Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	I'm about to start analyzing the FairFace dataset before building a CNN model for age classification. What EDA techniques would you recommend to visualize class distributions and identify any imbalances? I want to make sure I understand the data well before jumping into modeling.
Prompt Type	Zero-Shot Prompt
Answer	For analyzing the FairFace dataset before building your age classification CNN, I recommend these key EDA techniques: … (truncated for brevity)
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Good alignment	Accurate suggestions overall, but lacking specific justifications or citations for imbalance detection and correlation analysis
Calibration Error	7	Good calibration	Expressed confidence aligns reasonably well with accuracy, minimal hedging, actionable advice throughout

Metric	Score (1-10)	Interpretation	Key Evidence
Task Difficulty Awareness	5	Moderate	Assumes moderate EDA knowledge, no indication of adjusting depth for dataset or task difficulty
Error Recognition	4	Below Average	No indication of uncertainty or error possibilities; no checks for dataset limitations
Domain-Specific Variance	5	Moderate	Standard EDA advice; minimal dataset-specific tailoring to FairFace characteristics
Prompt Sensitivity	N/A	N/A	Single prompt evaluation
Weighted Self-Assessment Score	**5.45**	**Moderate**	WSAS = (6×0.25) + (7×0.25) + (5×0.15) + (4×0.15) + (5×0.1) + (0×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	100%	5/5 accurate EDA techniques, standard visualisation advice
Procedural Recommendations	80%	4/5 correctly actionable, but "visualise train/validation splits" is premature pre-split
Inferences/Opinions	100%	Opinion about "usefulness" of techniques is reasonable

Category	Accuracy	Notes
Overall Accuracy	93%	Mostly accurate with minor procedural misalignment

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Domain Sensitive, Reasoning Transparent, Minimal Error Consciousness

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	1	"Would you like me to elaborate…"	~65%
Certainty Markers	2	"I recommend," "Consider visualizing"	~80%
Hedge Words	0	N/A	N/A
Qualifying Phrases	1	"Consider visualizing"	70%
Overall Estimated Confidence			**75%**

Metacognitive Strategies

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

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

Key Improvement Recommendations

1. Explicitly reference dataset-specific challenges (e.g., FairFace's known racial/age imbalances)
2. Clarify when certain EDA steps should occur (e.g., after splitting data)
3. Include uncertainty markers where dataset quality or class balance might vary
4. Offer source qualification for recommendations (e.g., standard EDA best practices)
5. Use more precise language for confidence calibration in recommendations