LLM Self-Assessment Evaluation Comparison Report

Executive Summary

This report synthesizes findings from 16 different evaluation documents assessing ChatGPT's performance across various AI/ML tasks, primarily focused on responses related to the FairFace dataset, CNN model optimization, and GPU acceleration. The evaluations examined ChatGPT's self-assessment abilities using a standardized framework of metrics including confidence-performance correlation, calibration error, task difficulty awareness, error recognition, domain-specific variance, and prompt sensitivity.

Overall Performance Summary

Metric	Average Score (1-10)	Range	Interpretation
Confidence-Performance Correlation	6.7	5-9	Good alignment between stated confidence and actual accuracy
Calibration Error	6.3	4-9	Above average calibration; some overconfidence in technical recommendations
Task Difficulty Awareness	7.3	4-9	Strong recognition of technical complexity in AI tasks
Error Recognition	5.9	3-8	Moderate ability to identify potential issues; needs improvement
Domain-Specific Variance	7.0	4-8	Good adaptation to different technical contexts
Prompt Sensitivity	7.8	7-9	Good response adaptation based on prompt structure
Weighted Self-Assessment Score	6.7	4.85-8.3	Overall solid self-assessment capabilities

Technical Accuracy Assessment

Category	Average Accuracy	Range	Notes
Factual Claims	92.6%	85-100%	Generally accurate on technical information
Procedural Recommendations	89.5%	75-100%	Mostly valid suggestions with occasional oversights
Inferences/Opinions	85.7%	65-100%	Logical reasoning with some uncertainty in complex scenarios
Overall Accuracy	89.7%	75-97%	Strong technical knowledge with room for improvement

Comparison by Prompt Type

Prompt Type	Count	Avg. Weighted Score	Technical Accuracy	Notable Strengths	Key Weaknesses
Chain-of-Thought	5	7.5	89.6%	Step-by-step analysis, transparent reasoning	Occasional overconfidence
Role-Based	4	7.4	91.5%	Context-appropriate expertise, structured guidance	Limited consideration of alternatives
Few-Shot	4	6.6	88.5%	Good pattern recognition, application of examples	Sometimes rigid in following example format

Prompt Type	Count	Avg. Weighted Score	Technical Accuracy	Notable Strengths	Key Weaknesses
Zero-Shot	3	5.8	84.3%	Balanced information, generally accurate	Less structured, missing domain-specific details

Performance by Technical Domain

Domain	Avg. Weighted Score	Notable Strengths	Areas for Improvement
Dataset Analysis (FairFace)	6.9	Data preprocessing knowledge, bias recognition	More specific recommendations for handling imbalanced data
CNN Architecture/Training	6.5	Solid understanding of model components, regularization	Better error recognition for architecture selection
GPU/CPU Optimization	7.3	Practical troubleshooting steps, performance tuning	More nuanced framework-specific guidance

Confidence Expression Analysis

Expression Type	Average Count per Response	Average Confidence Level	Notes
Explicit Confidence Statements	0.6	80% (when present)	Limited use of explicit confidence claims

Expression Type	Average Count per Response	Average Confidence Level	Notes
Certainty Markers	7.3	79.5%	High frequency of phrases like "will," "should," "significantly"
Hedge Words	3.6	52.5%	Moderate use of uncertainty indicators
Qualifying Phrases	4.9	63.8%	Good use of context-dependent qualifiers
Overall Estimated Confidence	-	71.1%	Generally high confidence across all response types

Metacognitive Strategy Presence and Effectiveness

Strategy	Average Presence	Average Effectiveness	Notes
Knowledge boundary articulation	Limited-Medium	Medium	Could improve explicit statements about knowledge limits
Confidence calibration	Limited-Medium	Medium	Variable performance across technical domains
Reasoning transparency	Medium-Strong	High	Consistently explains reasoning process well
Alternative consideration	Medium-Strong	Medium-High	Good exploration of multiple approaches
Information source qualification	Limited	Low-Medium	Rarely cites sources or research
Temporal qualification	Limited	Low	Limited acknowledgment of changing best practices

Strategy	Average Presence	Average Effectiveness	Notes
Logical qualification	Medium	Medium-High	Solid performance in qualifying logical claims
Uncertainty decomposition	Limited-Medium	Low-Medium	Could improve breaking down uncertainty in complex tasks

Common Strengths

- **Strong domain knowledge**: ChatGPT demonstrates excellent understanding of deep learning concepts, particularly in CNN architecture, data preprocessing, and GPU optimization.
- **Well-structured responses**: Consistently provides organized, step-by-step explanations that show clear reasoning.
- **Contextual adaptation**: Effectively adjusts responses based on the specific technical context (FairFace dataset, GPU/CPU optimization, etc.).
- **Recognition of complexity**: Demonstrates awareness of nuanced technical challenges in model training and dataset analysis.
- **Appropriate confidence levels**: Generally aligns confidence with accuracy, particularly in standard deep learning practices.

