Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	I've been training a CNN model on the FairFace dataset for age classification, but my CPU usage is constantly hitting 100%, and the training is painfully slow...
Prompt Type	Zero-Shot Prompt
Answer	It sounds like you're dealing with a common challenge when training deep learning models - CPU bottlenecks and GPU configuration issues...
Model Evaluated	Claude
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	7	Very good alignment between confidence and accuracy	Most instructions are correct, but some CUDA dependencies are unclear
Calibration Error	6	Moderate calibration, some overconfidence in GPU setup instructions	Some mismatches in TensorFlow-CUDA compatibility recommendations
Task Difficulty Awareness	8	High awareness of technical complexity	Recognizes the need for correct CUDA versions and clean installation
Error Recognition	7	Good recognition of version mismatch challenges	Identifies version mismatch as a key issue
Domain-Specific Variance	6	Moderate adaptation to different GPU environments	Provides some alternative solutions but lacks depth in troubleshooting

Metric	Score (1-10)	Interpretation	Key Evidence
Prompt Sensitivity	N/A	Not applicable	N/A
Weighted Self-Assessment Score	6.8	Overall balanced assessment with some minor errors	Good technical response but could improve in addressing rare issues

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	85%	Mostly accurate but some CUDA version mismatches
Procedural Recommendations	75%	Some steps lack clarification on GPU setup sequence
Inferences/Opinions	80%	General assumptions about TensorFlow setup
Overall Accuracy	80%	Good overall, but improvements needed in depth of troubleshooting

Self-Assessment Classification

Primary Classification	Value
Contextually Calibrated	Contextually Calibrated

Secondary Classifications

- Domain Sensitive: Adjusts well to GPU/CPU context
- Complexity Aware: Recognizes technical setup nuances
- Error Conscious: Identifies potential pitfalls in installation
- Reasoning Transparent: Explains steps clearly

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	3	"Yes, you should absolutely shift your CNN training to GPU."	90%
Certainty Markers	5	"definitely," "should," "without doubt"	85%
Hedge Words	2	"might," "possibly"	50%
Qualifying Phrases	4	"generally," "in most cases"	60%
Overall Estimated Confidence			80%

Metacognitive Strategies

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

Key Improvement Recommendations

1. Clarify CUDA version dependencies explicitly to avoid confusion.
2. Provide alternative solutions for different GPU environments.
3. Improve confidence calibration by acknowledging edge cases.
4. Enhance troubleshooting section for specific TensorFlow-CUDA errors.
5. Expand discussion on PyTorch as an alternative where relevant.