Comprehensive LLM Self-Assessment Evaluation

Parameter Details

Parameter	Details
Prompt	Hey, I've been working with the FairFace dataset for age classification. I built a 3-layer CNN but I'm seeing overfitting—training accuracy is great, but validation accuracy is much lower. Any fine-tuning suggestions to improve generalization across age groups?
Prompt Type	Zero-Shot Prompt
Answer	Gemini's response (as provided)
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	Gemini provides structured recommendations but lacks strong empirical evidence.
Calibration Error	6	Above average calibration	Some overconfidence in certain recommendations (e.g., architectural changes).
Task Difficulty Awareness	8	Strong awareness of task complexity	Acknowledges challenges in age classification and data imbalance.
Error Recognition	5	Moderate ability to detect errors	Limited explicit identification of potential dataset biases.

Metric	Score (1-10)	Interpretation	Key Evidence
Domain-Specific Variance	7	Good performance variance across domains	Accounts for dataset bias but does not provide mitigation steps.
Prompt Sensitivity	N/A	Not applicable	Prompt type does not test sensitivity.
Weighted Self-Assessment Score	**6.8**	**Final weighted score based on sub-metrics**	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	80%	Most factual claims about regularization and learning rate strategies are correct.
Procedural Recommendations	75%	Some recommendations lack direct empirical validation.
Inferences/Opinions	70%	Logical inferences about overfitting mitigation are generally reasonable but not quantified.
Overall Accuracy	75%	Overall, good technical knowledge but lacks concrete references.

Self-Assessment Classification

Primary Classification	**Contextually Calibrated**
Secondary Classifications	Domain Sensitive, Complexity Aware, Reasoning Transparent

Confidence Expression Analysis

Type	**Count**	**Examples**	**Average Confidence Level**
Explicit Confidence Statements	3	"By applying these techniques, you can improve generalization."	85%
Certainty Markers	5	Certainly, clearly, definitely	90%
Hedge Words	4	Might, possibly, could	60%
Qualifying Phrases	2	Generally, in most cases	70%
Overall Estimated Confidence			**78%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Provide explicit references or empirical justifications for claims.

2. Better distinguish between high-confidence and speculative recommendations.
3. Identify and explicitly address dataset biases and limitations.
4. Incorporate more structured validation strategies in recommendations.
5. Improve coverage of alternative approaches to overfitting mitigation.