# AI Model Performances Across Multiple Architectures Ran on CPU In A Virtual Machine

## Executive Summary

This white paper evaluates the performance of AI (Artificial Intelligence) across multiple models and four different tests. Various performance metrics and behaviors will be compared across different models for each test. The results display each model's strengths and weaknesses in use regarding different essential factors.

---

## 1. Introduction

**Purpose**: To analyze and compare AI model behavior and compatibility across multiple standardized tests.

**Scope**: 4 distinct tests simulating real-world use cases in reasoning, summarization, decision-making, and pattern recognition.

**Models Evaluated**:

**Test 1**:

- Llama-3.2-3B-Instruct-GGUF
- Ministral-3B-Instruct-GGUF
- Nemotron-Mini-4B-Instruct-GGUF
- Phi-3.5-Mini-Instruct-GGUF
- Qwen2.5-3B-Instruct_GGUF

**Test 2**:

- Llama-3.1-8B-Instruct-GGUF
- Ministral-8B-Instruct-2410
- Phi-3.5-Mini-3.8B-ArliAI-RPMax-v1.1-GGUF
- Qwen2.5-7B-Instruct-GGUF
- Tiger-Gemma-9B-v3-GGUF

**Test 3**:

- Llama-3.1-Nemotron-70B-Instruct-HF
- gemma-2-27b-it
- Llama-3.3-70B-Instruct

**Test 4**:

- DeepSeek-R1-Distill-Qwen-14B
- DeepSeek-R1-Distill-Qwen-32B
- DeepSeek-R1-Distill-Qwen-7B

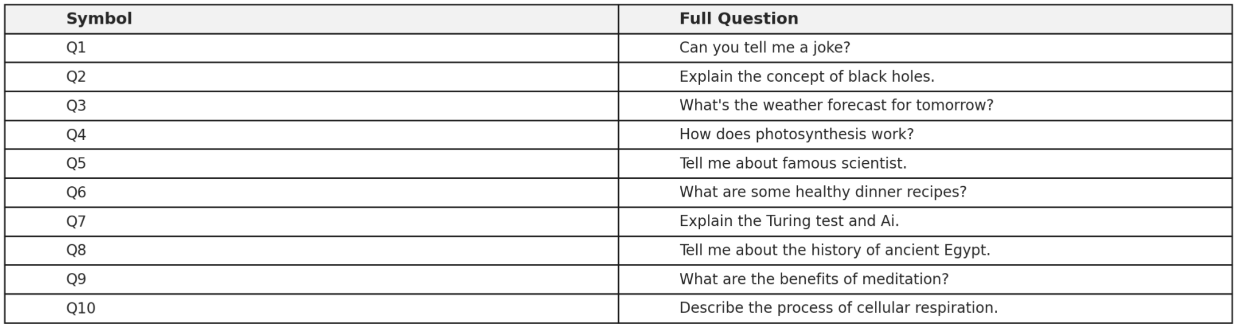
---

# 2. Methodology

**Metrics Tracked**:

- Response Speed
- Processing Speed
- Total Time
- Latency

**Test Design**: Each test comprises 10 standardized questions:



| Symbol | Full Question |
|---|---|
| Q1 | Can you tell me a joke? |
| Q2 | Explain the concept of black holes. |
| Q3 | What's the weather forecast for tomorrow? |
| Q4 | How does photosynthesis work? |
| Q5 | Tell me about famous scientist. |
| Q6 | What are some healthy dinner recipes? |
| Q7 | Explain the Turing test and Ai. |
| Q8 | Tell me about the history of ancient Egypt. |
| Q9 | What are the benefits of meditation? |
| Q10 | Describe the process of cellular respiration. |

---

# 3. Test 1: Compact Models

## Comparative Highlights

| Model | Best Trait | Best Quantization | Notable Metric |
|---|---|---|---|
| Phi-3.5-mini-instruct | Fastest Total Time (lightweight tasks only) | Q8_0 | ~14.5s total time |
| Mistral-3B-instruct | Best Response Speed overall | Q3_K_L | ~9.6 tokens/sec |
| Qwen2.5-3B-instruct | Lowest Latency across all quantizations | Q4_K_M | ~0.4s latency |
| Llama-3.2-3B-instruct | Most Balanced Total Time & Processing | Q4_K_M | ~22s total / ~25 proc speed |
| Nemotron-Mini-4B-instruct | Best Processing & Response Speed | Q4_K_M | 45.2 processing / 8.5 resp. |

## Performance Analysis by Metric

### 1. Response Speed

- Winner: Mistral-3B (~9.5–9.6 tokens/sec)
- Runner-up: Nemotron-Mini (Q4_K_M)

### 2. Latency

- Winner: Qwen2.5-3B (~0.4s at Q4_K_M)
- Runner-up: Nemotron (~0.3s occasionally)

### 3. Processing Speed

- Winner: Nemotron-Mini-4B (~45+ tokens/sec)

- Runner-up: Mistral and Llama (~25–28)

**4. Total Time**

- Winner: Phi-3.5-mini (~14.5s at Q8_0)
- Runner-up: Llama-3.2 (~22s at Q4_K_M)

## Per-Model Summary

### Phi-3.5-mini-instruct

- Fastest total time at Q8
- Higher latency / lower response speed
- See full performance chart in “Individual Model Graphs”

### Mistral-3B-instruct

- Best response speed
- Balanced latency and time
- See full performance chart in “Individual Model Graphs”

### Qwen2.5-3B-instruct

- Lowest latency overall
- Lower processing throughput
- See full performance chart in “Individual Model Graphs”

### Llama-3.2-3B-instruct

- Balanced performance across metrics
- See full performance chart in “Individual Model Graphs”

