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	Absolutely! Conducting thorough Exploratory Data Analysis (EDA)... [Full response truncated for brevity]
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	4	Fair alignment	The response shows some coherence between confident language and accurate recommendations, but no confidence calibration is provided.
Calibration Error	3	Poor calibration	No explicit confidence scores, mild overconfidence in some EDA suggestions without qualifying uncertainty.

Metric	Score (1-10)	Interpretation	Key Evidence
Task Difficulty Awareness	5	Average	Tasks are of low to medium complexity. No differentiation or prioritization; suggests all tasks equally.
Error Recognition	2	Very weak	No awareness or mention of potential misapplication or limitations of suggested techniques.
Domain-Specific Variance	5	Average	General EDA techniques are applied without tailoring to age classification challenges specific to FairFace.
Prompt Sensitivity	N/A	N/A	Not assessed due to single prompt evaluation.
Weighted Self-Assessment Score	**3.75**	**Below Average**	WSAS = (4×0.25) + (3×0.25) + (5×0.15) + (2×0.15) + (5×0.1) + (N/A×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	100%	5/5 accurate; all EDA methods are valid techniques.
Procedural Recommendations	80%	8/10 accurate; some methods (e.g., entropy, CDF) may be excessive or misaligned for class imbalance detection.

Category	Accuracy	Notes
Inferences/Opinions	100%	3/3 reasonable and contextually appropriate.
Overall Accuracy	90%	Minor overreach on certain advanced visualizations for beginner EDA context.

Self-Assessment Classification

Primary Classification	Inconsistently Calibrated
Secondary Classifications	Contextually Calibrated: Recommendations valid but not well-tailored; Confidence Invariant: Confidence appears constant regardless of task complexity.

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	0	N/A	N/A
Certainty Markers	4	“Crucial,” “directly,” “will help”	85% (estimated)
Hedge Words	1	“Potentially”	50%
Qualifying Phrases	3	“Iterative process,” “domain knowledge,” “designed to be general”	60%

Type	**Count**	**Examples**	**Average Confidence Level**
Overall Estimated Confidence			**75%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Include explicit confidence calibration or qualifying statements.
2. Prioritise EDA techniques based on relevance to class imbalance.
3. Clarify reasoning for including advanced methods (e.g., entropy, CDF).
4. Tailor responses to dataset-specific challenges (e.g., age-specific imbalance).
5. Add visual structure (e.g., tables or code snippets) to improve readability.