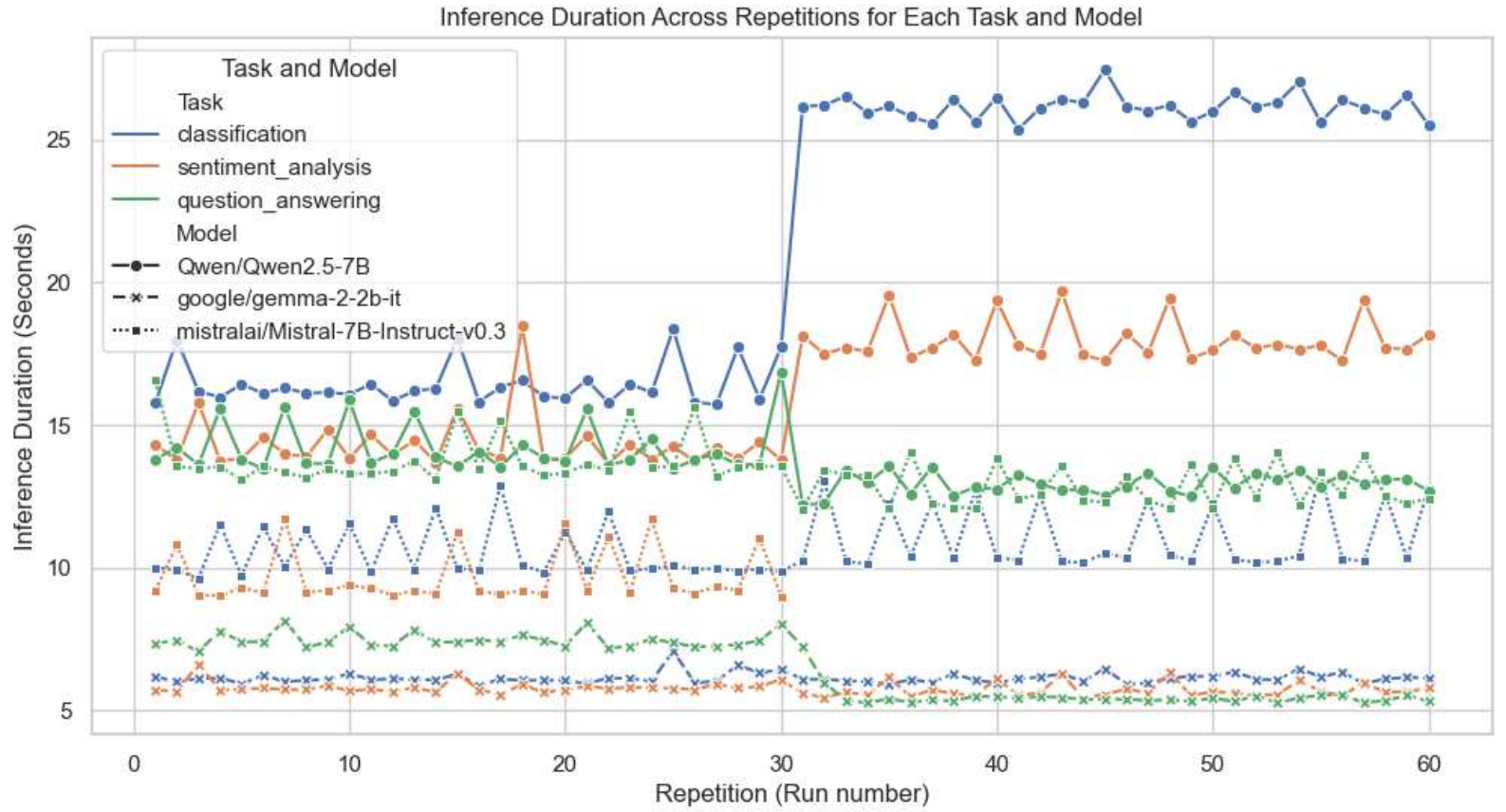
