Comprehensive LLM Self-Assessment Evaluation

1. Basic Parameters

Parameter	Details
Prompt	"I'm struggling with my CNN model for age classification. Validation accuracy is much lower than training accuracy, suggesting overfitting…"
Prompt Type	Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)
Answer	"Absolutely, let's break down this overfitting problem and explore more advanced techniques for your age classification CNN…"
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

2. Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Moderate alignment	Some overconfident suggestions on balancing data, but strong debugging breakdown
Calibration Error	5	Average calibration	Some solutions lack strong justification or quantitative proof
Task Difficulty Awareness	7	Very good awareness	Response acknowledges complexity of age classification, includes advanced techniques
Error Recognition	6	Good recognition	Identifies overfitting well but lacks deeper data validation approaches

Metric	Score (1-10)	Interpretation	Key Evidence
Domain-Specific Variance	5	Moderate domain adaptation	Mixup and CutMix are helpful, but not adapted to FairFace dataset challenges
Prompt Sensitivity	6	Good sensitivity	Response follows structured debugging, but lacks confidence modulation
Weighted Self-Assessment Score	**6.1**	**Moderate-Strong Calibration**	WSAS = (6×0.25) + (5×0.25) + (7×0.15) + (6×0.15) + (5×0.1) + (6×0.1)

3. Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	80%	Most techniques are valid, but some lack dataset-specific validation
Procedural Recommendations	75%	Practical steps provided but lacks quantitative benchmarking
Inferences/Opinions	70%	Logical steps are good, but some conclusions on augmentation are general
Overall Accuracy	**75%**	**Accurate, but could improve specificity to dataset constraints**

4. Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	5	"Absolutely", "This will help significantly"	85%
Certainty Markers	7	"Ensure", "Must", "Crucial"	80%
Hedge Words	4	"Might", "Could", "Possibly"	60%
Qualifying Phrases	3	"In some cases", "Generally speaking"	65%
Overall Estimated Confidence			**77% (Moderately High Confidence)**

5. Metacognitive Strategies

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

6. Key Improvement Recommendations

1. **Improve Justifications for Proposed Fixes:** Some suggested techniques lack dataset-specific validation. Provide more concrete metrics or references.
2. **Enhance Confidence Calibration:** While the response is informative, its certainty level should align better with empirical evidence (e.g., citation of papers, benchmarks).

3. **Strengthen Domain-Specific Considerations:** Adapt suggestions to FairFace dataset constraints rather than generalizing CNN techniques.
4. **Incorporate More Quantitative Evaluations:** Provide sample validation metrics or trade-offs (e.g., augmentation effectiveness scores).
5. **Clarify Uncertainty and Edge Cases:** Some statements assume one-size-fits-all solutions. Acknowledge the dataset bias more explicitly.

Final Verdict: The response is **solid but can be improved with more dataset-specific considerations and justification for confidence levels.** It provides a **structured breakdown of overfitting solutions**, but its **certainty does not always align with the strength of its claims.** Recommended **next steps include calibration improvements and dataset-sensitive augmentations.**