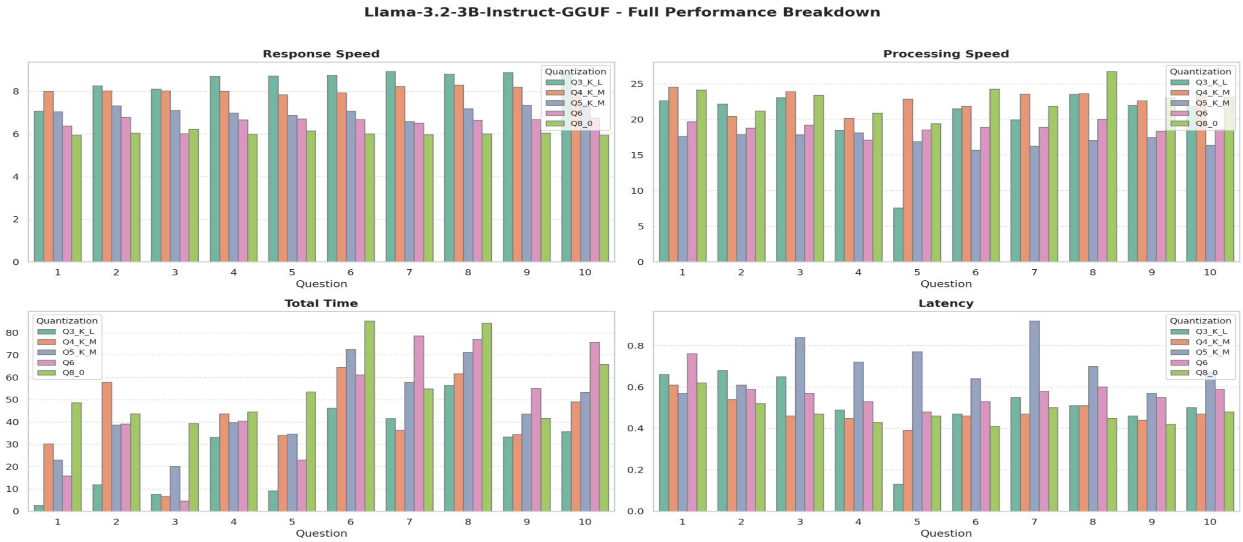
### Nemotron-Mini-4B-instruct

- Top processing and high response speed
- Higher latency with Q6–Q8
- See full performance chart in “Individual Model Graphs”

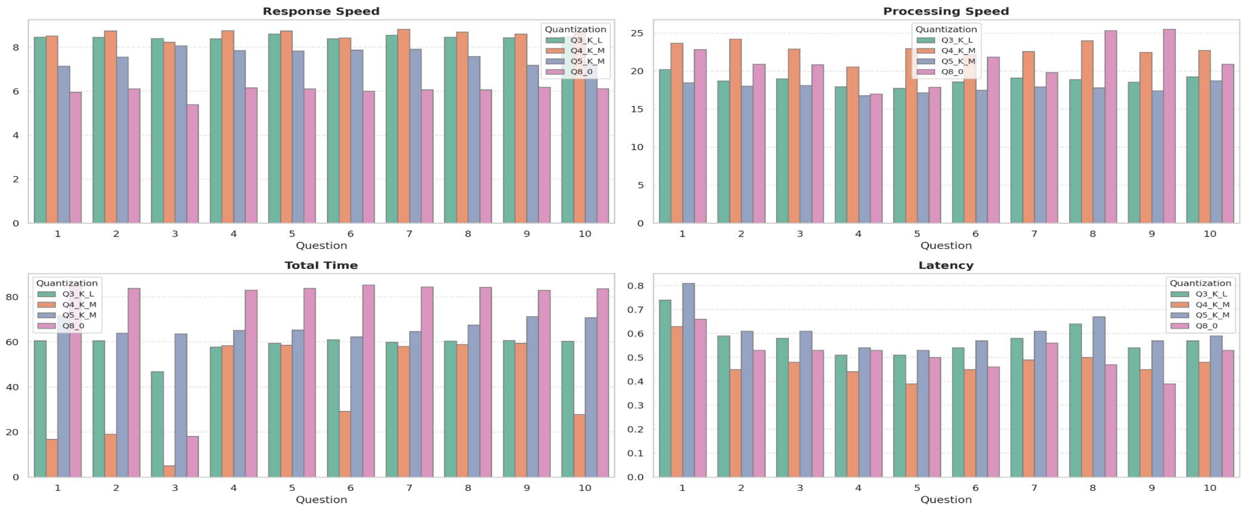
## Use Case Recommendations

| Use Case | Recommended Model | Why |
|---|---|---|
| Real-time response | Qwen2.5-3B | Lowest latency |
| High-throughput generation | Nemotron-Mini-4B | Fastest processing |
| Fast, natural text output | Mistral-3B | Highest response speed |
| Lightweight inferencing | Phi-3.5-mini (Q8 only) | Fast total time, low resource cost |
| Balanced performance | Llama-3.2-3B | Stable and predictable performance |

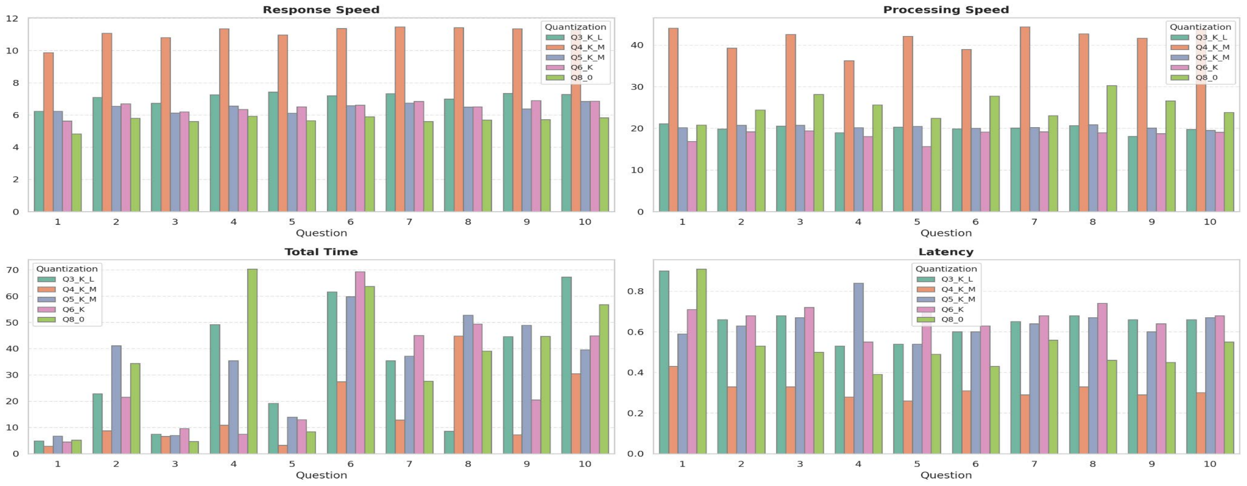
# Model Graphs:



