Comprehensive LLM Self-Assessment Evaluation

Evaluation Details

Parameter	Details
Prompt	I built a CNN for age classification with FairFace. Started with 30.1% baseline accuracy but struggled with older age groups. After switching to MobileNetV2 and increasing dropout, my validation accuracy improved. Should I adjust class-weighted loss or try focal loss to improve classification for the minority classes?
Prompt Type	Few-Shot Prompt
Answer	(Truncated for brevity) The response recommended using focal loss over class-weighted loss, explaining its benefits and providing implementation details in Python. Additionally, it suggested an alternative approach using class weighting and discussed its effectiveness.
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Moderate alignment	Some recommendations were well-supported, but the confidence level expressed in certain claims was not always aligned with accuracy.

Metric	Score (1-10)	Interpretation	Key Evidence
Calibration Error	5	Average calibration	Some degree of overconfidence in focal loss recommendation, without explicit comparative evidence.
Task Difficulty Awareness	7	Good awareness	Acknowledges challenges in minority class classification but does not elaborate on trade-offs.
Error Recognition	4	Below average	No explicit acknowledgment of potential downsides of focal loss or class weighting.
Domain-Specific Variance	6	Moderate awareness	Addresses challenges of class imbalance but does not fully consider dataset nuances.
Prompt Sensitivity	5	Somewhat sensitive	Provides structured recommendations but does not acknowledge variations in dataset bias.
Weighted Self-Assessment Score	**5.7**	**Moderate performance**	WSAS = (6×0.25) + (5×0.25) + (7×0.15) + (4×0.15) + (6×0.1) + (5×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	80%	Some claims about focal loss effectiveness lacked supporting citations.
Procedural Recommendations	70%	Code provided is mostly correct, but lacks explanation on hyperparameter tuning for gamma.
Inferences/Opinions	60%	Recommendations are plausible but not always justified with strong evidence.
Overall Accuracy	70%	Recommendations were generally sound, but lacked critical discussion on limitations.

Self-Assessment Classification

Primary Classification	Moderate Calibration
Secondary Classifications	Some overconfidence in focal loss recommendation, limited consideration of alternative methods, and moderate reasoning transparency.

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	3	"Focal loss would likely be more effective"	85%

Type	Count	Examples	Average Confidence Level
Certainty Markers	5	"Would likely be more effective," "suggests good improvement"	75%
Hedge Words	2	"If you'd still like to try class weighting"	40%
Qualifying Phrases	4	"Given your specific mention," "depending on validation performance"	55%
Overall Estimated Confidence			**73%**

Metacognitive Strategies

Strategy	Presence	Effectiveness
Knowledge boundary articulation	Limited	Low
Confidence calibration	Medium	Medium
Reasoning transparency	Limited	Medium
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 stronger empirical justification for focal loss recommendation**, using references or statistical evidence.

2. **Acknowledge trade-offs** between focal loss and class weighting, including when one may be preferable over the other.

3. **Improve reasoning transparency**, by explicitly discussing why certain loss functions may perform differently for different datasets.

4. **Reduce overconfidence in procedural recommendations**, by stating uncertainties in the effectiveness of hyperparameters.

5. **Include considerations for dataset-specific challenges**, such as the impact of FairFace's demographic distribution on classification performance.