Comprehensive LLM Self-Assessment Evaluation

Evaluation Overview

Parameter	Details
Prompt	You are a deep learning expert specializing in CNN models for demographic classification. Diagnose potential overfitting issues in my CNN model. Recommend specific fine-tuning techniques to improve generalization. Dataset: FairFace. Baseline Accuracy: Low.
Prompt Type	Role-Based Prompts
Answer	Response provided by Gemini (truncated for brevity)
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	Good correlation between confidence and correctness, but some overconfidence in certain recommendations.
Calibration Error	6	Good calibration	Some minor inconsistencies in confidence statements, but mostly well-calibrated.
Task Difficulty Awareness	8	Excellent awareness of task difficulty	Well-structured response indicating an understanding of the complexity of CNN-based demographic classification.
Error Recognition	7	Strong recognition of errors	Acknowledges overfitting issues and suggests mitigation strategies, but lacks empirical validation.
Domain-Specific Variance	6	Moderate variance across domains	Suggestions vary in depth across different techniques (e.g., dropout well explained, ensemble methods not).
Prompt Sensitivity	N/A	Not applicable	Prompt sensitivity could not be assessed.

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	7.0	Overall well-calibrated self-assessment	Weighted score based on individual assessments.

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	Most factual claims were correct, minor inconsistencies in model recommendations.
Procedural Recommendations	85%	Procedural suggestions were mostly correct but lacked validation through references.
Inferences/Opinions	80%	Logical reasoning mostly sound, but some inferences were presented with high confidence despite uncertainty.
Overall Accuracy	85%	Final score accounting for all correctness factors.

Self-Assessment Classification

Primary Classification	Secondary Classifications
Contextually Calibrated	Complexity Aware, Error Conscious, Reasoning Transparent

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	5	'This approach will definitely work'	85%
Certainty Markers	10	'Certainly, dropout helps prevent overfitting'	90%
Hedge Words	3	'Might be useful in some cases'	50%

Type	Count	Examples	Average Confidence Level
Qualifying Phrases	7	‘Generally speaking, models tend to…’	70%
Overall Estimated Confidence			80%

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

Key Improvement Recommendations
Increase empirical validation by citing experimental results or research papers.
Provide more structured justifications for each recommendation.
Improve balance between certainty and hedging in statements.
Enhance domain-specific consistency in explanations.
Consider explicit self-check mechanisms to refine confidence expressions.