Exploring Open-Source LLMs with Ollama

CS 5393 - Midterm Exam

https://github.com/cgkarahalios

April 10, 2025

TinyLlama

- Small-sized model (1.1B parameters)
- Optimized for efficiency and speed on limited resources

Mistral

- Medium-sized model (7B parameters)
- Balanced performance and resource requirements

LLaMA3.1

- Larger, more capable model (8B parameters)
- Advanced reasoning and instruction following

Experimental Setup & Methodology

- Testing Categories: General Q&A, Text Summarization, Code Generation, Creative Writing
- Difficulty Levels: Easy, Medium, and Hard prompts
- Prompt Engineering: Basic, Few-shot, Chain-of-thought, Role-based, Self-consistency
- Metrics: Response time, Response length, Output quality assessment

Performance Overview - Response Times

- TinyLlama: 5.47s (Fastest: Summarization Easy - 0.96s, Slowest: Creative Easy - 14.56s)
- Mistral: 19.85s (Fastest: Summarization Easy - 6.29s, Slowest: Creative Hard - 37.87s)
- LLaMA3.1: 84.69s (Fastest: Summarization Easy - 5.24s, Slowest: Code Medium - 233.58s)

Response Time by Difficulty and Model

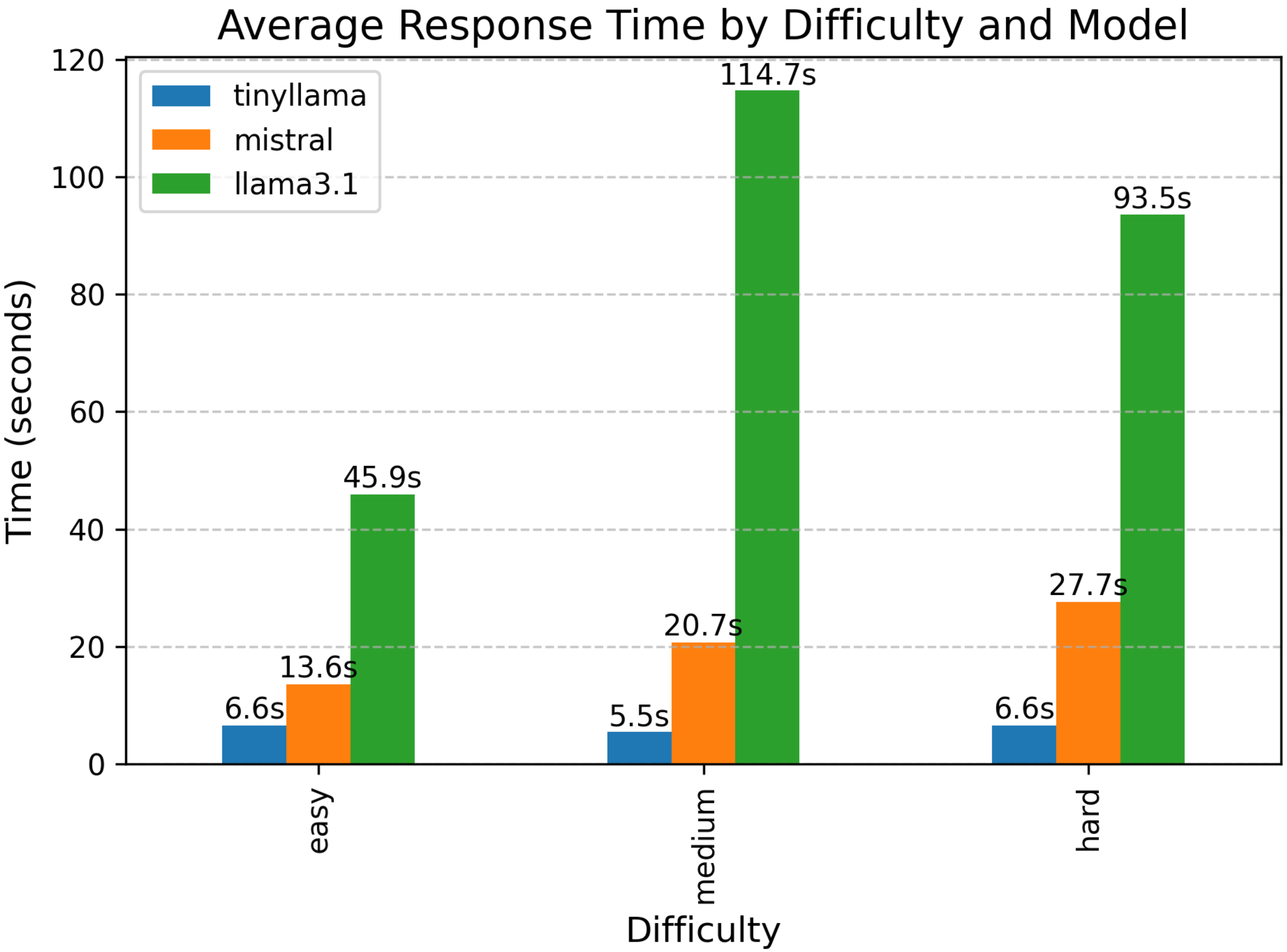

Fig 1: Average Response Time by Difficulty and Model

Key Observations

- TinyLlama significantly faster (3–15x) but with quality tradeoffs
- LLaMA3.1 has highest output quality but longest response times
- Response time increases with task complexity across all models
- Creative writing and code generation are the most time-intensive

