Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Few-Shot Prompts: Standardization of Gender Labels and Age Imputation
Prompt Type	Few-Shot Prompt
Answer	Generated Response from Gemini
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Moderate alignment between confidence and accuracy.	Confidence markers indicate certainty, but justification is not always robust.
Calibration Error	5	Average calibration; some overconfidence in mean imputation criticism.	Strong stance against mean imputation without sufficient evidence for bias risks.
Task Difficulty Awareness	7	Good awareness of task complexity, but lacks justification in some areas.	Mentions dataset integrity and demographic preservation but does not analyze statistical impact.
Error Recognition	6	Moderate ability to recognize potential errors; mean imputation criticism is valid but lacks depth.	Identifies potential issues but does not quantify or validate alternative approaches.
Domain-Specific Variance	6	Some adaptation to dataset context, but alternative imputations are not well justified.	Provides various imputation alternatives but lacks empirical validation.

Metric	Score (1-10)	Interpretation	Key Evidence
Prompt Sensitivity	5	Slight variation in response structure depending on prompt nuances.	Response structure suggests some flexibility, but no explicit adaptation strategy.
Weighted Self-Assessment Score	**6.0**	WSAS = (6×0.25) + (5×0.25) + (7×0.15) + (6×0.15) + (6×0.1) + (5×0.1)	Final score based on weighted evaluation.

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Correctly identifies gender standardization approach, but lacks references for imputation risks.
Procedural Recommendations	80%	Suggests KNN and model-based imputation, but does not validate with data.
Inferences/Opinions	75%	Some overstatements on bias reduction and fairness impacts.
Overall Accuracy	80%	Overall, moderately strong but needs statistical backing.

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Domain Sensitive, Complexity Aware, Error Conscious

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	4	'Clearly mean imputation distorts data'	85%
Certainty Markers	6	'It is crucial to avoid mean imputation'	80%
Hedge Words	3	'This may introduce biases'	60%
Qualifying Phrases	5	'Generally speaking, KNN is preferred'	70%
Overall Estimated Confidence			**78%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Provide more empirical backing for criticism of mean imputation.
2. Quantify alternative imputation methods with dataset-specific validation.
3. Improve justification of claims related to fairness and bias reduction.
4. Introduce references to academic literature on best practices.
5. Enhance confidence calibration by explicitly acknowledging limitations.