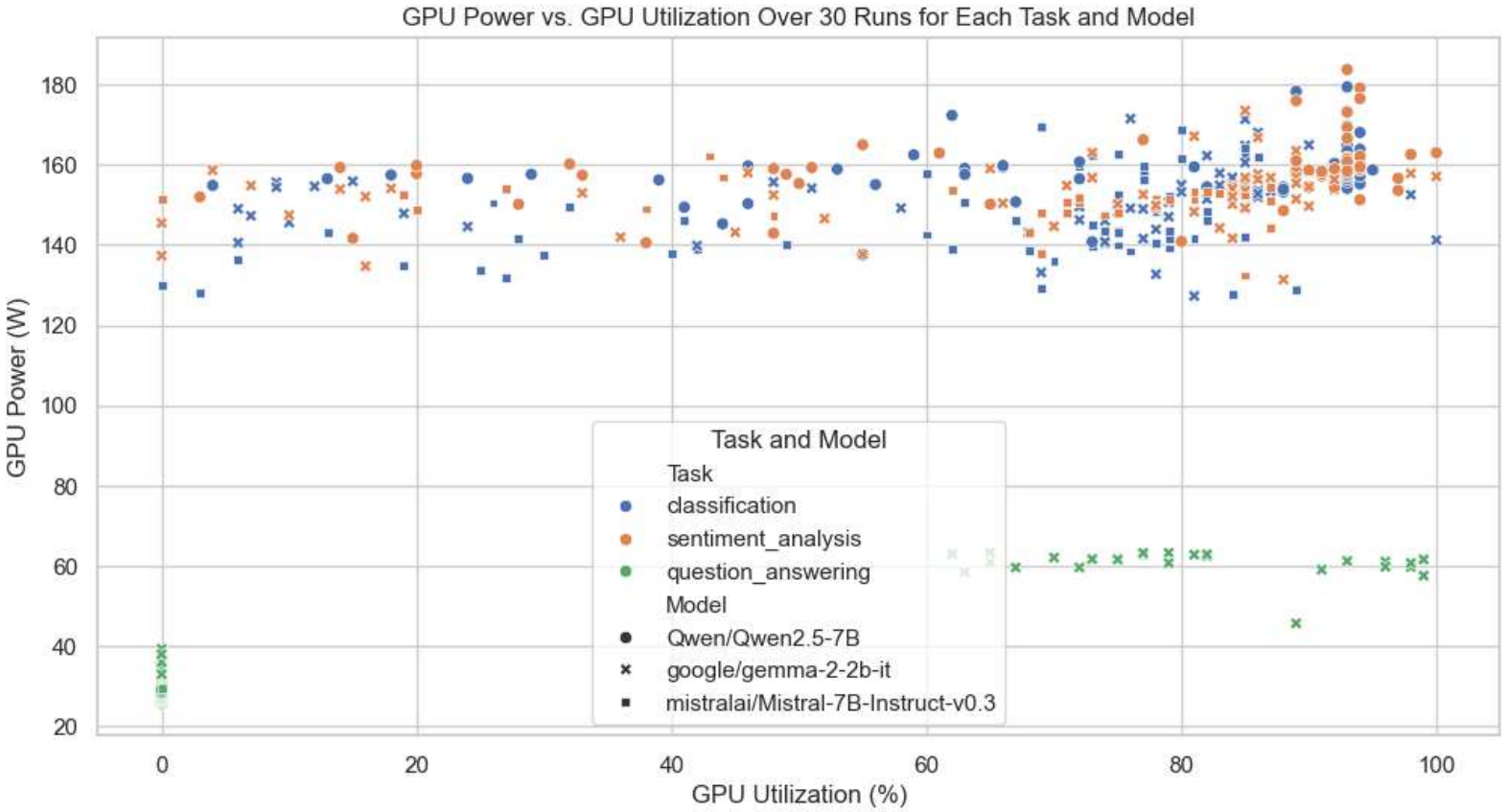