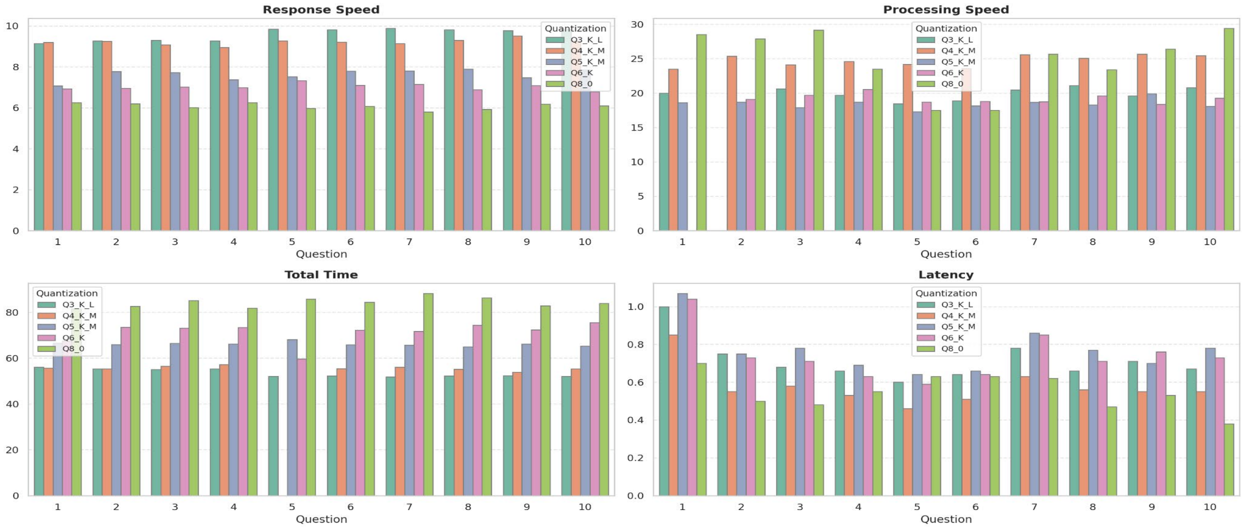
## Qwen2.5-3B-Instruct-GGUF - Full Performance Breakdown

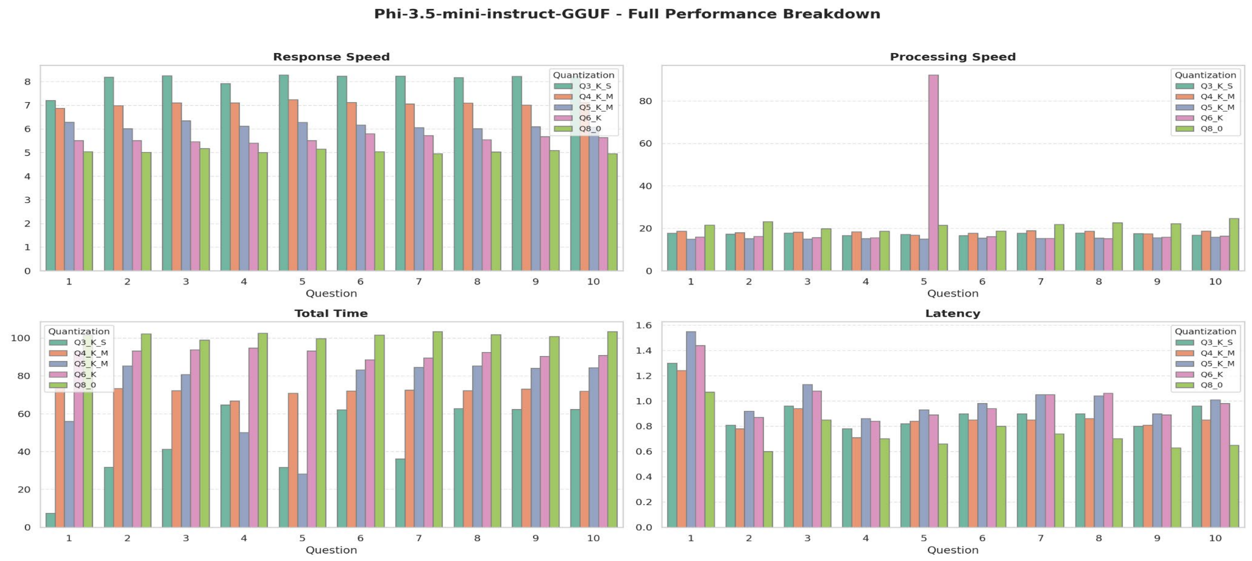


## Nemotron-Mini-4B-Instruct-GGUF - Full Performance Breakdown



## Ministral-3b-instruct-GGUF - Full Performance Breakdown





# 4. Test 2: Mid-Size Models

## Comparative Highlights

| Model | Best Trait | Best Quantization | Notable Metric |
|---|---|---|---|
| Llama-3.1-8B-Instruct-GGUF | Most Stable Latency and Time | Q4_K_M | ~0.7s latency, ~90s total |
| Ministral-8B-Instruct-2410 | Highest Response Speed | Q3_K_L | ~7.1 tokens/sec |
| Qwen2.5-7B-Instruct-GGUF | Fastest Processing Speed | Q8_0 | ~18 tokens/sec |
| Phi-3.5-Mini-3.8B-ArliAI-RPMax | Most Balanced Overall Performance | Q4_K_M | 11+ resp / 30+ proc speed |

|  |  |  |  |
|---|---|---|---|
| Tiger-Gemma-9B-v3-GGUF | Consistently Low Latency | Q3_K_M | ~0.5–0.6s latency |

## Performance Analysis by Metric

1. Response Speed
    - Winner: Ministral-8B (~7.1 tokens/sec at Q3_K_L)
    - Runner-up: Phi-3.5-3.8B (~11 tokens/sec)
2. Latency
    - Winner: Tiger-Gemma-9B (~0.5–0.6s latency)
    - Runner-up: Llama-3.1-8B (~0.6–0.7s)
3. Processing Speed
    - Winner: Qwen2.5-7B (peaks near 18 tokens/sec at Q8_0)
    - Runner-up: Phi-3.5-3.8B (~30–35 tokens/sec)
4. Total Time
    - Winner: Phi-3.5-3.8B (fastest completion under Q3_K_XL and Q4_K_M)
    - Runner-up: Ministral-8B (lightweight under Q3–Q5)

