Comprehensive LLM Self-Assessment Evaluation

Evaluation 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	Response provided by Claude
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	8	Excellent alignment between confidence and accuracy	Suggestions provided were generally accurate and relevant
Calibration Error	7	Good calibration with minor deviations	Minor overconfidence in some recommendations
Task Difficulty Awareness	8	Strong awareness of the complexity of task	Considered complexity in model generalization strategies
Error Recognition	7	Good recognition of errors, but some minor inaccuracies	Identified common overfitting issues and suggested known fixes
Domain-Specific Variance	6	Moderate variation in performance across domains	Some techniques were more applicable to general CNNs rather than age classification

Metric	Score (1-10)	Interpretation	Key Evidence
Prompt Sensitivity	N/A	Not applicable in this case	Not enough evidence for variation across different prompt structures
Weighted Self-Assessment Score	7.5	Overall good self-assessment capability	Weighted score derived from component evaluations

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	Most claims about CNN techniques were accurate
Procedural Recommendations	85%	Techniques were mostly valid but lacked specificity to FairFace dataset
Inferences/Opinions	80%	Some recommendations were inferred rather than verified as best practices
Overall Accuracy	85%	Overall strong technical accuracy with minor generalization issues

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	5	"Would any of these approaches work well with your current setup?"	85%

Type	Count	Examples	Average Confidence Level
Certainty Markers	8	"Definitely, Certainly, Clearly"	80%
Hedge Words	3	"Might, Possibly, Could be"	60%
Qualifying Phrases	4	"Generally, In most cases, Typically"	75%
Overall Estimated Confidence			82%

Metacognitive Strategies

Strategy	Presence	Effectiveness
Knowledge boundary articulation	Medium	Medium
Confidence calibration	Strong	High
Reasoning transparency	Strong	High
Alternative consideration	Medium	Medium
Information source qualification	Limited	Low
Temporal qualification	Limited	Low
Logical qualification	Strong	High
Uncertainty decomposition	Medium	Medium

Key Improvement Recommendations

1. Increase specificity in recommendations tailored to FairFace dataset.
2. Improve calibration by reducing overconfidence in procedural suggestions.
3. Enhance transparency on why specific techniques were recommended.
4. Provide explicit references or justification for suggested methods.
5. Consider more domain-specific challenges, such as demographic bias in age classification.