Llama3.1 Response Time Heatmap

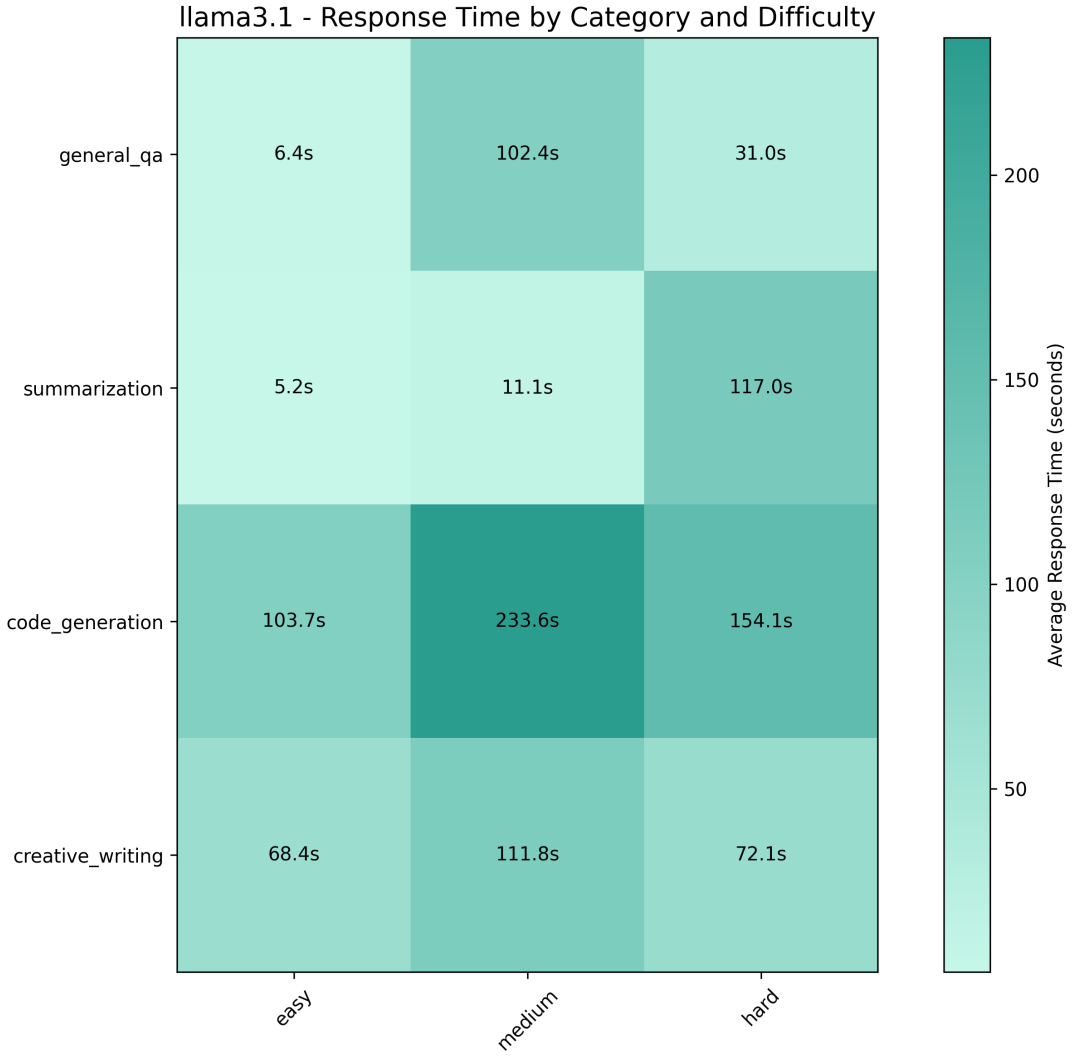

Fig 2: Llama3.1 - Response Time by Category and Difficulty

Mistral Response Time Heatmap

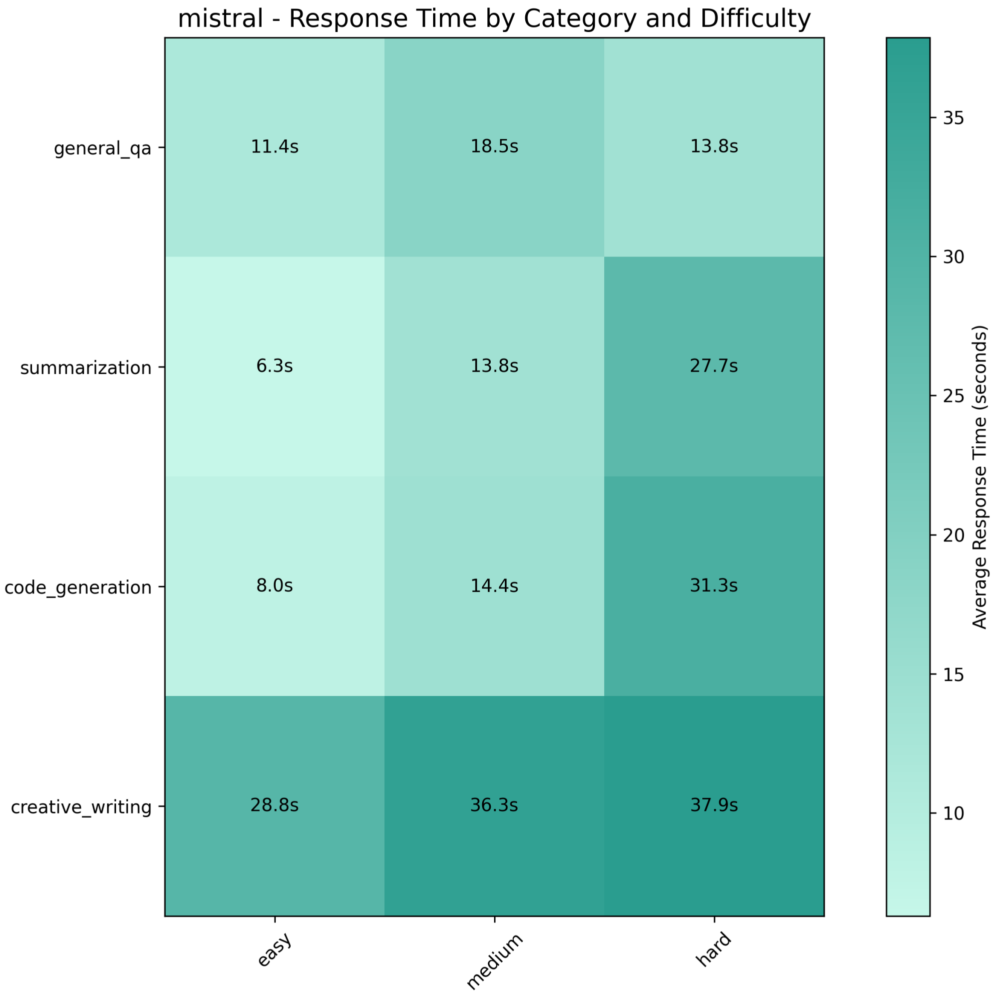

Fig 3: Mistral - Response Time by Category and Difficulty

Key Insights

- Llama3.1 shows extreme variance with code generation tasks (103–233s)
- Mistral maintains more consistent performance across categories
- TinyLlama demonstrates the most uniform response times
- All models show performance degradation with increased difficulty

Response Length Comparison

- TinyLlama: 347 tokens (Shortest: QA Easy - 11, Longest: Creative Easy - 842)
- Mistral: 231 tokens (Shortest: QA Easy - 6, Longest: Creative Hard - 493)
- LLaMA3.1: 361 tokens (Shortest: QA Easy - 6, Longest: Creative Medium - 621)