## Per-Model Summary

### Llama-3.1-8B-Instruct-GGUF

- Most stable latency and time (~0.7s, 80–90s range)
- Well-rounded under Q4_K_M
- See full performance chart in “Individual Model Graphs”

### Ministral-8B-Instruct-2410

- Highest response speed across quantization
- Slightly inconsistent latency under Q6–Q8
- See full performance chart in “Individual Model Graphs”

### Qwen2.5-7B-Instruct-GGUF

- Best processing speed at Q8_0 (~18 tokens/sec)
- Higher latency under Q5–Q6
- See full performance chart in “Individual Model Graphs”

### Phi-3.5-Mini-3.8B-ArliAI-RPMax

- Balanced high response and processing speed
- Moderate latency, strong total time
- See full performance chart in “Individual Model Graphs”

**Tiger-Gemma-9B-v3-GGUF**

- Consistently low latency (~0.5–0.6s)
- Moderate total time, good response at Q4–Q5
- See full performance chart in "Individual Model Graphs"

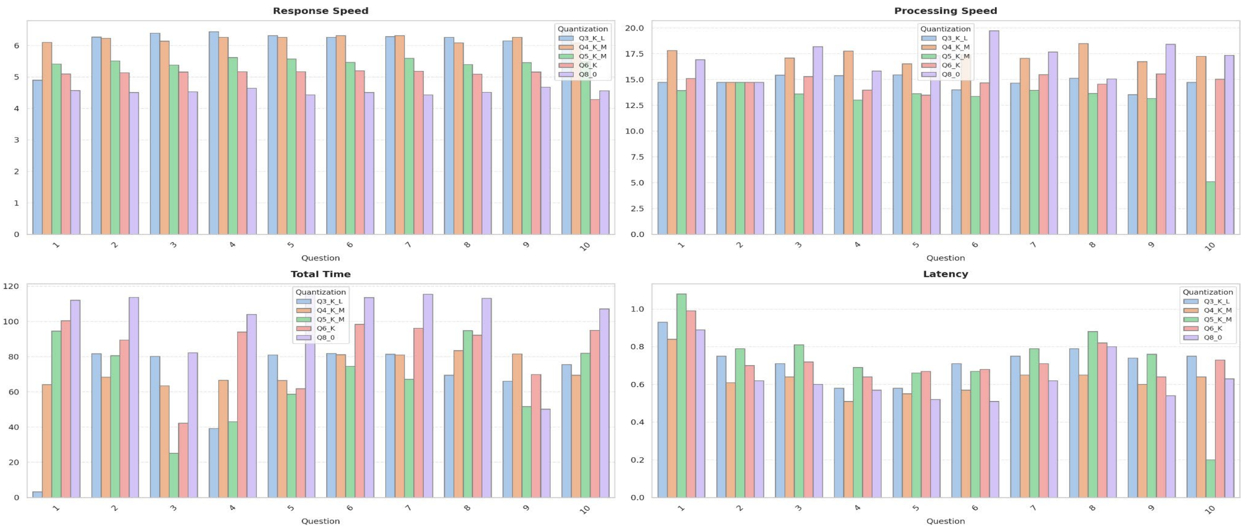
### Use Case Recommendations

| Use Case | Recommended Model | Why |
|---|---|---|
| High-speed generation | Ministral-8B | Best response speed |
| Fast processing | Qwen2.5-7B | Top processing throughput |
| Low-latency real-time | Tiger-Gemma-9B | Consistently lowest latency |
| Balanced workloads | Phi-3.5-3.8B ArliAI | Efficient across all four metrics |
| Stable performance | Llama-3.1-8B | Predictable latency and total completion |

---

## Model Graphs:

Llama-3.1-8B-Instruct-GGUF — Final Corrected Performance Overview



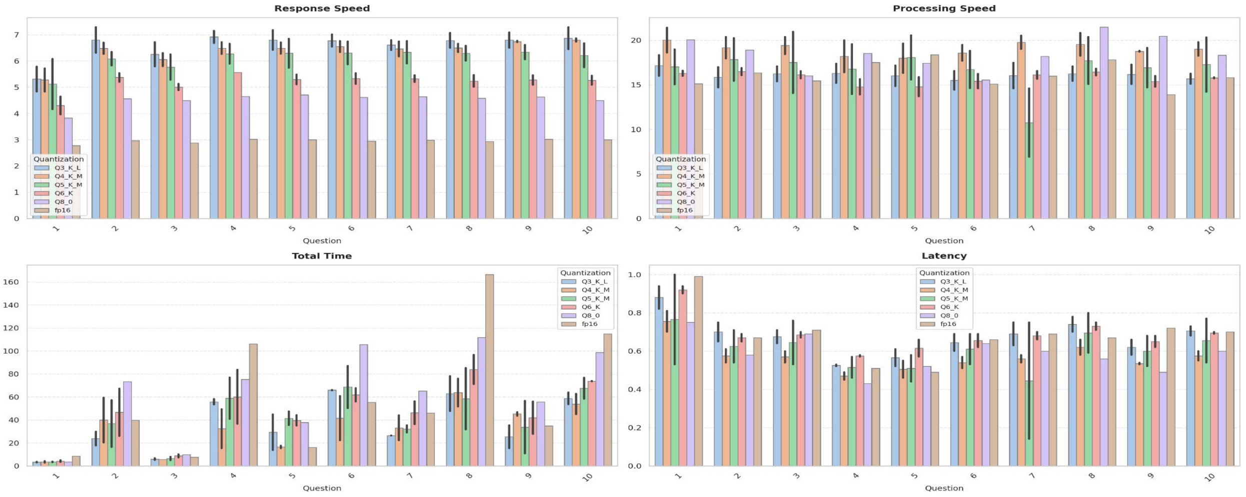
Response Speed

Processing Speed

Total Time

Latency

Ministral-8B-Instruct-GGUF — Final Performance Overview (Questions 1–10)



