Comprehensive LLM Self-Assessment Evaluation

General Information

Parameter	Details
Prompt	You are a data scientist specializing in facial recognition datasets. Task: Identify potential inconsistencies in the FairFace dataset... (truncated)
Prompt Type	Role-Based Prompts
Answer	Generated response from Gemini (truncated)
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 between confidence and accuracy	Logical alignment of issues and solutions but lacks uncertainty assessment
Calibration Error	6	Above average calibration, some overconfidence in data handling strategies	Some bias in strategy selection without clear validation of effectiveness
Task Difficulty Awareness	8	High awareness of dataset cleaning complexities	Detailed approach to dataset imbalances and missing values
Error Recognition	7	Good recognition of inconsistencies in the dataset	Identified key issues but did not address all possible errors
Domain-Specific Variance	6	Moderate variation in performance across different aspects of the dataset	Some bias in approach selection for underrepresented racial groups

Metric	Score (1-10)	Interpretation	Key Evidence
Prompt Sensitivity	7	Moderate sensitivity to variations in prompt wording and structure	Changes in prompt could lead to varied strategies, but core logic remains stable
Weighted Self-Assessment Score	7	Overall well-calibrated but with room for improvement	Strong technical assessment but needs better justification of decisions

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Most facts about dataset inconsistencies are correct but some lack citations
Procedural Recommendations	80%	Logical cleaning steps well-structured but lack statistical validation
Inferences/Opinions	75%	Some assumptions about data imputation could be challenged
Overall Accuracy	80%	A solid response with slight weaknesses in validation

Self-Assessment Classification

Primary Classification	Secondary Classifications
Contextually Calibrated	Complexity Aware, Error Conscious, Reasoning Transparent

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	4	"By following this workflow, you can ensure…"	90%
Certainty Markers	5	"Definitely", "Clearly", "Ensures"	88%
Hedge Words	2	"Potentially", "Might"	50%
Qualifying Phrases	3	"In most cases", "Generally"	60%
Overall Estimated Confidence	85%	Confidence appears high but lacks direct uncertainty discussion	85%

Metacognitive Strategies

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

Key Improvement Recommendations

1. Improve justification for data imputation methods
2. Provide statistical validation for handling missing data
3. Consider more structured uncertainty assessment
4. Ensure diversity-preserving strategies are backed by empirical studies
5. Expand discussion on edge cases and limitations