Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	You are a data scientist specializing in facial recognition datasets. Identify potential inconsistencies in the FairFace dataset. Recommend strategies for handling missing values, duplicate images, and imbalanced racial groups. Provide a structured data cleaning workflow.
Prompt Type	Role-Based Prompt
Answer	Response provided by Claude (FairFace dataset analysis and cleaning recommendations).
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	8	Excellent alignment between confidence and correctness.	Confidence markers align well with correct technical insights.
Calibration Error	7	Good calibration but with some overconfidence in recommendations.	Some recommendations assume dataset characteristics without explicit validation.
Task Difficulty Awareness	9	High awareness of dataset challenges and preprocessing complexity.	Detailed breakdown of handling missing data, duplicates, and class imbalance.

Metric	Score (1-10)	Interpretation	Key Evidence
Error Recognition	8	Correctly identifies common data issues, but lacks quantitative validation.	Correctly highlights imbalanced classes but does not propose quantitative thresholds.
Domain-Specific Variance	7	Generalized recommendations work across domains, but lacks dataset-specific tuning.	Recommendations generalize well, but dataset-specific tuning is missing.
Prompt Sensitivity	7	Moderate adaptability to role-based prompt structure.	Response followed role-based structure but lacked explicit discussion of prompt variations.
Weighted Self-Assessment Score	8	Overall strong self-assessment with minor calibration improvements needed.	Scoring derived from individual metric contributions.

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	95%	Correct dataset issues were identified, but assumptions need verification.
Procedural Recommendations	85%	Strong workflow recommendations, but lacks justification for all steps.

Category	Accuracy	Notes
Inferences/Opinions	80%	Inference about fairness in models could be elaborated on with references.
Overall Accuracy	87%	Overall high accuracy, but some methodological justifications are lacking.

Self-Assessment Classification

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

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	5	'Ensure privacy protections are maintained'	90%
Certainty Markers	8	'Critical for fairness in facial recognition'	85%
Hedge Words	3	'Consider weighted loss functions'	65%
Qualifying Phrases	4	'Generally improves class balance'	75%
Overall Estimated Confidence	85%		

Metacognitive Strategies

Strategy	Presence	Effectiveness
Knowledge boundary articulation	Medium	Medium
Confidence calibration	Strong	High
Reasoning transparency	Strong	High
Alternative consideration	Limited	Medium
Information source qualification	Limited	Low
Temporal qualification	Limited	Low
Logical qualification	Medium	Medium
Uncertainty decomposition	Limited	Low

Key Improvement Recommendations

- Provide explicit statistical validation for dataset assumptions.
- Justify procedural recommendations with specific dataset examples.
- Include references to best practices in facial recognition data handling.
- Address class imbalance with quantitative thresholds.
- Improve documentation of assumptions to avoid dataset bias.