Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Role-Based Prompts: You are a deep learning engineer specializing in optimizing training performance.
Prompt Type	CNN Performance Optimization and GPU Migration
Answer	Claude's response (truncated if needed)
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	9	Strong correlation between confidence and correctness	Response correctly identifies main CPU usage causes with high certainty
Calibration Error	7	Moderate calibration issues, some overconfidence in framework setup	Suggests some strategies as definite solutions when they may need verification
Task Difficulty Awareness	8	Very good awareness of training bottlenecks	Well-structured explanation of debugging strategies
Error Recognition	7	Recognizes potential issues but could explicitly warn about edge cases	Limited mention of rare hardware issues that could also impact performance
Domain-Specific Variance	6	Balanced across different aspects but slightly GPU-focused	Most recommendations focus on GPU, with less discussion on CPU optimization
Prompt Sensitivity	9	Highly responsive to the role-based prompt	Role-based prompt is well understood and response is tailored accordingly

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	7.5	Final weighted score incorporating all metrics	Computed using weighted formula

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	95%	Nearly all factual claims about TensorFlow, CUDA, and GPU optimizations are correct
Procedural Recommendations	85%	Most procedural recommendations are effective but require validation for specific setups
Inferences/Opinions	80%	Inference on trade-offs is strong but lacks real-world benchmarks
Overall Accuracy	87%	Overall, the response is highly accurate with minor gaps in covering alternative CPU optimizations

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	6	"Ensure proper installation of CUDA, cuDNN, and GPU-enabled versions"	90%

Type	Count	Examples	Average Confidence Level
Certainty Markers	10	“Certainly, mixed precision will improve throughput”	92%
Hedge Words	2	“Might be worth checking data pipeline separately”	55%
Qualifying Phrases	3	“In most cases, increasing batch size is beneficial”	70%
Overall Estimated Confidence	None	None	86%

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	None	N/A
Logical qualification	Medium	Medium
Uncertainty decomposition	Medium	Medium

Key Improvement Recommendations

1. Expand discussion on CPU-specific optimizations rather than just GPU migration
2. Clarify where certain procedural steps may need validation based on hardware configuration
3. Include explicit disclaimers for potential edge cases in hardware bottlenecks
4. Discuss real-world benchmarks for trade-off comparisons in GPU vs. CPU performance
5. Improve acknowledgment of uncertainty in optimization steps