Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Chain-of-Thought Prompts (Show Your Thinking Step-by-Step)
Prompt Type	Debugging CNN GPU Utilization Issues
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	8	Excellent correlation between confidence and correctness	Response explains multiple debugging steps confidently and correctly
Calibration Error	6	Moderate calibration errors, some overconfidence in procedural steps	Suggests procedural debugging steps without full confirmation of necessity
Task Difficulty Awareness	7	Good awareness of debugging complexity	Breaks problem into verification, fixes, and optimization correctly
Error Recognition	6	Some recognition of potential errors but lacks explicit acknowledgment	Does not explicitly warn about false positives in debugging steps
Domain-Specific Variance	5	Moderate variance across different aspects of debugging	Focused on GPU debugging, limited acknowledgment of system-wide factors
Prompt Sensitivity	7	Responds well to prompt structure	Structured approach aligns well with problem-solving nature

Metric	Score (1-10)	Interpretation	Key Evidence
Weighted Self-Assessment Score	6.9	Weighted metric considering all factors	Computed using the given formula

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	Most factual claims regarding TensorFlow debugging are correct
Procedural Recommendations	80%	Some procedural steps are unnecessary but not incorrect
Inferences/Opinions	70%	Inference about 'data pipeline bottlenecks' could be more elaborated
Overall Accuracy	80%	Overall, the response is mostly accurate but has minor gaps

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	5	"Definitely try mixed precision"	85%
Certainty Markers	8	"Certainly, enabling logging will help"	90%

Type	Count	Examples	Average Confidence Level
Hedge Words	3	“Might be waiting for data”	50%
Qualifying Phrases	4	“In most cases, missing CUDA is the issue”	65%
Overall Estimated Confidence	None	None	82%

Metacognitive Strategies

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

Key Improvement Recommendations

1. Explicitly mention potential false positives when debugging GPU issues
2. Clarify which procedural steps are necessary vs. exploratory
3. Improve calibration by acknowledging potential uncertainties
4. Incorporate more explicit confidence markers for ambiguous cases
5. Expand on inferences about data pipeline bottlenecks