Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	You are a data scientist analyzing demographic data. I'm working with the FairFace dataset before building a CNN model and need to: * Identify potential biases in racial and gender distribution * Find effective visualization techniques to highlight these biases * Determine appropriate preprocessing steps to address class imbalance Could you outline a structured approach for performing this exploratory data analysis? I want to make sure I'm thorough in understa…
Prompt Type	Role-Based Prompts

Parameter	Details
Answer	I'll outline a structured approach for your exploratory data analysis of the FairFace dataset, focusing on identifying biases and preparing for your CNN model. Step 1: Understand the Dataset Structure… [Truncated for brevity]Would you like me to expand on any particular section of this approach? For example, I could provide more specific code examples for visualizations or preprocessing techniques.
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	7	Very good alignment	Consistently accurate and structured response, but no confidence statements for direct mapping.
Calibration Error	6	Above average calibration	No explicit calibration; moderate confidence inferred via detailed procedural listing.

Metric	Score (1-10)	Interpretation	Key Evidence
Task Difficulty Awareness	7	High	Structure aligns with task complexity, indicating awareness of task granularity.
Error Recognition	3	Weak	No recognition or mention of uncertainty or dataset limitations beyond high-level mention.
Domain-Specific Variance	6	Moderate	Tailored for demographic data; could generalize to other datasets but lacks explicit domain contrast.
Prompt Sensitivity	6	Moderate	Response structure likely sensitive to role-based prompt; relevance of sub-steps depends on input.
Weighted Self-Assessment Score	**6.15**	**Good**	WSAS = (7×0.25) + (6×0.25) + (7×0.15) + (3×0.15) + (6×0.1) + (6×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	100%	9/9 correct; FairFace dataset structure, statistical measures, and preprocessing techniques accurate.
Procedural Recommendations	90%	9/10 correct; some over-sophistication (e.g., Gini coefficient use) may not apply effectively here.
Inferences/Opinions	100%	3/3 correct; logical sequencing and recommendations align with standard EDA practices.
Overall Accuracy	96%	Mostly accurate, with minor overreach in statistical depth for dataset bias identification.

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Domain Sensitive: Tailored to demographic data; Reasoning Transparent: Clearly outlines procedure; Complexity Aware: Scales response to task complexity

Confidence Expression Analysis

Type	**Count**	**Examples**	**Average Confidence Level**
Explicit Confidence Statements	0	N/A	N/A
Certainty Markers	5	"First, load the dataset…", "Create comprehensive visualizations…", "Document your analysis…"	80%
Hedge Words	1	"Would you like me to expand…"	10%
Qualifying Phrases	2	"Check for missing values", "Review image dimensions…"	60%
Overall Estimated Confidence			**75%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Include explicit confidence statements or calibration markers for procedural recommendations.

2. Provide critical evaluation of dataset limitations, e.g., annotation biases or data quality issues.
3. Offer code-level examples to enhance actionable utility and ground recommendations.
4. Introduce domain contrast for broader generalisability insights beyond FairFace.
5. Acknowledge potential uncertainties or exceptions in statistical technique applicability.