ensure precision.

**Enhance Response Speed and Consistency:** The response time is inconsistent, particularly for Kinyarwanda and Chinese prompts. Upgrading to a higher bandwidth server is recommended to address this latency and improve overall system performance. Efforts should also be made to ensure responses are consistent and similarly structured across all supported languages.

**Refine Accuracy and Reduce Hallucinations:** The model occasionally provides incorrect page numbers for its sources and has shown instances of hallucination, especially in non-English prompts. It is crucial to correct the referencing mechanism and conduct more thorough testing to resolve version control issues that may be causing hallucinations before the official launch.