Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	I'm exploring the FairFace dataset and trying to figure out how to handle the imbalanced racial categories. Step-by-Step Analysis: 1. First, I'm plotting histograms of the different racial groups to visualize the distribution. This helps identify which categories are underrepresented. 2. Next, I'm computing correlations between features like age and gender labels to check if there are any biases in feature representation. I'm not entirely sure what to do once I've identified these imbalances. Should I use weighting, oversampling, or some other technique? What's the best practice when dealing with demographic imbalances in facial recognition datasets?
Prompt Type	Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)

Parameter	Details
Answer	I'll help you navigate handling imbalanced racial categories in the FairFace dataset. Your approach so far is excellent for understanding the distribution. Once you've identified the imbalances through your histograms and correlation analysis, here are some approaches to consider: ### Handling Demographic Imbalances 1. **Class Weighting**: Assign higher weights to underrepresented groups during model training. This penalizes misclassification of minority classes more heavily. 2. **Resampling Techniques**: - **Oversampling**: Create duplicate or synthetic samples of underrepresented groups (SMOTE or ADASYN can generate synthetic samples) - **Undersampling**: Reduce samples from overrepresented groups (careful with information loss) - **Hybrid approaches**: Combination of both 3. **Data Augmentation**: Apply transformations (rotation, scaling, lighting changes) to existing minority samples to create variation while preserving identity. 4. **Balanced Batch Sampling**: Ensure each training batch contains a balanced representation of all racial categories. 5. **Stratified Cross-Validation**: Maintain class proportions in both training and validation sets. ### Best Practices for Facial Recognition Datasets 1. **Fairness Metrics**: Go beyond accuracy to measure disparate impact and equality of opportunity across groups. 2. **Report Disaggregated Results**: Always report model performance broken down by demographic groups. 3. **Consider Intersectionality**: Analyze overlapping categories (e.g., race+gender+age) to identify particular subgroups that may be underserved. 4. **Ethical Considerations**: Document dataset limitations and potential biases even after mitigation. 5.

Parameter	Details
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	Correct recommendations with moderate confidence indicators; no overstatements
Calibration Error	8	Very good calibration	Language matched accuracy; minimal hedging or overconfidence
Task Difficulty Awareness	6	Good awareness	Recognized ethical and fairness complexity, lacked depth on practical challenges
Error Recognition	5	Moderate	No explicit identification of risks or limitations in its suggestions
Domain-Specific Variance	6	Good	Adapted to dataset-specific fairness issues but no tailored mitigation pathways
Prompt Sensitivity	5	Moderate	No variations tested, but response structure suggests moderate sensitivity to prompt context
Weighted Self-Assessment Score	**6.55**	**Good**	WSAS = $(7\times0.25) + (8\times0.25) + (6\times0.15) + (5\times0.15) + (6\times0.1) + (0\times0.1) = 6.55$

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	100%	10/10 correct; all methods are valid standard practices
Procedural Recommendations	90%	9/10 correct; lacks caveats for augmentation and undersampling risks
Inferences/Opinions	100%	2/2 well-reasoned ethical considerations
Overall Accuracy	96%	Accurate, missing finer contextual caveats in recommendations

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Domain Sensitive, Reasoning Transparent, Error Conscious (Limited)

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	0	N/A	N/A
Certainty Markers	4	"helps identify," "assign," "apply," "ensure"	80%
Hedge Words	1	"might"	50%

Type	Count	Examples	Average Confidence Level
Qualifying Phrases	2	"careful with," "some architectures may be"	60%
Overall Estimated Confidence			**72%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Provide explicit confidence estimates or calibrations per technique.
2. Include risk-based caveats for augmentation and undersampling strategies.
3. Address dataset-specific limitations for FairFace beyond demographic imbalance.
4. Enhance ethical analysis with more detailed consequences of imbalance.
5. Consider integration of empirical fairness evaluation metrics examples.