AwardBench: Evaluating AI Performance in Government Contracting

Awarded AI Research Team GovCon AI Initiative

June 20, 2025

Abstract

We present AwardBench, a comprehensive evaluation framework designed to assess AI model performance in government contracting contexts. Our benchmark evaluates models across five critical dimensions: compliance accuracy, proposal quality, workflow effectiveness, retrieval accuracy, and overall efficiency. Through rigorous testing on real-world GovCon scenarios, we demonstrate that specialized AI platforms significantly outperform general-purpose models, achieving up to 94.7% overall accuracy compared to 72.4% for generic solutions. This paper introduces our methodology, presents detailed performance metrics, and provides actionable insights for organizations selecting AI solutions for government contracting workflows.

1 Introduction

The government contracting landscape presents unique challenges for artificial intelligence systems. With over $600 billion in annual federal contracts and stringent compliance requirements under the Federal Acquisition Regulation (FAR) and Defense Federal Acquisition Regulation Supplement (DFARS), AI models must demonstrate exceptional accuracy, domain knowledge, and reliability.

AwardBench addresses the critical need for standardized evaluation metrics in this space. Unlike general-purpose AI benchmarks, our framework specifically targets the complexities of government contracting, including regulatory compliance, proposal generation, and workflow automation.

2 Methodology

Our evaluation methodology encompasses five key performance indicators:

1. **Compliance Accuracy**: Measures the model's ability to correctly interpret FAR/DFARS clauses and identify compliance requirements
2. **Proposal Quality**: Assesses the technical accuracy and win theme alignment of generated proposals
3. **Workflow Effectiveness**: Evaluates end-to-end process automation capabilities
4. **Retrieval Accuracy**: Tests document retrieval precision and context utilization
5. **Overall Efficiency**: Combines speed, cost optimization, and resource utilization metrics

Each model undergoes testing across 1,000+ real-world scenarios derived from actual government solicitations, past performance data, and compliance challenges.

3 Results

Our evaluation reveals significant performance disparities between specialized and general-purpose AI models:

Elite Performance Tier (99th Percentile)

- Awarded AI Platform: 94.7% overall score
- Exceptional compliance accuracy (98%)
- Superior domain knowledge integration

Advanced Capability Tier (85-90th Percentile)

- Claude 3.7 Sonnet: 88.3% overall score
- GPT-4o: 87.2% overall score
- Strong general capabilities with moderate GovCon understanding

Professional Standard Tier (70th Percentile)

- Generic ChatGPT: 72.4% overall score
- Basic functionality without specialized training

4 Discussion

The results demonstrate that domain-specific training and architecture significantly impact AI performance in government contracting contexts. Key findings include:

1. **Specialized Knowledge Matters**: Models trained on GovCon data outperform general models by an average of 22.3%
2. **Compliance is Critical**: The top-performing model achieved 98% compliance accuracy, compared to 65% for generic solutions
3. **Integration Capabilities**: Elite tier models demonstrated superior workflow integration and automation potential

These findings suggest that organizations should prioritize specialized AI platforms for mission-critical GovCon applications.

5 Conclusion

AwardBench establishes a new standard for evaluating AI performance in government contracting. Our framework provides organizations with objective metrics to assess and select AI solutions that meet the unique demands of federal procurement.

As the GovCon landscape continues to evolve, we anticipate regular updates to our benchmark methodology, incorporating new regulatory requirements and emerging use cases. We invite the community to contribute to this ongoing effort to advance AI capabilities in government contracting.