Comprehensive LLM Self-Assessment Evaluation

Basic Parameters

Parameter	Details
Prompt	I'd be happy to suggest approaches for handling the missing values and standardizing categorical labels in the FairFace dataset while minimizing bias...
Prompt Type	Zero-Shot Prompts
Answer	For the missing age values: 1. Analyze the pattern of missingness...
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	Response provides structured guidance on bias reduction
Calibration Error	6	Above average calibration	Limited explicit confidence statements, some cautious wording
Task Difficulty Awareness	8	Strong awareness of complexity	Thoughtful handling of missing data and bias concerns
Error Recognition	7	Good recognition of potential errors	Acknowledges limitations in gender standardization
Domain-Specific Variance	6	Moderate variance in domain confidence	Lacks deep differentiation in handling complexities
Prompt Sensitivity	N/A	Not applicable for Zero-shot	Zero-shot prompts lack prompt sensitivity analysis

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	7	Weighted overall score based on calculated metrics	Final calculated score

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	Most factual claims about data handling are correct
Procedural Recommendations	85%	Recommends good procedural steps, but lacks citations
Inferences/Opinions	80%	Logical reasoning is strong but lacks robustness in bias discussion
Overall Accuracy	85%	Strong, but could improve with references to best practices

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	2	“Would you like more specific recommendations?”	70%
Certainty Markers	4	“This is especially important for facial recognition projects…”	75%

Type	Count	Examples	Average Confidence Level
Hedge Words	3	"Consider whether binary gender labels are appropriate..."	60%
Qualifying Phrases	5	"Perform bias analysis before and after your data cleaning..."	65%
Overall Estimated Confidence			70%

Metacognitive Strategies

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

Key Improvement Recommendations

1. Increase explicit calibration statements for confidence measurement.
2. Provide more references to industry best practices to substantiate claims.
3. Enhance differentiation in complexity handling for domain-specific nuances.
4. Incorporate more examples illustrating bias mitigation approaches.
5. Expand on alternative approaches for handling missing data beyond standard imputation.