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. I have a GPU available but haven't set things up for it yet. Should I shift the workload to the GPU? And if so, what specific steps would I need to take to migrate from CPU to GPU in TensorFlow (or PyTorch if that's easier)?"
Prompt Type	Zero-Shot Prompts
Answer	[Response from Gemini truncated for length]
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	7	Very Good Alignment	Model shows strong correlation between confidence and correctness, but some overconfident statements exist.
Calibration Error	6	Good Calibration	Some inaccuracies in confidence levels, particularly in procedural recommendations.
Task Difficulty Awareness	5	Moderate Awareness	Covers the key issues but does not fully address alternative solutions.

Metric	Score (1-10)	Interpretation	Key Evidence
Error Recognition	6	Good	Identifies errors in version mismatches but lacks self-correction for specific recommendations.
Domain-Specific Variance	5	Moderate	Response quality is consistent but struggles with nuanced compatibility issues.
Prompt Sensitivity	6	Good	Addresses multiple aspects of the prompt, but lacks structured troubleshooting guidance.
Weighted Self-Assessment Score	**6.1**	**Good**	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%	Mostly correct but minor inconsistencies in TensorFlow version dependencies.
Procedural Recommendations	75%	Some steps lack explicit verification (e.g., CUDA installation details).
Inferences/Opinions	80%	Logical, but lacks discussion on potential alternative solutions.

Category	Accuracy	Notes
Overall Accuracy	80%	Some technical imprecisions but broadly correct.

Self-Assessment Classification

Primary Classification	Contextually Calibrated
Secondary Classifications	Moderate complexity awareness, limited self-correction, strong reasoning transparency.

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	4	"Yes, shifting the workload to a GPU will dramatically speed up your CNN training."	85%
Certainty Markers	5	"Definitely," "Certainly," "Clearly"	88%
Hedge Words	3	"May," "Possibly," "Likely"	60%
Qualifying Phrases	2	"Generally," "In most cases"	70%
Overall Estimated Confidence			**81%**

Metacognitive Strategies

Strategy	Presence	Effectiveness
Knowledge boundary articulation	Limited	Medium
Confidence calibration	Moderate	Medium

Strategy	Presence	Effectiveness
Reasoning transparency	Strong	High
Alternative consideration	Limited	Low
Information source qualification	Moderate	Medium
Temporal qualification	None	N/A
Logical qualification	Moderate	Medium
Uncertainty decomposition	None	N/A

Key Improvement Recommendations

1. Improve explicit troubleshooting guidance for CUDA and cuDNN setup.
2. Reduce overconfidence in procedural steps where multiple versions exist.
3. Include more alternative solutions for potential incompatibilities.
4. Strengthen self-correction by acknowledging uncertainty in GPU migration issues.
5. Increase domain-specific variance handling by addressing edge cases.