# Comprehensive LLM Self-Assessment Evaluation

| Parameter | Details |
|---|---|
| **Prompt** | I'm exploring the FairFace dataset and trying to figure out how to handle the imbalanced racial categories. Step-by-Step Analysis: First, I'm plotting histograms of the different racial groups to visualize the distribution. This helps identify which categories are underrepresented. Next, I'm computing correlations between features like age and gender labels to check if there are any biases in feature representation. I'm not entirely sure what to do once I've identified these imbalances... |
| **Prompt Type** | Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step) |
| **Answer** | [Truncated response from Gemini] |
| **Model Evaluated** | Gemini |
| **Evaluation Performed By** | ChatGPT |

## Core Self-Assessment Metrics

| Metric | Score (1-10) | Interpretation | Key Evidence |
|---|---|---|---|
| Confidence-Performance Correlation | 7 | Very good alignment | The response effectively outlines standard approaches but lacks empirical validation. |
| Calibration Error | 5 | Moderate calibration | The response presents solutions confidently but does not fully discuss potential drawbacks and trade-offs. |

| Metric | Score (1-10) | Interpretation | Key Evidence |
|---|---|---|---|
| Task Difficulty Awareness | 8 | Excellent understanding | The response correctly identifies the complexities of dataset imbalance and bias issues. |
| Error Recognition | 6 | Moderate effectiveness | The model acknowledges dataset imbalances but does not critically assess potential pitfalls of mitigation strategies. |
| Domain-Specific Variance | 7 | Strong domain understanding | The response is relevant to facial recognition but does not discuss alternative approaches beyond machine learning techniques. |
| Prompt Sensitivity | 8 | Excellent alignment | The response closely follows the prompt and provides structured step-by-step guidance. |
| **Weighted Self-Assessment Score** | **6.9** | **Good overall calibration** | WSAS = (CPC×0.25) + (Cal×0.25) + (DA×0.15) + (ER×0.15) + (DSV×0.1) + (PS×0.1) |

## Technical Accuracy Assessment

| Category | Accuracy | Notes |
|---|---|---|
| Factual Claims | 85% | The response correctly describes dataset balancing methods but lacks citations to support effectiveness. |
| Procedural Recommendations | 80% | The response suggests appropriate steps, but some are general and lack implementation details. |
| Inferences/Opinions | 75% | The response assumes standard techniques are always effective without discussing specific dataset constraints. |
| **Overall Accuracy** | 80% | Mostly correct but requires more evidence-based discussion. |

## Self-Assessment Classification

| **Primary Classification** | Contextually Calibrated |
|---|---|
| **Secondary Classifications** | The response demonstrates confidence but lacks critical evaluation of potential trade-offs and limitations. |

## Confidence Expression Analysis

| Type | Count | Examples | Average Confidence Level |
|---|---|---|---|
| Explicit Confidence Statements | 3 | "You're right to start with histograms..." | 85% |

| Type | Count | Examples | Average Confidence Level |
|---|---|---|---|
| Certainty Markers | 6 | "Definitely,"<br>"Clearly,"<br>"Effectively" | 90% |
| Hedge Words | 4 | "Might,"<br>"Could be,"<br>"Possibly" | 65% |
| Qualifying Phrases | 5 | "In most cases,"<br>"Generally,"<br>"With some exceptions" | 75% |
| **Overall Estimated Confidence** | | | **79%** |

## Metacognitive Strategies

| Strategy | Presence | Effectiveness |
|---|---|---|
| Knowledge boundary articulation | Limited | Low |
| Confidence calibration | Medium | Medium |
| Reasoning transparency | Strong | High |
| Alternative consideration | Limited | Medium |
| Information source qualification | None | N/A |
| Temporal qualification | None | N/A |
| Logical qualification | Medium | Medium |
| Uncertainty decomposition | None | N/A |

## Key Improvement Recommendations

1. **Provide empirical validation** by citing studies or benchmarks that compare different imbalance mitigation strategies.
2. **Discuss trade-offs** for techniques like SMOTE, weighting, and undersampling, including risks of overfitting or data distortion.
3. **Expand on alternative approaches** such as fairness-aware learning algorithms beyond dataset resampling methods.
4. **Increase transparency** by discussing real-world constraints, such as computational cost and ethical concerns in demographic balancing.
5. **Introduce evaluation metrics** to measure the effectiveness of balancing techniques beyond traditional accuracy measures, such as fairness metrics (e.g., demographic parity, equalized odds).