Comparative Analysis of LLM Self-Assessment Capabilities

Executive Summary

This report presents a comprehensive comparison of three leading large language models (LLMs)—ChatGPT, Gemini, and Claude—focusing specifically on their self-assessment capabilities in AI/ML tasks. Based on evaluation reports where each model was tested on identical tasks related to the FairFace dataset analysis, CNN model optimization, and GPU acceleration across 16 different assessments, we analyze how accurately these models can evaluate their own knowledge and confidence.

Our findings reveal distinct patterns of strengths and weaknesses across the models. ChatGPT demonstrates superior technical accuracy (89.7%) but moderate self-assessment capabilities. Claude shows the best confidence-performance correlation (7.44/10) and overall calibration. Gemini presents balanced but generally lower performance across most metrics. Notably, all models struggle significantly with error recognition, highlighting a universal area for improvement in current LLM technology.

Study Methodology

Each model was evaluated using a standardized framework measuring performance across metrics including:

- **Confidence-performance correlation**: How well a model's expressed confidence matches its actual accuracy
- **Calibration error**: The magnitude of miscalibration in confidence levels
- **Task difficulty awareness**: Recognition of complexity in technical tasks
- **Error recognition**: Ability to identify potential issues or limitations
- **Domain-specific variance**: Adaptation to different technical contexts
- **Prompt sensitivity**: Response adaptation based on prompt structure

Models were tested using four prompt types (Chain-of-Thought, Role-Based, Few-Shot, and Zero-Shot) across three technical domains (Dataset Analysis, CNN Architecture/Training, and GPU/CPU Optimization).

Overall Performance Comparison

Metric	ChatGPT	Gemini	Claude	Insights
Weighted Self-Assessment Score	6.70/10	6.32/10	6.78/10	Claude demonstrates slightly better overall self-assessment

Metric	ChatGPT	Gemini	Claude	Insights
Technical Accuracy	89.7%	81.9%	85.6%	ChatGPT shows highest technical knowledge
Confidence-Performance Correlation	6.70/10	6.53/10	7.44/10	Claude shows superior alignment between confidence and accuracy
Calibration Error	6.30/10	5.56/10	6.69/10	Claude demonstrates best calibration
Task Difficulty Awareness	7.30/10	7.06/10	7.19/10	All models recognize complexity well
Error Recognition	5.90/10	5.44/10	5.88/10	Universal weakness across all models
Domain-Specific Variance	7.00/10	6.13/10	6.25/10	ChatGPT shows better adaptation across domains
Prompt Sensitivity	7.80/10	6.67/10	6.86/10	ChatGPT responds best to different prompt structures

Technical Accuracy Analysis

Each model was evaluated on the accuracy of different types of statements:

Category	ChatGPT	Gemini	Claude	Insights
Factual Claims	92.6%	87.3%	91.1%	ChatGPT and Claude excel in factual accuracy
Procedural Recommendations	89.5%	79.3%	82.5%	ChatGPT provides most reliable procedural guidance

Category	ChatGPT	Gemini	Claude	Insights
Inferences/Opinions	85.7%	77.2%	83.1%	ChatGPT shows strongest reasoning capabilities
Overall Accuracy	89.7%	81.9%	85.6%	ChatGPT leads overall but all models show strong knowledge

Performance by Technical Domain

Dataset Analysis (FairFace)

Model	Score	Strengths	Weaknesses
ChatGPT	6.9/10	Data preprocessing, bias recognition	Handling imbalanced data
Gemini	6.5/10	Data preprocessing, bias recognition	Categorical data recommendations
Claude	7.18/10	Data preprocessing, bias recognition	Categorical data recommendations

Key Insight: Claude demonstrates superior performance in dataset analysis tasks, suggesting particular strength in data preprocessing and understanding demographic data.

CNN Architecture/Training

Model	Score	Strengths	Weaknesses
ChatGPT	6.5/10	Model components, regularization	Architecture selection error recognition
Gemini	6.32/10	Model components, regularization	Architecture selection error recognition
Claude	6.43/10	Model components	Architecture selection error recognition

Key Insight: All models perform similarly in CNN architecture tasks, with a universal weakness in recognizing potential errors in architecture selection.

GPU/CPU Optimization

Model	Score	Strengths	Weaknesses
ChatGPT	7.3/10	Practical troubleshooting, performance tuning	Framework-specific guidance
Gemini	6.6/10	Practical troubleshooting, performance tuning	Framework-specific guidance
Claude	7.23/10	Practical troubleshooting, performance tuning	Framework-specific guidance

Key Insight: Claude and ChatGPT both excel in GPU/CPU optimization, with nearly identical performance in this domain.

Prompt Type Effectiveness

Performance by Prompt Type

Prompt Type	ChatGPT	Gemini	Claude	Insights
Chain-of-Thought	7.5/10	6.77/10	6.85/10	ChatGPT excels with step-by-step reasoning
Role-Based	7.4/10	6.77/10	7.15/10	ChatGPT and Claude perform well with expert roles
Few-Shot	6.6/10	6.57/10	6.68/10	Similar performance across models
Zero-Shot	5.8/10	5.98/10	6.13/10	Claude performs best without examples or context

Key Insight: ChatGPT responds exceptionally well to Chain-of-Thought prompts, while Claude performs best with Role-Based prompts. All models perform worst with Zero-Shot prompts, highlighting the importance of structured prompting.

