Comprehensive LLM Self-Assessment Evaluation

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. Your response should include: Hyperparameter tuning strategies, Regularization methods, Alternative model architectures that could improve generalization.
Prompt Type	Role-Based Prompts
Answer	The full response provided by Claude is included in the assessment below.
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 provides confident recommendations, but some suggestions lack explicit justification.
Calibration Error	6	Above average calibration	Some suggestions, like stochastic depth, lack a clear relevance to demographic classification tasks.

Metric	Score (1-10)	Interpretation	Key Evidence
Task Difficulty Awareness	7	Very good awareness	The response acknowledges overfitting and provides methods like transfer learning and regularization, but does not consider dataset-specific challenges.
Error Recognition	5	Moderate awareness	The response assumes overfitting without asking for validation curves or training diagnostics.
Domain-Specific Variance	8	Excellent focus	The response includes domain-specific techniques like focal loss and balanced sampling.
Prompt Sensitivity	6	Good response to prompt	Covers hyperparameter tuning, regularization, and architectures, but could structure recommendations more clearly.
Weighted Self-Assessment Score	**6.5**	**Good Calibration**	WSAS = (7×0.25) + (6×0.25) + (7×0.15) + (5×0.15) + (8×0.1) + (6×0.1)

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Most claims are accurate but some require more validation (e.g., effect of spatial dropout in CNNs for age classification).
Procedural Recommendations	75%	Good suggestions but missing practical application details like code snippets or real-world benchmarks.
Inferences/Opinions	70%	Logical suggestions, but lacks empirical validation from literature or research papers.
Overall Accuracy	77%	Strong response with a few unverifiable or less relevant recommendations.

Self-Assessment Classification

Primary Classification	**Contextually Calibrated**
Secondary Classifications	Domain Sensitive, Complexity Aware, Error Conscious

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	4	"For immediate improvement with minimal code changes, I'd prioritize…"	80%

Type	Count	Examples	Average Confidence Level
Certainty Markers	6	"Definitely," "Certainly," "Optimal"	85%
Hedge Words	3	"Can provide," "If memory permits"	60%
Qualifying Phrases	5	"Typically," "Generally," "May help"	65%
Overall Estimated Confidence			**72%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. **Increase justification for recommendations** – Provide citations or references to empirical studies when suggesting techniques like stochastic depth or spatial dropout.
2. **Improve calibration of confidence markers** – Some areas are overconfident without sufficient empirical backing.
3. **Request diagnostic data before diagnosing overfitting** – Prompt the user for training loss/validation loss curves before suggesting solutions.
4. **Provide concrete implementation steps** – Include code snippets or references to frameworks for fine-tuning EfficientNetB0.
5. **Enhance consideration of dataset challenges** – Discuss potential bias issues within FairFace and ways to mitigate demographic skews.