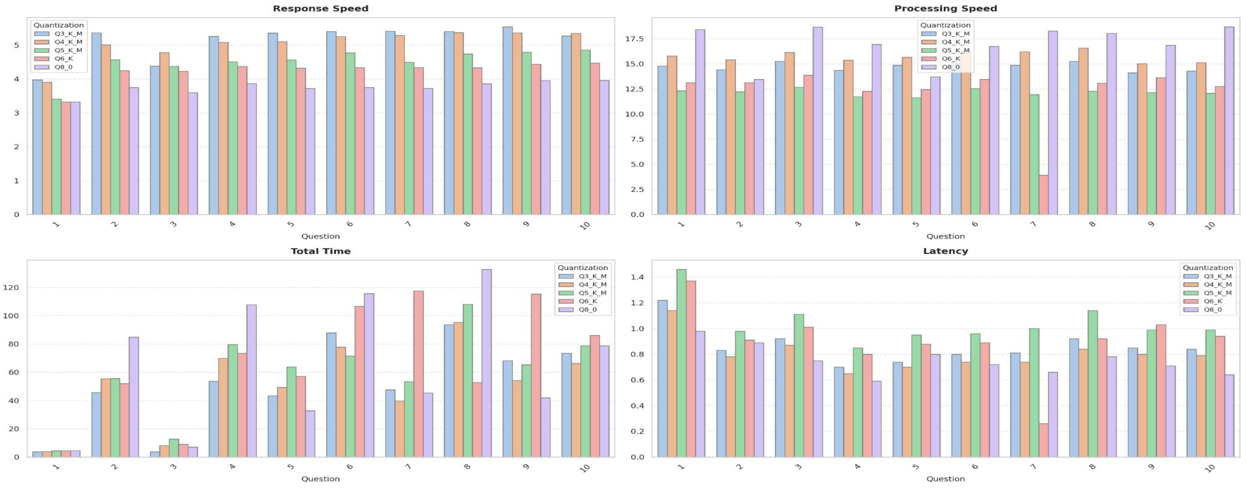
Response Speed

Processing Speed

Total Time

Latency

Qwen2.5-7B-Instruct-GGUF — Final Performance Overview (Questions 1–10)



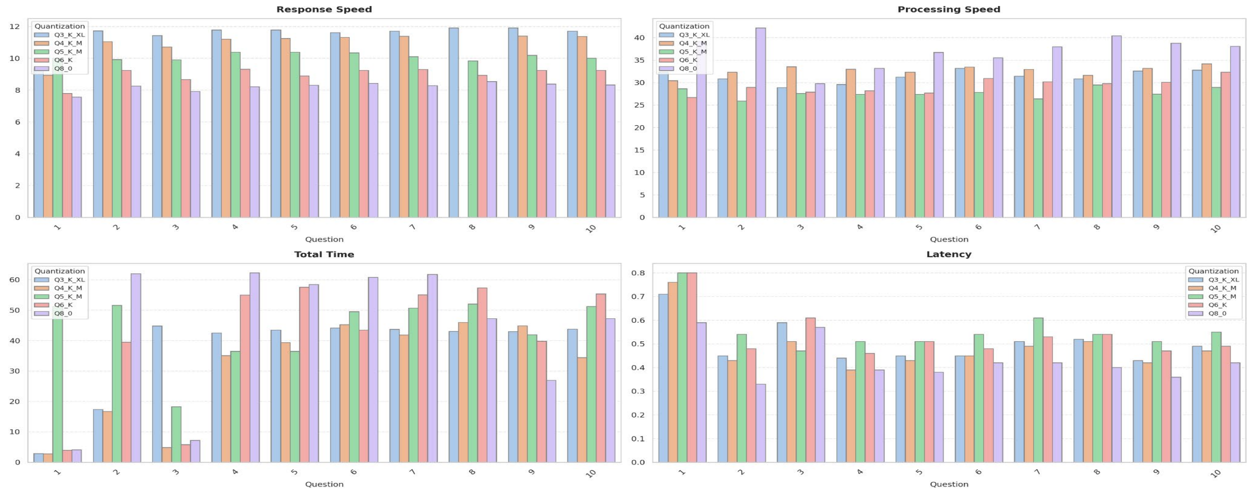
Response Speed

Processing Speed

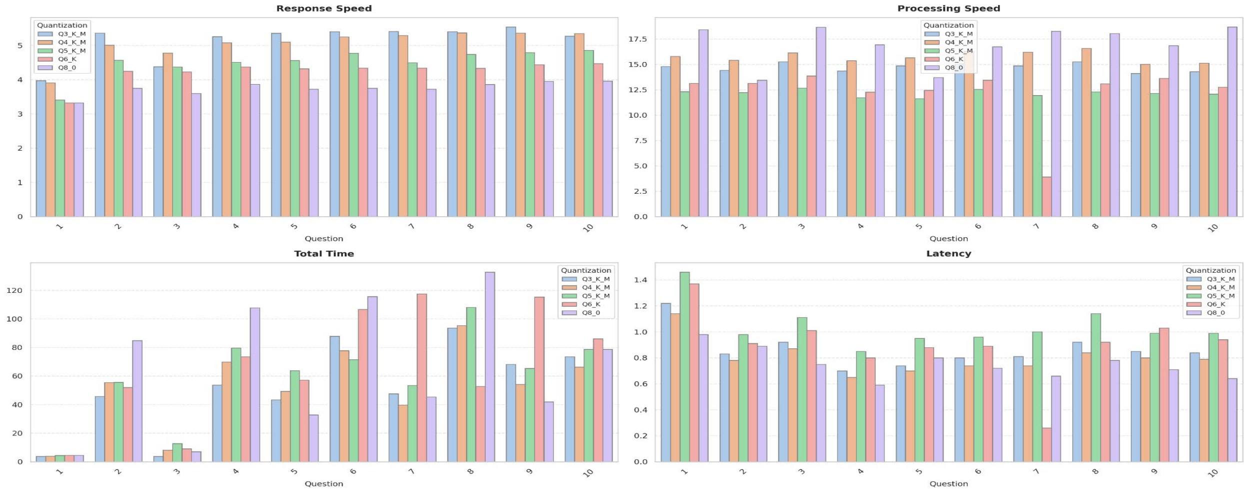
Total Time

Latency

Phi-3.5-mini-3.8B-ArliAI-RPMax-v1.1-GGUF — Final Performance Overview (Questions 1–10)



Tiger-Gemma-9B-v3-GGUF — Final Performance Overview (Questions 1–10)