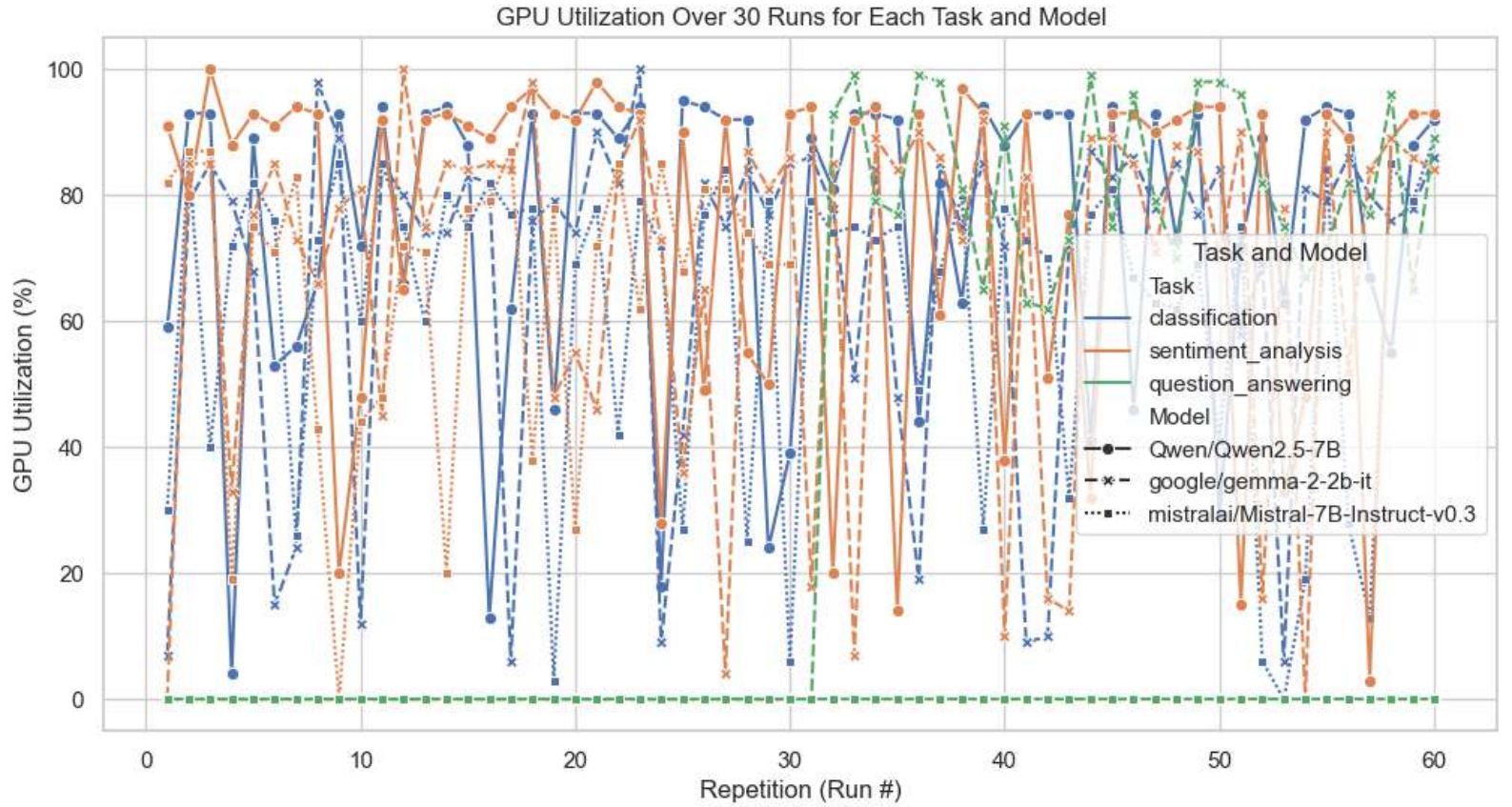
