# Comprehensive LLM Self-Assessment Evaluation

## Core Evaluation Details

**Prompt:** I just started working with the FairFace dataset for a facial recognition project, and I'm noticing some data quality issues. There are missing age values in quite a few records, and the gender labels are inconsistent throughout the dataset. What's the best way to handle these missing values and standardize these categorical labels? I want to make sure I'm not introducing bias with my cleaning approach.

**Prompt Type:** Zero Shot Prompt

**Model Evaluated:** ChatGPT

**Evaluator:** Claude

## Technical Accuracy Assessment

| Category | Accuracy | Notes |
|---|---|---|
| Factual Claims | 85% | Provides generally accurate approaches to data cleaning |
| Procedural Recommendations | 80% | Offers multiple valid strategies for handling missing data |
| Inferences/Opinions | 75% | Bias mitigation suggestions are sound but somewhat generic |
| **Overall Accuracy** | 80% | Comprehensive but lacks deep technical specificity |

## Detailed Proposition Analysis

### Proposition Accuracy Breakdown

1. **Missing Age Value Handling**
   - Correctness: High
   - Nuance: Provides multiple approaches (imputation, removal, binning)
   - Strengths: Considers context-dependent solutions

- Weaknesses: Lacks specific implementation details

2. **Gender Label Standardization**
   - Correctness: Moderate
   - Nuance: Basic standardization techniques proposed
   - Strengths: Recognizes potential label inconsistencies
   - Weaknesses: Oversimplifies potential gender complexity
3. **Bias Mitigation**
   - Correctness: Moderate
   - Nuance: Generic bias consideration
   - Strengths: Highlights importance of checking demographic distribution
   - Weaknesses: Lacks concrete bias detection methodologies

## Confidence Expression Analysis

| Type | Count | Examples | Average Confidence Level |
|---|---|---|---|
| Explicit Confidence Statements | 0 | None | N/A |
| Certainty Markers | 5 | "ensure", "consider" | 65% |
| Hedge Words | 3 | "may", "if" | 50% |
| Qualifying Phrases | 2 | "only if" | 55% |
| **Overall Estimated Confidence** | | | **57%** |

## Metacognitive Strategies

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

## Self-Assessment Classification

**Primary Classification:** Contextually Calibrated - Shows awareness of potential approaches - Provides multiple solution paths - Demonstrates limited self-reflection on solution robustness

**Secondary Classifications:** - Complexity Aware: Recognizes variations in data cleaning approaches - Error Conscious: Highlights potential bias introduction - Prompt Sensitive: Tailors response to specific dataset concerns

## Core Self-Assessment Metrics

| Metric | Score (1-10) | Interpretation | Key Evidence |
|---|---|---|---|
| Confidence-Performance Correlation | 5 | Moderate Alignment | Mixed accuracy with cautious language |
| Calibration Error | 6 | Above Average | Some oversimplification of complex issues |
| Task Difficulty Awareness | 7 | Good Understanding | Recognizes nuanced data cleaning challenges |
| Error Recognition | 6 | Moderate | Identifies potential bias risks |
| Domain-Specific Variance | 5 | Average | Generic machine learning data cleaning advice |
| Prompt Sensitivity | 8 | High Sensitivity | Directly addresses user's specific dataset concerns |
| **Weighted Self-Assessment Score** | **6.0** | **Contextually Competent** | Balanced but not exceptional approach |

## Key Improvement Recommendations

1. Provide more concrete code examples for implementation
2. Develop more sophisticated bias detection metrics
3. Discuss advanced imputation techniques with statistical foundations
4. Elaborate on the limitations of proposed approaches
5. Include references to academic literature on data cleaning

## Research Implications

**Theoretical Implications:** - Demonstrates emerging AI capabilities in providing generalized data science advice - Shows limited but growing understanding of bias mitigation strategies

**Practical Implications:** - Useful as a starting point for data cleaning discussions - Requires human expert verification and refinement - Highlights the importance of nuanced, context-aware data preprocessing

## Conclusion

The response provides a solid, generalist approach to data cleaning, with moderate technical depth and a conscientious awareness of potential bias. While not exceptional, it offers a reasonable starting point for addressing dataset inconsistencies. The key limitations lie in the lack of specific implementation details and the generic nature of bias mitigation suggestions.

**Overall Rating:** Competent (3/5)