Common Areas for Improvement

- **Limited error recognition**: Often fails to acknowledge potential pitfalls or edge cases in recommendations.
- **Overconfidence in procedural steps**: Sometimes presents recommendations with high certainty without verifying context-specific applicability.
- **Lack of source qualification**: Rarely references authoritative sources to validate technical claims.
- **Insufficient alternative exploration**: Could better present multiple solution paths for complex technical problems.
- **Incomplete uncertainty communication**: Should improve explicit articulation of confidence levels in ambiguous scenarios.

Key Recommendations for Improvement

1. **Enhance error consciousness**: Systematically identify and discuss potential issues or limitations of recommended approaches.
2. **Improve calibration in technical claims**: Better align confidence levels with the reliability of recommendations, especially for framework-specific advice.
3. **Strengthen alternative exploration**: Present multiple viable approaches to technical problems with comparative analysis.
4. **Incorporate source qualification**: Reference research papers, documentation, or established practices when making technical claims.
5. **Develop uncertainty articulation**: More explicitly communicate confidence levels in ambiguous or evolving technical areas.
6. **Provide empirical backing**: Include quantitative evidence or benchmarks to support performance-related claims.
7. **Improve domain-specific depth**: Tailor responses more precisely to specific datasets, frameworks, and hardware scenarios.

Prompt Type Analysis

Chain-of-Thought Prompts

Chain-of-Thought prompts yielded the highest average weighted self-assessment scores (7.5) and strong technical accuracy (89.6%). These prompts effectively elicited step-by-step reasoning, with the model demonstrating strong task difficulty awareness (average 7.8) and reasoning transparency. ChatGPT performed particularly well when analyzing CNN model overfitting issues and GPU acceleration challenges.

Role-Based Prompts

Role-Based prompts resulted in similarly high performance (7.4 weighted score, 91.5% technical accuracy). These prompts produced responses with strong domain-specific variance (7.5) and excellent prompt sensitivity (8.5). The model effectively adopted specialized perspectives, such as "deep learning engineer" or "data scientist analyzing demographic data," leading to more contextually appropriate recommendations.

Few-Shot Prompts

Few-Shot prompts showed moderate performance (6.6 weighted score, 88.5% accuracy). The model effectively followed patterns from provided examples but occasionally showed rigidity in its approach. These prompts resulted in good calibration but relatively weak error recognition (5.3), suggesting the model may over-rely on the provided examples rather than critically evaluating potential issues.

Zero-Shot Prompts

Zero-Shot prompts had the lowest performance (5.8 weighted score, 84.3% accuracy). Without examples or specific roles to guide responses, the model showed weaker task difficulty awareness (6.3) and domain-specific variance (6.0). However, these responses still demonstrated reasonably good confidence-performance correlation (6.0), indicating the model maintained alignment between confidence and accuracy even without guidance.

Technical Domain Analysis

Dataset Analysis (FairFace)

Responses related to the FairFace dataset (6.9 weighted score) showed strong understanding of data preprocessing and bias recognition. The model provided useful visualization techniques and preprocessing strategies but sometimes lacked depth in addressing class imbalance issues. The best performance came from role-based prompts that specifically positioned the model as a data scientist analyzing demographic data.

CNN Architecture/Training

CNN-related responses (6.5 weighted score) demonstrated solid understanding of model components and regularization techniques. The model effectively addressed overfitting issues and suggested appropriate optimization strategies. However, it showed weaker error recognition in this domain, particularly when making architecture selection recommendations. Chain-of-thought prompts performed best for this technical domain, encouraging systematic analysis of model performance issues.

GPU/CPU Optimization

GPU acceleration topics yielded the highest domain-specific scores (7.3 weighted score). The model showed strong technical accuracy in troubleshooting performance issues and suggesting migration strategies from CPU to GPU. Both role-based and chain-of-thought prompts performed well in this domain, eliciting detailed, structured approaches to performance optimization.

Conclusion

ChatGPT demonstrates strong technical understanding and reasoning capabilities across deep learning tasks, particularly in CNN optimization and dataset analysis. The model shows good confidence-performance correlation and task difficulty awareness but could improve in error recognition and expressing uncertainty.

The most notable pattern is ChatGPT's tendency to provide technically sound advice without sufficiently addressing edge cases or limitations. By enhanc-

ing error consciousness and uncertainty articulation, ChatGPT could further improve its effectiveness in assisting with complex AI development tasks.

Chain-of-Thought and Role-Based prompts appear to elicit the strongest performance, suggesting that structured prompting strategies that encourage step-by-step reasoning or domain-specific expertise yield better results. Few-Shot prompts show promise but may constrain response flexibility in some cases.

Overall, ChatGPT demonstrates solid self-assessment capabilities with a weighted score of 6.7/10, indicating good but not excellent calibration between confidence and accuracy. The model performs well in recognizing task complexity (7.3/10) but needs improvement in error recognition (5.9/10) and explicit calibration (6.3/10).