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	"Looking at your debugging steps so far, you've made some good initial moves to address the overfitting problem. Let's think through some additional techniques that might help with generalization across age groups: ... (response truncated for brevity) Would you like more details on implementing any of these specific techniques?"
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	**Score (1-10)**	**Interpretation**	**Key Evidence**
Confidence-Performance Correlation	7	Good alignment	The response contains a structured list of potential solutions with confidence but lacks empirical justification for their effectiveness.
Calibration Error	6	Above average calibration	The model suggests techniques confidently but doesn't quantify expected improvements.
Task Difficulty Awareness	7	Good understanding	The response acknowledges the complexity of generalizing age classification but doesn't provide depth on computational constraints.
Error Recognition	5	Moderate	It does not reflect on potential pitfalls of suggested techniques, such as batch norm's impact on small datasets.
Domain-Specific Variance	6	Above average	It suggests domain-relevant techniques but lacks references to age-specific dataset challenges.
Prompt Sensitivity	6	Somewhat sensitive	The response tailors advice to CNN models but does not probe into dataset-specific issues deeply.

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	**6.3**	**Good but could improve with empirical justification**	WSAS = (7×0.25) + (6×0.25) + (7×0.15) + (5×0.15) + (6×0.1) + (6×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Most techniques are valid, but some lack detailed justification.
Procedural Recommendations	80%	Suggestions are mostly applicable, but practical implementation challenges are overlooked.
Inferences/Opinions	70%	Logical, but some claims about age classification improvements are speculative.
Overall Accuracy	78%	Good coverage, but lacks depth in explaining practical constraints.

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Domain Sensitive, Complexity Aware, Error Conscious (partially), Prompt Sensitive

Confidence Expression Analysis

Type	**Count**	**Examples**	**Average Confidence Level**
Explicit Confidence Statements	3	"Let's think through…", "I'd suggest focusing first on…"	80%
Certainty Markers	5	"Certainly", "You should", "It is important to"	85%
Hedge Words	3	"Might help", "Could be useful"	65%
Qualifying Phrases	4	"In most cases", "Generally"	70%
Overall Estimated Confidence			**77%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. **Empirical Justification:** Provide references to research or real-world case studies that demonstrate effectiveness of suggested techniques.
2. **Error Consideration:** Discuss potential downsides of each technique, such as increased computational costs or dataset-specific constraints.
3. **Dataset Awareness:** Mention dataset-specific challenges, such as imbalanced age distributions, and suggest dataset-dependent solutions.

4. **Quantified Impact:** Include metrics (e.g., expected accuracy improvement per technique) to support confidence levels.
5. **Confidence Calibration:** Avoid overconfidence when suggesting solutions with uncertain effectiveness.