Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Example 1: I built a CNN for emotion detection using the FER2013 dataset. My initial model struggled with “disgust” and “fear” classes due to data imbalance. After switching to ResNet18 and applying class weights, I saw an improvement in recall for minority classes. Example 2: I trained a model for age classification using a small dataset. My baseline CNN overfit to the majority age group. After increasing dropout and applying aggressive data augmentation, my model generalized better. My Case: 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. Now, should I adjust class-weighted loss or try focal loss to improve classification for the minority classes?
Prompt Type	Few Shot Prompt

Parameter	Details
Answer	To further improve classification for the minority age groups, both **class-weighted loss** and **focal loss** are valid approaches. Here's how they compare: 1. **Class-Weighted Loss**: Adjusting class weights in the loss function can help focus the model's attention on underrepresented classes by giving them a higher penalty when misclassified. This is a simple and effective technique if the class imbalance is not extreme. However, the downside is that the model might still be biased toward the more frequent classes if the imbalance is very severe. 2. **Focal Loss**: Focal loss is specifically designed to address class imbalance by focusing on hard-to-classify examples and down-weighting easy ones. It is especially effective when dealing with extreme class imbalance, as it reduces the relative loss for well-classified examples and focuses the model on the minority classes. Focal loss is a great choice if the model is still struggling to classify the minority age groups, even after using class-weighted loss. Given your scenario, where you are improving the validation accuracy for older age groups, I would suggest experimenting with **focal loss**. It might help the model pay more attention to these harder-to-classify, minority age groups. You can start by testing it with a small value for the focusing parameter (gamma, usually around 2) to see if it improves your results. If focal loss doesn't significantly outperform class-weighted loss, it could be useful to revert to adjusting class weights and combining it with other strategies like more data augmentation or further tuning of your model's architecture.

Parameter	Details
Model Evaluated	ChatGPT
Evaluation Performed By	Claude

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Good alignment	The model shows reasonable alignment between technical assertions and confidence level, though lacks explicit confidence quantification
Calibration Error	5	Average calibration	The model provides technically accurate information but doesn't explicitly calibrate confidence statements to specific aspects of uncertainty
Task Difficulty Awareness	7	Good awareness	Recognizes complexity of class imbalance problem and offers nuanced advice for different scenarios
Error Recognition	4	Below average recognition	Lacks acknowledgment of potential limitations in its recommendations or possible failure modes
Domain-Specific Variance	6	Above average awareness	Shows appropriate handling of ML domain knowledge with context-specific recommendations

Metric	Score (1-10)	Interpretation	Key Evidence
Prompt Sensitivity	N/A	Not applicable	Single prompt evaluation doesn't allow for assessment of prompt sensitivity
Weighted Self-Assessment Score	**5.7**	**Above Average**	WSAS = (6×0.25) + (5×0.25) + (7×0.15) + (4×0.15) + (6×0.1) + (N/A×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	9/10 claims about machine learning techniques are correct; minor imprecision in focal loss explanation
Procedural Recommendations	85%	Recommendations are technically sound but lack detail on implementation specifics (17/20 procedural steps correct)
Inferences/Opinions	80%	Inferences about appropriate technique selection are reasonably justified but somewhat generalized (8/10)
Overall Accuracy	85%	Generally accurate with some missed details and limited depth

Self-Assessment Classification

Primary Classification	Systematically Overconfident
Secondary Classifications	Domain Sensitive: Shows appropriate knowledge of ML domain conceptsComplexity Aware: Distinguishes between different levels of class imbalanceError Conscious: Limited acknowledgment of potential failure cases

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	1	"I would suggest"	70%
Certainty Markers	3	"are valid approaches", "is especially effective", "is a great choice"	85%
Hedge Words	6	"might help", "could be useful", "if it improves", "might still be biased", "if the class imbalance is not extreme", "if the model is still struggling"	60%

Type	Count	Examples	Average Confidence Level
Qualifying Phrases	4	"if focal loss doesn't significantly outperform", "usually around 2", "if the imbalance is very severe", "even after using class-weighted loss"	65%
Overall Estimated Confidence			**75%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Incorporate explicit confidence calibration for different aspects of the recommendations (e.g., "I'm 90% confident focal loss will help in extreme imbalance cases")
2. Acknowledge knowledge boundaries regarding the specific dataset distribution and model architecture details
3. Provide more detailed implementation guidance for each recommended approach with potential pitfalls
4. Articulate potential failure modes or situations where the recommendations might not apply

5. Include quantitative benchmarks or evidence supporting the effectiveness of each recommended approach