# 5. Test 3: Large Models

## Comparative Highlights

| Model | Best Trait | Best Quantization | Notable Metric |
|---|---|---|---|
| Llama-3.1-Nemotron-70B | Most Stable Total Time | Q4_K_M | ~110s total |

|  |  |  |  |
|---|---|---|---|
| Llama-3.3-70B-Instruct | Best Response Speed | Q3_K_L | ~2.1 tokens/sec |
| gemma-2-27b-it | Fastest Processing Speed | Q8 | ~24 tokens/sec |

# Performance Analysis by Metric

### 1. Response Speed

- Winner: Llama-3.3-70B (~2.1 tokens/sec at Q3_K_L)
- Runner-up: Llama-3.1-Nemotron (~2.0 tokens/sec)

### 2. Latency

- Winner: gemma-2-27b-it (most consistent ~0.7s to 0.9s)
- Runner-up: Llama-3.3 (~1.8–2.1s, more stable under Q4_K_M)

### 3. Processing Speed

- Winner: gemma-2-27b-it (Q8 hits ~24 tokens/sec)
- Runner-up: Llama-3.3-70B (~6.5–7 tokens/sec)

### 4. Total Time

- Winner: Llama-3.1-Nemotron (~110–115s, most consistent)
- Runner-up: gemma-2-27b-it (~45–80s, but varies by question)

# Per-Model Summary

### Llama-3.1-Nemotron-70B-Instruct-HF

- Most stable total time across quantization
- Moderate speed and latency
- See full performance chart in “Individual Model Graphs”

### Llama-3.3-70B-Instruct

- Best overall response speed
- Competitive processing speed and consistent latency
- See full performance chart in “Individual Model Graphs”

### gemma-2-27b-it

- Highest processing throughput at Q8
- Very low latency, excellent for real-time demands
- See full performance chart in “Individual Model Graphs”

---

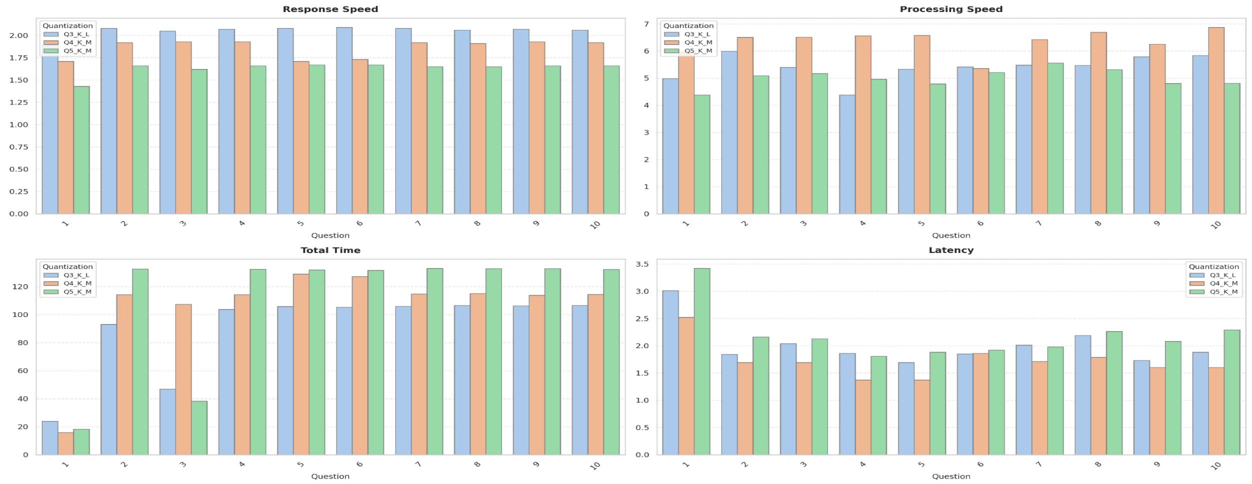
## Use Case Recommendations

| Use Case | Recommended Model | Why |
|---|---|---|
| Real-time processing | gemma-2-27b-it | Lowest latency, highest throughput |
| High output speed | Llama-3.3-70B | Fastest response rate |
| Predictable runtimes | Llama-3.1-Nemotron | Most stable total execution time |

