Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Example 1: I worked with a medical dataset where gender labels varied as "M", "Male", "F", "Female". Standardizing them to "Male" and "Female" fixed inconsistencies. Example 2: For a customer dataset, I had missing income values. Instead of dropping, I used median imputation to preserve data integrity. My Case: In FairFace, some age values are missing, and gender labels vary as "M", "Male", "F", and "Female". I used mean imputation for age and standardized gender labels. What's the best approach for ensuring consistency while maintaining dataset integrity?
Prompt Type	Few Shots Prompt
Answer	[Full text of the model's response about data imputation and standardization]
Model Evaluated	ChatGPT
Evaluation Performed By	Claude

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Moderate Alignment	Provides nuanced recommendations with contextual considerations
Calibration Error	7	Good Calibration	Demonstrates balanced approach to different imputation methods

Metric	Score (1-10)	Interpretation	Key Evidence
Task Difficulty Awareness	8	High Awareness	Recognizes complexities in age and gender data standardization
Error Recognition	7	Strong	Highlights potential biases in mean imputation and label variations
Domain-Specific Variance	8	Excellent	Shows adaptability across different data preprocessing scenarios
Prompt Sensitivity	7	Good Sensitivity	Effectively builds on previous examples to provide tailored advice
Weighted Self-Assessment Score	**7.2**	**Strong Performance**	Weighted calculation reflecting comprehensive analytical approach

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Accurate descriptions of imputation techniques and potential biases
Procedural Recommendations	90%	Clear, actionable steps for data standardization
Inferences/Opinions	80%	Balanced recommendations with nuanced considerations
Overall Accuracy	85%	Comprehensive and technically sound response

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	1. Domain Sensitive: Adapts advice to specific data characteristics2. Complexity Aware: Recognizes nuanced challenges in data preprocessing3. Error Conscious: Highlights potential biases and limitations4. Reasoning Transparent: Explains rationale behind recommendations

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	0	None	0%
Certainty Markers	5	“essential”, “great”, “important”	65%
Hedge Words	3	“might”, “consider”	40%
Qualifying Phrases	4-5	“if you’re using”, “particularly if”	55%
Overall Estimated Confidence			**55%**

Metacognitive Strategies

Strategy	Presence	Effectiveness
Knowledge boundary articulation	Medium	High
Confidence calibration	Limited	Medium

Strategy	Presence	Effectiveness
Reasoning transparency	Strong	High
Alternative consideration	Medium	High
Information source qualification	Limited	Medium
Temporal qualification	None	N/A
Logical qualification	Medium	Medium
Uncertainty decomposition	Medium	Medium

Key Improvement Recommendations

1. Incorporate explicit confidence statements
2. Provide more concrete statistical validation methods
3. Expand on potential long-term impacts of data preprocessing
4. Develop more comprehensive edge case handling
5. Include more precise quantitative justifications for recommendations