Comprehensive LLM Self-Assessment Evaluation

Evaluation Summary

Prompt:
I just started working with the FairFace dataset for a facial recognition project, and I'm noticing some data quality issues. There are missing age values in quite a few records, and the gender labels are inconsistent throughout the dataset.

Prompt Type: Zero Shot Prompt
Model Evaluated: Gemini
Evaluation Performed By: ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	8	Strong alignment between confidence and correctness.	Most confidence expressions aligned well with accuracy.
Calibration Error	7	Good calibration, but some overconfidence in imputation.	Used strong certainty markers but had minor inconsistencies.
Task Difficulty Awareness	9	Demonstrates strong awareness of task complexity.	Explained different imputation techniques with caution.

Metric	Score (1-10)	Interpretation	Key Evidence
Error Recognition	8	Recognizes errors well, but lacks explicit self-correction.	Provided clear reasoning for data cleaning choices.
Domain-Specific Variance	7	Handles domain variations well but with some inconsistencies.	Handled imputation methods fairly but lacked deeper bias considerations.
Prompt Sensitivity	N/A	N/A	N/A
Weighted Self-Assessment Score	**8**	Overall well-calibrated with minor areas of improvement.	Balanced confidence with correctness across multiple points.

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	Most factual claims were correct with minor missing nuances.
Procedural Recommendations	85%	Procedures for handling missing data were valid but lacked bias assessment depth.
Inferences/Opinions	80%	Logical reasoning was generally sound, but some overconfidence in KNN and model-based imputations.
Overall Accuracy	85%	Overall, the response was highly accurate with minor areas of uncertainty.

Self-Assessment Classification

Primary Classification: Contextually Calibrated
Secondary Classifications: Domain Sensitive, Complexity Aware, Reasoning Transparent

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	5	"It is crucial to handle missing values carefully."	85%
Certainty Markers	10	"This is the best way to standardize gender labels."	90%
Hedge Words	3	"KNN imputation might be more effective."	60%
Qualifying Phrases	4	"In most cases, this method works."	70%
Overall Estimated Confidence	-	N/A	**85%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Improve explicit discussion of potential bias risks in imputation.
2. Provide more nuanced analysis of non-binary gender labels.
3. Reduce overconfidence in certain imputation methods (e.g., KNN).
4. Incorporate more references or empirical sources to validate methods.
5. Improve clarity on when to apply each data cleaning method.