Response Length by Category

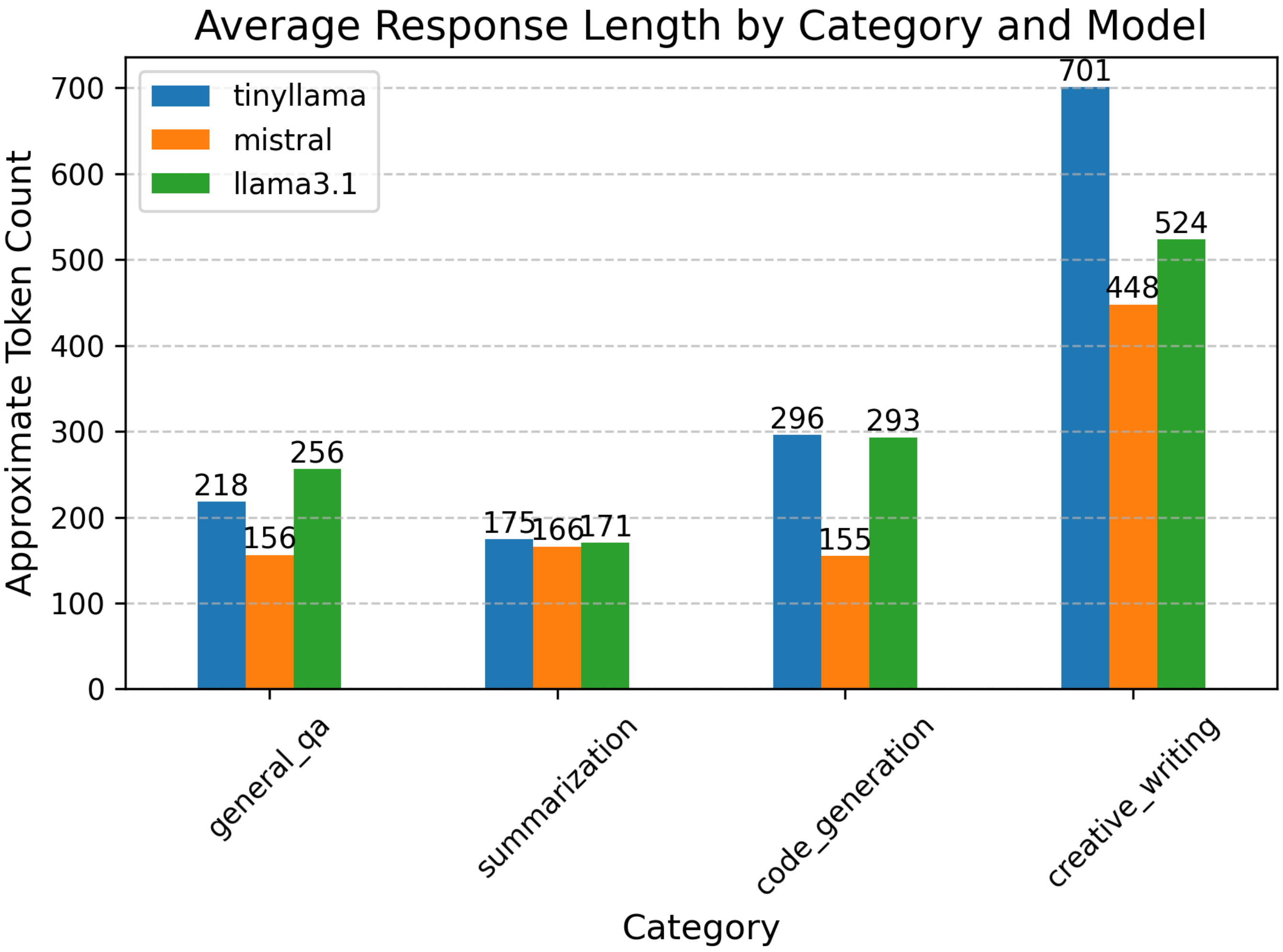

Fig 4: Average Response Lengths by Category and Model

Key Insights

- Response length correlates with difficulty level across all models
- LLaMA3.1 generates the longest responses on average
- Creative writing tasks consistently produced the longest outputs
- Simple general knowledge questions yielded the shortest responses