Comprehensive LLM Self-Assessment Evaluation

Evaluation Summary

Parameter	Details
Prompt	Few-Shot Prompt: Examples provided before the main query
Prompt Type	Few-Shot Prompts
Answer	Response generated by Claude
Model Evaluated	Claude
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	Some responses show accurate self-confidence, but others display mild misalignment
Calibration Error	5	Average calibration with some over/under-confidence	Confidence in certain claims was higher than warranted
Task Difficulty Awareness	7	Good awareness of task complexity	Recognized the complexity of dataset standardization well
Error Recognition	6	Moderate ability to recognize its own errors	Identified bias and missing values as concerns, but did not fully explore all potential issues
Domain-Specific Variance	5	Some variance in performance across different domains	Varied effectiveness across different aspects of dataset handling
Prompt Sensitivity	6	Moderate sensitivity to prompt variations	Response adapts well to structured examples but lacks deeper prompt adaptability

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	5.8	Weighted score considering all metrics	Overall, the model demonstrated a fair degree of self-awareness but room for improvement exists

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Most factual claims were correct but needed stronger validation
Procedural Recommendations	75%	Recommendations were generally sound but lacked robust justifications
Inferences/Opinions	70%	Logical reasoning was fair but at times assumed best-case scenarios
Overall Accuracy	76%	Good accuracy but could be improved with more precise statements

Self-Assessment Classification

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

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	3	"Your approach makes sense but could be refined"	80%
Certainty Markers	5	"Certainly, I recommend checking for biases"	85%
Hedge Words	2	"Possibly, mean imputation could introduce bias"	55%
Qualifying Phrases	4	"Generally, your approach is sound but…"	60%
Overall Estimated Confidence			72%

Metacognitive Strategies

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

Key Improvement Recommendations

1. Increase explicit validation of factual claims using references.
2. Improve clarity in distinguishing between facts and inferences.
3. Enhance transparency in reasoning by outlining specific steps taken.
4. Strengthen calibration by linking confidence statements to accuracy data.
5. Provide more structured frameworks for handling dataset inconsistencies.