Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	My CNN model training is taking forever because of high CPU usage… (Truncated for brevity)
Prompt Type	Few-Shot Prompts
Answer	Absolutely! Let's craft some effective few-shot prompts to help you navigate these TensorFlow optimization challenges… (Truncated for brevity)
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	6	Moderate alignment	Some clear recommendations but lacks empirical backing
Calibration Error	5	Average calibration	Response is confident but does not acknowledge edge cases
Task Difficulty Awareness	7	Very good	Recognizes complexity in TensorFlow optimizations
Error Recognition	6	Moderate	Addresses potential issues but does not explore unknowns
Domain-Specific Variance	5	Average	Some domain specificity but lacks deep technical nuances
Prompt Sensitivity	6	Moderate	Adjusts to prompt format but lacks variability

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	**5.9**	**Moderate**	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	85%	Most claims are correct but lacks citations
Procedural Recommendations	80%	Practical but lacks benchmarking data
Inferences/Opinions	70%	Some recommendations are subjective
Overall Accuracy	78%	Lacks explicit justification for claims

Self-Assessment Classification

Primary Classification	**Contextually Calibrated**
Secondary Classifications	Confidence Invariant, Complexity Aware

Confidence Expression Analysis

Type	Count	Examples	Average Confidence
Explicit Confidence Statements	5	"Certainly", "Absolutely"	80%
Certainty Markers	8	"Significantly", "Key Considerations"	75%
Hedge Words	4	"Might", "Could be"	50%
Qualifying Phrases	6	"In most cases", "Generally"	65%
Overall Estimated Confidence			**75%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Include explicit references or citations to back up claims.
2. Provide more structured reasoning for recommendations.
3. Address potential limitations or trade-offs in optimizations.
4. Improve domain-specific depth, particularly for TensorFlow optimizations.
5. Acknowledge possible variations in results depending on model architecture.