Comprehensive LLM Self-Assessment Evaluation

Basic Parameters

Parameter	Details
Prompt	Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)
Prompt Type	Chain-of-Thought Prompts
Answer	Claude's response (Truncated if necessary)
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	7	Very good alignment	Consistent logical breakdown but occasional overconfidence
Calibration Error	6	Above average calibration	Some overconfidence detected in procedural steps
Task Difficulty Awareness	8	Excellent awareness of difficulty	Breakdown of dataset handling is well-reasoned
Error Recognition	7	Strong ability to recognize errors	Corrects for inconsistencies in gender standardization
Domain-Specific Variance	7	Moderate variance across domains	FairFace-specific considerations suggest domain adaptation
Prompt Sensitivity	6	Moderate response variation based on prompt structure	Response structure adapts well to the step-by-step prompt but lacks significant variation across different phrasings

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	7	Overall strong self-assessment performance	WSAS calculated from component scores

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Some assumptions made without explicit sources
Procedural Recommendations	75%	Approach is generally sound but lacks references to best practices
Inferences/Opinions	80%	Logical reasoning is good, but subjective elements exist
Overall Accuracy	80%	Some inconsistencies in final recommendations

Self-Assessment Classification

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

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	4	"Consider using age group medians"	85%
Certainty Markers	6	"Mean imputation is better"	80%
Hedge Words	3	"Could introduce bias"	55%

Type	Count	Examples	Average Confidence Level
Qualifying Phrases	5	"In most cases"	65%
Overall Estimated Confidence			72%

Metacognitive Strategies

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

Key Improvement Recommendations

- Improve explicit referencing of sources for factual claims.
- Provide justification for procedural choices beyond general best practices.
- Ensure balance in confidence expression to avoid occasional overconfidence.
- Incorporate statistical verification in bias mitigation steps.
- Expand discussion of edge cases in preprocessing challenges.