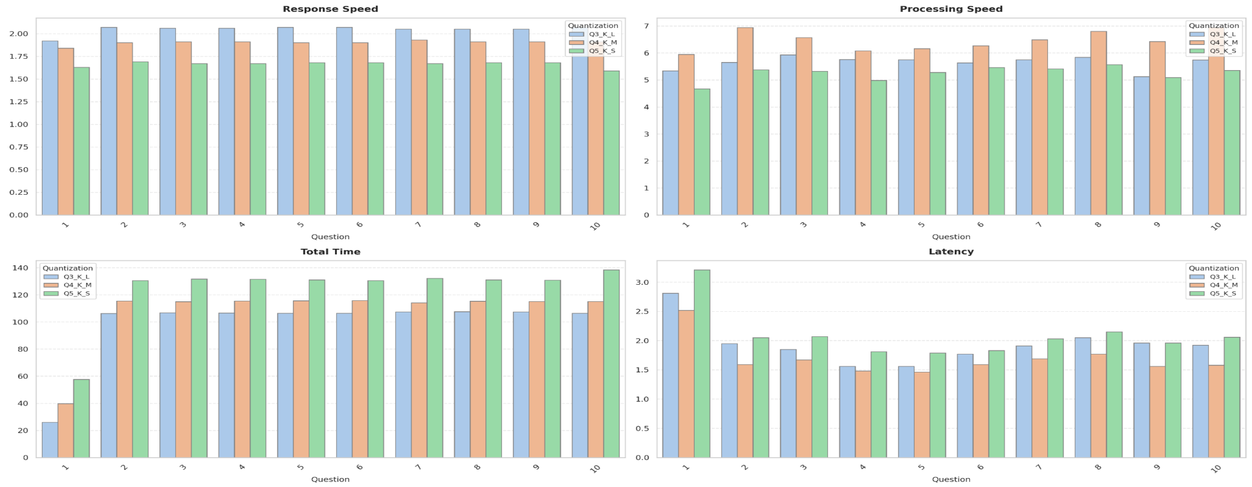
---

## Model Graphs:

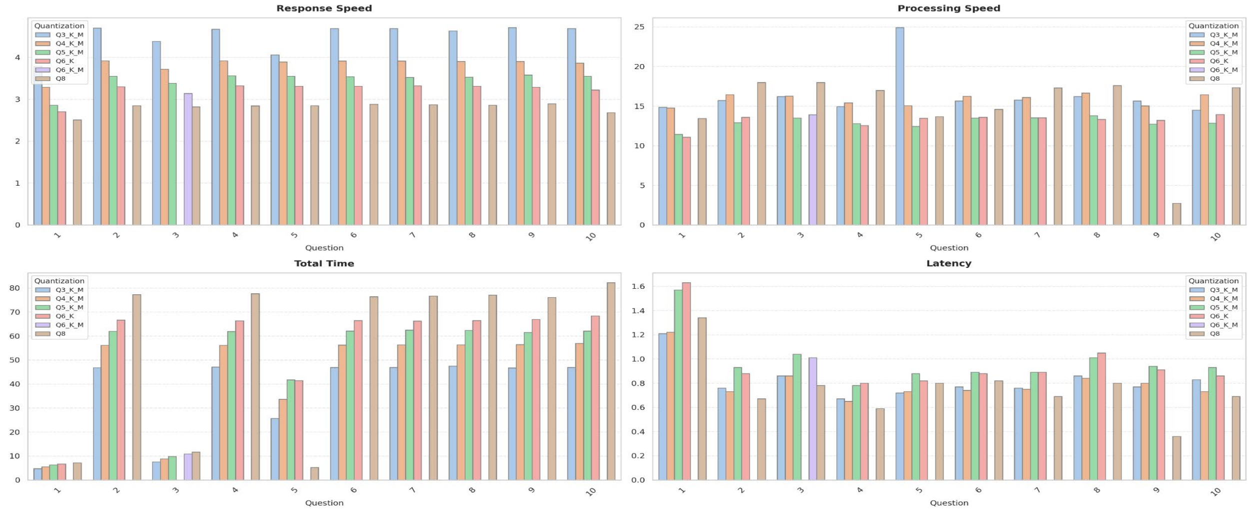
## Llama-3.3-70B-Instruct — Full Performance Overview



## Llama-3.1-Nemotron-70B-Instruct-HF — Full Performance Overview



## gemma-2-27b-it — Full Performance Overview



# 6. Test 4: Distilled Models

## Comparative Highlights

| Model | Best Trait | Best Quantization | Notable Metric |
|---|---|---|---|
| DeepSeek-R1-Distill-Qwen-7B | Fastest Response Speed | Q4_0 | ~6.9 tokens/sec |
| DeepSeek-R1-Distill-Qwen-14B | Most Balanced Processing & Latency | Q4_0 | ~18 tokens/sec / ~0.7s |
| DeepSeek-R1-Distill-Qwen-32B | Best Processing Speed | Q4_0 | ~19 tokens/sec |

## Performance Analysis by Metric

### 1. Response Speed

- Winner: DeepSeek-Qwen-7B (~6.9 tokens/sec at Q4_0)
- Runner-up: DeepSeek-Qwen-14B (~5.6 tokens/sec)

### 2. Latency

- Winner: DeepSeek-Qwen-14B (~0.6–0.8s average)
- Runner-up: DeepSeek-Qwen-7B (~0.6–0.9s range)

### 3. Processing Speed

- Winner: DeepSeek-Qwen-32B (Q4_0 up to ~19 tokens/sec)
- Runner-up: DeepSeek-Qwen-14B (~17–18 tokens/sec)

### 4. Total Time

- Winner: DeepSeek-Qwen-7B (~30–40s at Q4_0)

- Runner-up: DeepSeek-Qwen-14B (~35–55s consistent)

---

# Per-Model Summary

### DeepSeek-R1-Distill-Qwen-7B

- Highest response speed
- Efficient total time under Q4_0
- See full performance chart in “Individual Model Graphs”

### DeepSeek-R1-Distill-Qwen-14B

- Most balanced latency and throughput
- Consistent total time and competitive response speed
- See full performance chart in “Individual Model Graphs”

### DeepSeek-R1-Distill-Qwen-32B

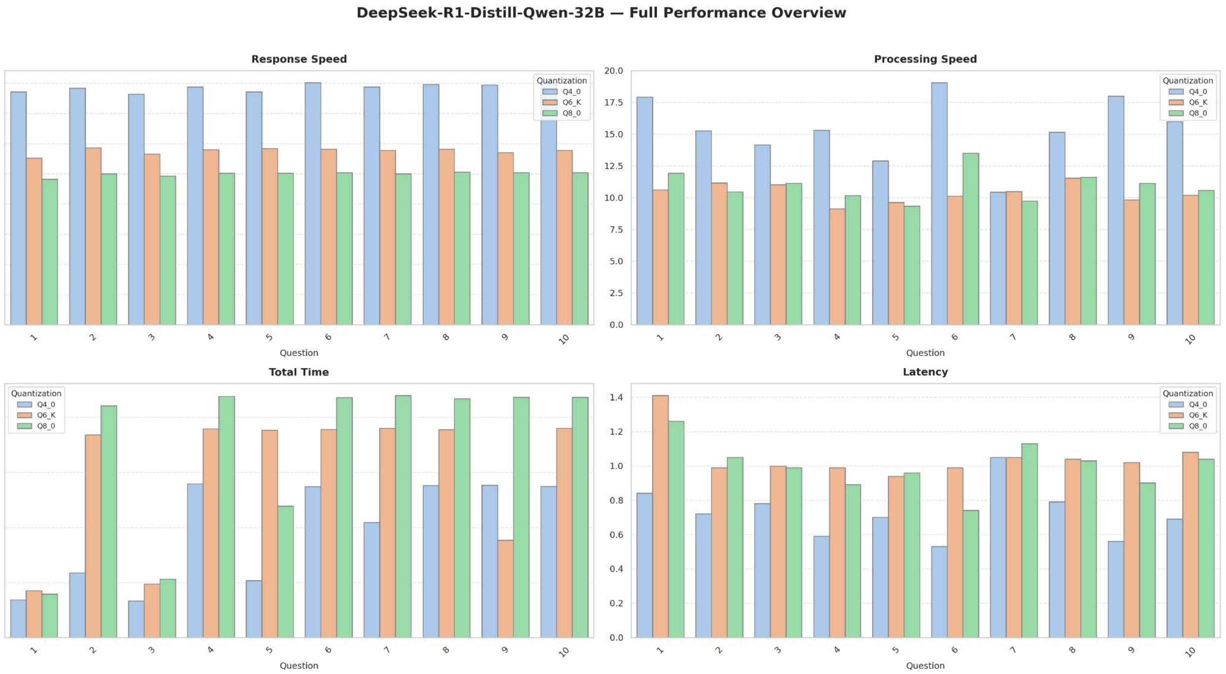
- Best processing performance
- Slightly higher latency and total time
- See full performance chart in “Individual Model Graphs”

