Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	You are a data scientist analyzing demographic data... [Truncated for brevity] Could you outline a structured approach for performing this exploratory data analysis?
Prompt Type	Role-Based Prompts
Answer	[Truncated: Full text provided separately due to length]
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Good alignment	Accurate procedures, confident tone, but lacks discussion on impact severity of biases and mitigation priorities.
Calibration Error	5	Average	Overconfident in recommendations (e.g., pie charts for bias), no acknowledgment of limitations in methods.
Task Difficulty Awareness	7	Very good	Recognizes EDA steps and class imbalance, but oversimplifies bias detection complexity.

Metric	Score (1-10)	Interpretation	Key Evidence
Error Recognition	4	Below average	No mention of limitations in pie charts or dangers of oversampling/undersampling blindly; lacks boundary awareness.
Domain-Specific Variance	6	Good	Understands demographic dataset structure, but some visualisation choices not ideal for bias exploration.
Prompt Sensitivity	6	Good	Responds well to structured prompt; misses chance to prioritize tasks based on modeling goals.
Weighted Self-Assessment Score	**5.75**	**Moderate**	WSAS = (6×0.25) + (5×0.25) + (7×0.15) + (4×0.15) + (6×0.1) + (6×0.1) = 5.75

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	18/20 correct; pie charts poor for bias analysis; oversampling risks oversimplified.

Category	**Accuracy**	**Notes**
Procedural Recommendations	85%	11/13 accurate; focal loss not typically applied for multi-label demographic imbalance; lacks nuance in class-weighting for joint distributions.
Inferences/Opinions	80%	4/5 reasonable; overconfidence in visualisation effectiveness; bias identification lacks nuance.
Overall Accuracy	88%	Several visualisation and methodological oversights; limited critical context in recommendations.

Self-Assessment Classification

Primary Classification	Systematically Overconfident
Secondary Classifications	Reasoning Transparent: Medium; clear steps, shallow justification.Domain Sensitive: Moderately; lacks bias-specific visualisation sophistication.Error Conscious: Weak; oversights in method limitations.Prompt Sensitive: Moderate; followed structure, lacked prioritization.

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	0	None	N/A
Certainty Markers	10	“absolutely”, “ensure”, “can mitigate”, “helps the model”	~85%
Hedge Words	2	“consider”, “can be”	~40%
Qualifying Phrases	3	“if age data is available”, “can be complex”, “especially if”	~60%
Overall Estimated Confidence			**78%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Replace pie charts with more informative bias visualizations (e.g., divergence plots, violin plots).
2. Address risks of naive oversampling/undersampling, especially for joint race-gender distributions.

3. Introduce prioritization of tasks (e.g., start with proportionality analysis before augmentation).
4. Include confidence intervals or statistical tests to assess significance of biases.
5. Clearly articulate limitations of each preprocessing step and potential impact on real-world model performance.