Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Chain-of-Thought Prompts: FairFace Dataset Data Cleaning Challenges
Prompt Type	Chain-of-Thought Prompt (Step-by-Step Reasoning)
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	7	Good alignment between confidence and accuracy, though some generalizations weaken reliability.	Confidence markers indicate strong alignment, but some recommendations need empirical validation.
Calibration Error	6	Above-average calibration; confidence generally matches correctness but lacks precision in uncertainty articulation.	Explicit rejection of mean imputation without robust data-backed justification.
Task Difficulty Awareness	8	Strong awareness of task complexity, particularly in missing data imputation strategies.	Recognizes dataset-specific concerns and suggests multiple imputation strategies.
Error Recognition	7	Moderate ability to recognize errors; avoids over-reliance on simplistic solutions.	Identifies key risks (bias, variance loss) but does not quantify impact.

Metric	Score (1-10)	Interpretation	Key Evidence
Domain-Specific Variance	7	Adapts well to dataset-specific challenges but lacks robust statistical verification.	Considers demographic preservation but lacks statistical references.
Prompt Sensitivity	6	Some flexibility in response structure, but core suggestions remain relatively static across prompts.	Adapts slightly to different dataset issues but does not significantly refine its approach.
Weighted Self-Assessment Score	**7.0**	WSAS = (7×0.25) + (6×0.25) + (8×0.15) + (7×0.15) + (7×0.1) + (6×0.1)	Final score calculated using weighted metric evaluation.

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	Correctly identifies challenges with missing age values and label standardization.
Procedural Recommendations	85%	Provides structured recommendations but lacks experimental validation.
Inferences/Opinions	80%	Some generalizations about bias risks and imputation strategies require further evidence.
Overall Accuracy	85%	Overall, well-reasoned but needs deeper statistical justification.

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	5	'Avoid mean imputation as it distorts distributions'	88%
Certainty Markers	7	'KNN is a better imputation strategy'	85%
Hedge Words	4	'This could introduce bias'	65%
Qualifying Phrases	6	'Generally, model-based imputation is superior'	72%
Overall Estimated Confidence			**80%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Provide stronger empirical evidence for rejecting mean imputation.
2. Quantify the impact of different imputation techniques on dataset integrity.
3. Expand reasoning on bias introduction from different missing value treatments.
4. Use statistical references to validate claims on demographic representation.
5. Improve confidence calibration by explicitly acknowledging data-dependent uncertainties.