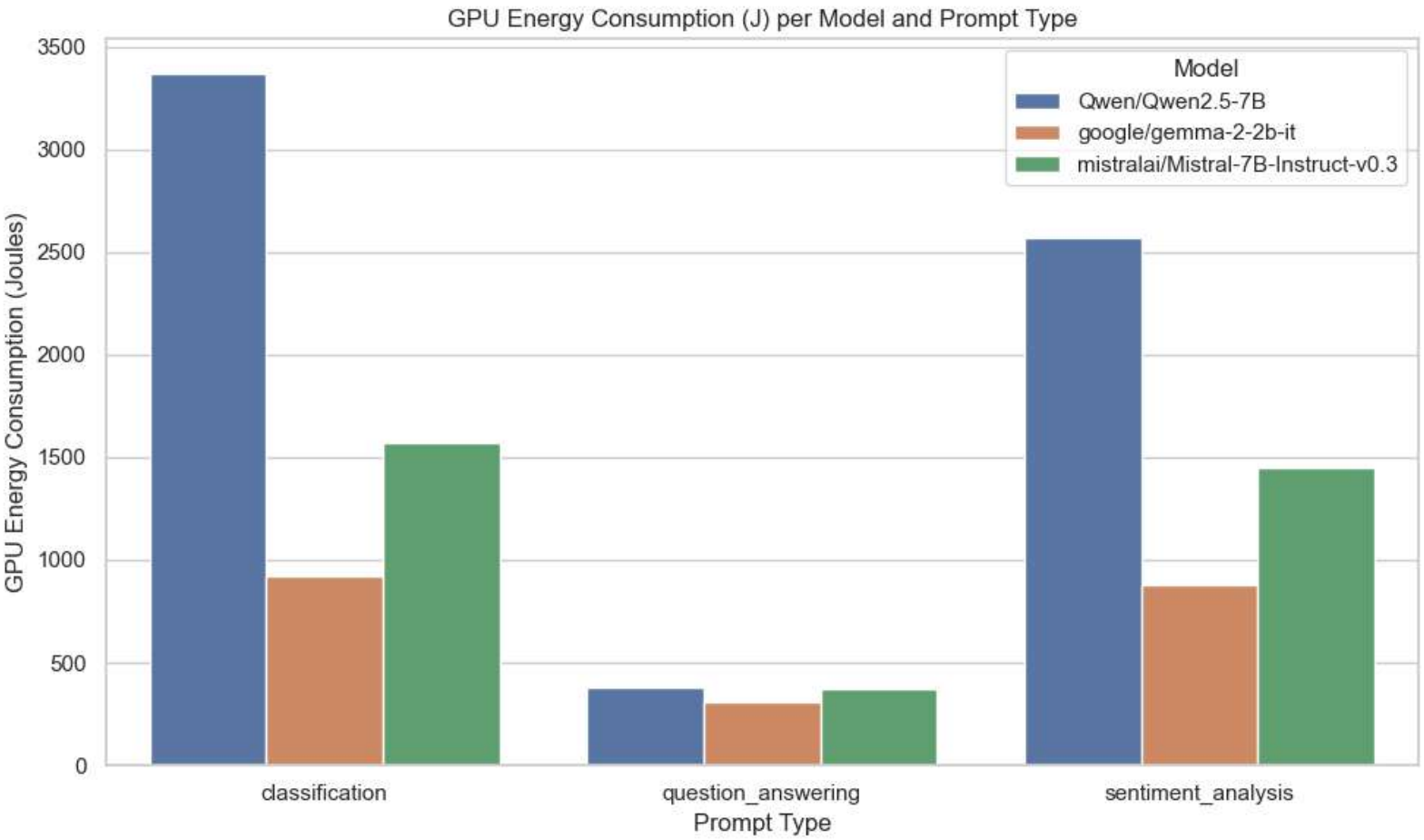