The Calibration-Accuracy Trade-off

An interesting pattern emerges when comparing confidence-performance correlation with technical accuracy:

- **Claude**: High calibration (7.44/10) but moderate accuracy (85.6%)
- **ChatGPT**: High accuracy (89.7%) but lower calibration (6.7/10)
- **Gemini**: Lower scores on both dimensions

This suggests a potential trade-off between raw technical knowledge and self-assessment capability in current LLM designs, where models might excel at one aspect at the expense of the other.

Common Strengths Across Models

Despite their differences, all three models demonstrate several common strengths:

1. **Strong domain knowledge**: All models show excellent understanding of deep learning concepts
2. **Well-structured responses**: Consistently provide organized, step-by-step explanations
3. **Contextual adaptation**: Effectively adjust responses based on technical context
4. **Recognition of complexity**: Demonstrate awareness of nuanced technical challenges
5. **Appropriate confidence levels**: Generally align confidence with accuracy in standard tasks

Universal Areas for Improvement

The evaluation identified several consistent weaknesses across all models:

1. **Limited error recognition**: All models fail to sufficiently acknowledge potential pitfalls or edge cases
2. **Overconfidence in procedural steps**: Recommendations often presented with high certainty without verification
3. **Lack of source qualification**: Rarely reference authoritative sources to validate technical claims
4. **Insufficient alternative exploration**: Limited presentation of multiple solution paths
5. **Incomplete uncertainty communication**: Inadequate expression of confidence levels in ambiguous scenarios

Model-Specific Insights

ChatGPT

- **Technical Leader**: Highest overall accuracy (89.7%) across all categories

- **Prompt Flexibility**: Exceptional performance with structured prompts (7.8/10 prompt sensitivity)
- **Domain Adaptability**: Most consistent performance across technical domains (7.0/10)
- **Balanced Output**: Strong technical knowledge with good but not excellent self-assessment
- **Key Improvement Area**: Error recognition and articulating uncertainty

Claude

- **Calibration Champion**: Best confidence-performance correlation (7.44/10) and calibration (6.69/10)
- **Domain Specialist**: Particularly strong in dataset analysis (7.18/10) and GPU optimization (7.23/10)
- **Role Adaptation**: Performs best with role-based prompts (7.15/10)
- **Well-Balanced**: Good overall weighted self-assessment score (6.78/10)
- **Key Improvement Area**: Domain-specific variance and alternative exploration

Gemini

- **Consistent Performer**: Most balanced performance across metrics, but generally lower scores
- **Task Awareness**: Strongest in recognizing task difficulty (7.06/10)
- **Prompt Type Consistency**: Less variation across prompt types than other models
- **Zero-Shot Resilience**: Relatively strong in zero-shot scenarios compared to other models
- **Key Improvement Area**: Overall technical accuracy and calibration error

Practical Recommendations

For LLM Users

1. **Match model to task**:
   - Choose ChatGPT when technical accuracy is paramount
   - Choose Claude when confidence calibration is critical
   - Choose Gemini when consistent performance across prompt types is needed
2. **Optimize prompting strategy**:
   - Use Chain-of-Thought prompts with ChatGPT
   - Use Role-Based prompts with Claude
   - Structure prompts clearly for all models
3. **Address universal weaknesses**:
   - Explicitly request identification of potential pitfalls or limitations

- Ask for confidence levels on recommendations
- Request multiple solution approaches
- Verify recommendations independently when stakes are high

For LLM Developers

1. **Enhance error recognition**: Develop techniques to better identify and communicate potential issues in recommendations.
2. **Improve calibration mechanisms**: Better align expressed confidence with actual performance, especially for technical recommendations.
3. **Develop uncertainty articulation**: Create more nuanced ways to express confidence levels in ambiguous or evolving technical areas.
4. **Strengthen source qualification**: Incorporate better referencing of authoritative sources when making technical claims.
5. **Improve alternative exploration**: Develop techniques to present multiple viable approaches to complex problems.

Conclusion

This comparative analysis reveals that while all three models demonstrate strong capabilities in AI/ML domains, each has distinct strengths and common limitations. Claude demonstrates superior self-assessment and calibration; ChatGPT shows the strongest technical accuracy and adaptability; and Gemini offers balanced but generally lower performance.

The most critical finding is the universal weakness in error recognition across all models, suggesting an important direction for future LLM development. By enhancing the ability to identify potential pitfalls and limitations, these models could provide more reliable assistance for complex technical tasks.

The analysis also highlights the importance of prompt engineering, with structured approaches like Chain-of-Thought and Role-Based prompts consistently outperforming Zero-Shot prompts across all models. This reinforces the need for users to carefully craft their queries to maximize model performance.

Overall, this comparison provides valuable insights for both LLM users seeking to select the most appropriate model for specific tasks and developers working to enhance LLM self-assessment capabilities in future iterations.