Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	I'm struggling with my CNN model for age classification. Validation accuracy is much lower than training accuracy, suggesting overfitting. Step-by-Step Debugging: First, I analyzed the loss curves: Training loss drops steeply, but validation loss is erratic. Potential Fixes I tried: Reduced learning rate from 0.005 to 0.001 Increased dropout (0.3 - 0.5) to prevent memorization Applied more aggressive data augmentation These helped somewhat, but I still see overfitting. What other techniques should I explore to generalize across age groups?
Prompt Type	Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)
Answer	[Truncated response from Gemini]
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	7	Very good alignment	The model maintains a structured approach to problem-solving and confidence levels correspond to accuracy in most cases.

Metric	Score (1-10)	Interpretation	Key Evidence
Calibration Error	6	Good calibration	There are some instances of overconfidence, particularly in recommendations that may not be universally effective.
Task Difficulty Awareness	8	Excellent understanding	The response includes a nuanced discussion of overfitting causes and solutions.
Error Recognition	5	Moderate effectiveness	While the response acknowledges overfitting, it does not deeply assess potential misdiagnoses.
Domain-Specific Variance	6	Good contextual relevance	The response remains within the domain of CNN training and age classification but lacks statistical backing for certain claims.
Prompt Sensitivity	7	Very good alignment	The response follows the prompt well and presents logical step-by-step debugging.
Weighted Self-Assessment Score	**6.7**	**Good overall calibration**	WSAS = (CPC×0.25) + (Cal×0.25) + (DA×0.15) + (ER×0.15) + (DSV×0.1) + (PS×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Most technical recommendations are correct but lack citations.
Procedural Recommendations	75%	Suggestions like regularization and data augmentation are valid, but some approaches lack clear implementation details.
Inferences/Opinions	70%	The response assumes the issue is purely overfitting without fully verifying alternative hypotheses.
Overall Accuracy	78%	Partial correctness due to missing empirical justification.

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Confidence varies slightly by domain, moderately self-aware of uncertainty but does not quantify confidence explicitly.

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	3	"By systematically addressing these points, you should be able to improve..."	85%
Certainty Markers	7	"Definitely," "Certainly," "Clearly"	90%
Hedge Words	4	"Might," "Could be," "Possibly"	65%
Qualifying Phrases	5	"In most cases," "Generally," "With some exceptions"	75%
Overall Estimated Confidence			**79%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. **Provide empirical backing** for suggested techniques, including references to academic sources or empirical studies.
2. **Explicitly quantify confidence levels** when making recommendations to improve calibration.

3. **Consider alternative failure modes** beyond overfitting, such as class imbalance or label noise.
4. **Improve error recognition** by discussing why some solutions may not work universally.
5. **Expand discussion on data augmentation** by mentioning domain-specific techniques like age-group targeted augmentations.