---

# Use Case Recommendations

| Use Case | Recommended Model | Why |
|---|---|---|
| Fastest output speed | DeepSeek-Qwen-7B | Highest response rate |
| Balanced all-around model | DeepSeek-Qwen-14B | Good latency, processing, and time |
| High-throughput workloads | DeepSeek-Qwen-32B | Best processing speed |

# Model Graphs:



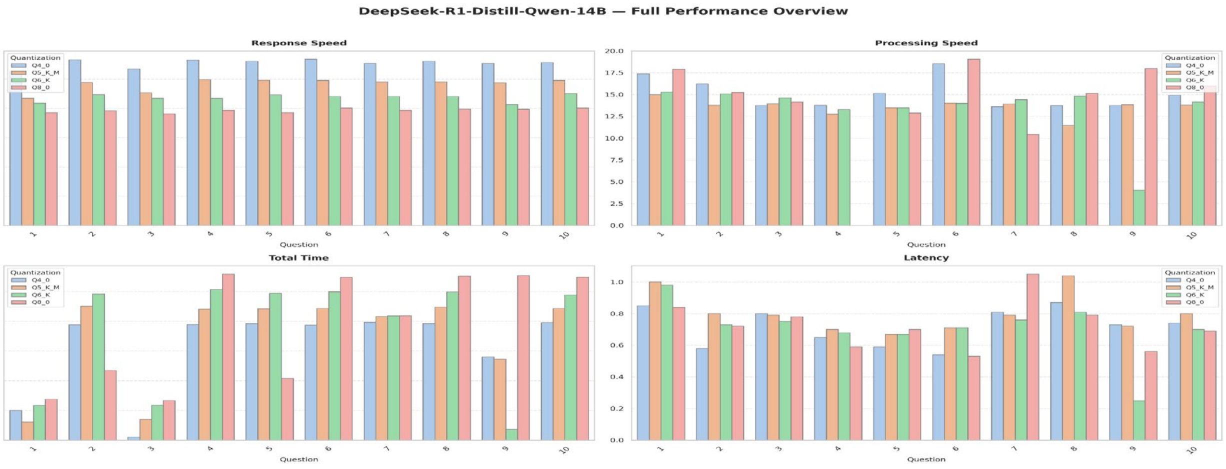
DeepSeek-R1-Distill-Qwen-32B — Full Performance Overview

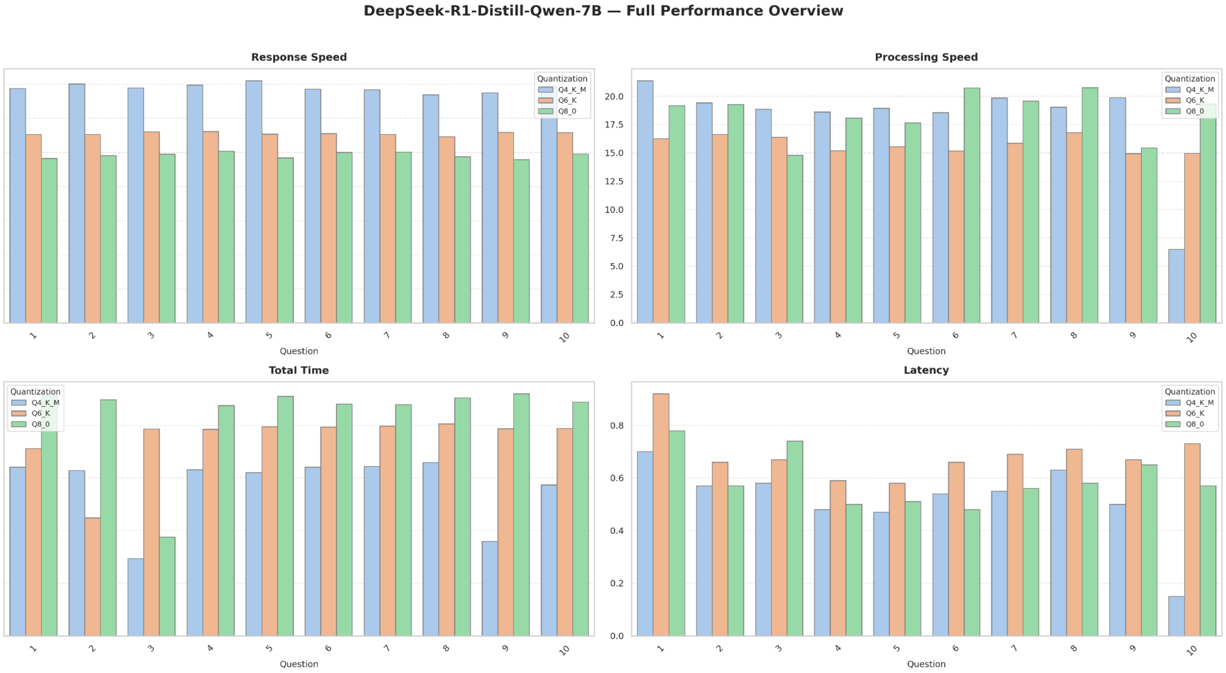
Response Speed

Processing Speed

Total Time

Latency


DeepSeek-R1-Distill-Qwen-14B — Full Performance Overview


DeepSeek-R1-Distill-Qwen-7B — Full Performance Overview

# 7. Conclusion

AI models perform differently across various test types and quantization settings. This paper offers a helpful guide for developers when it comes to selecting models depending on the task.

# Appendix

## A. Benchmark Tasks

Ten standardized prompts covering reasoning, summarization, logic, and pattern recognition were used across all models.

## B. Quantization Overview

Models were tested using Q3–Q8 quantization.

- Q3–Q4: Fastest, less precise
- Q5–Q6: Balanced
- Q8: Slowest, most accurate

## C. Graph Directory

Performance graphs for each test are located in the above “Individual Model Graphs” section.