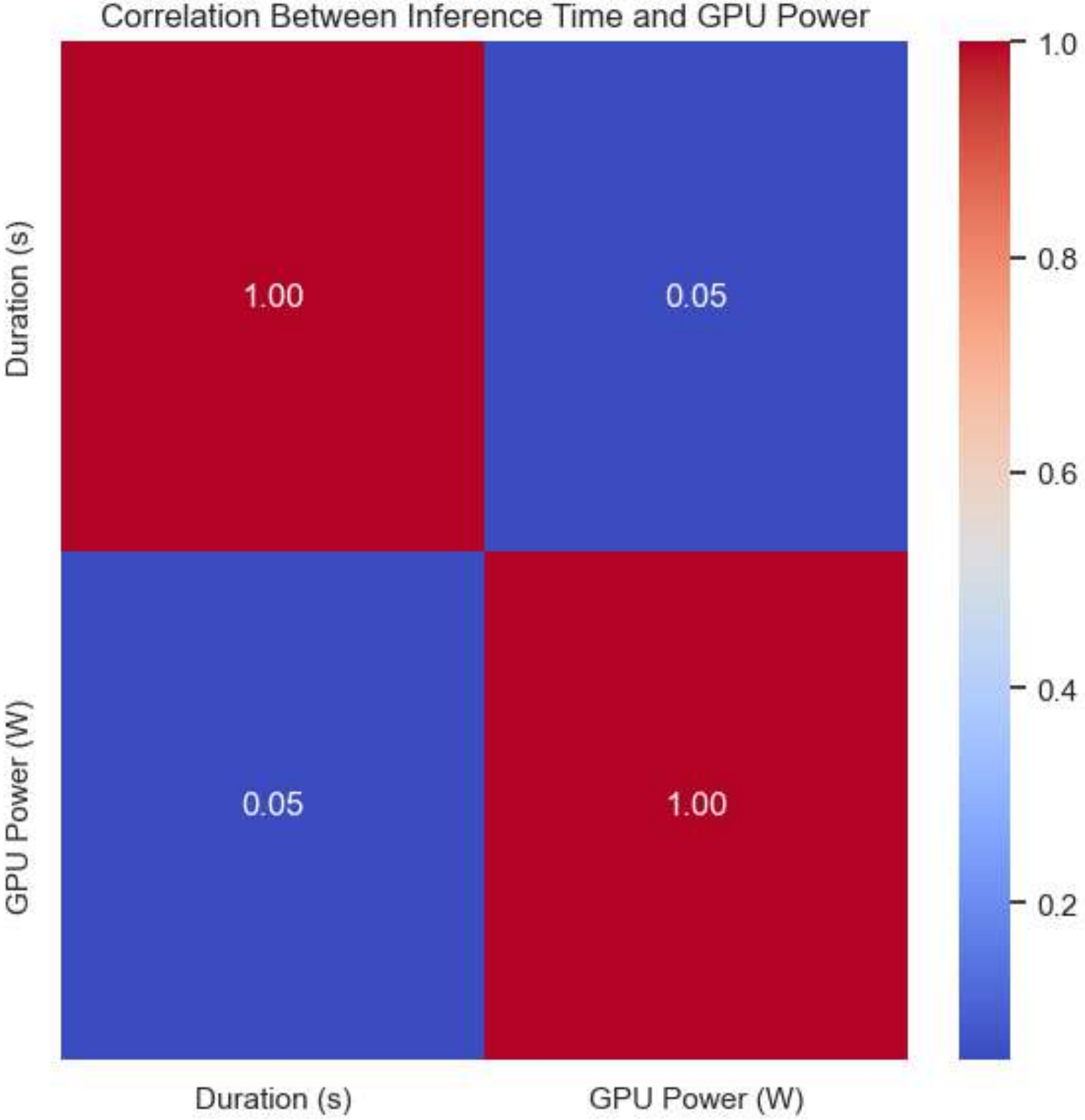
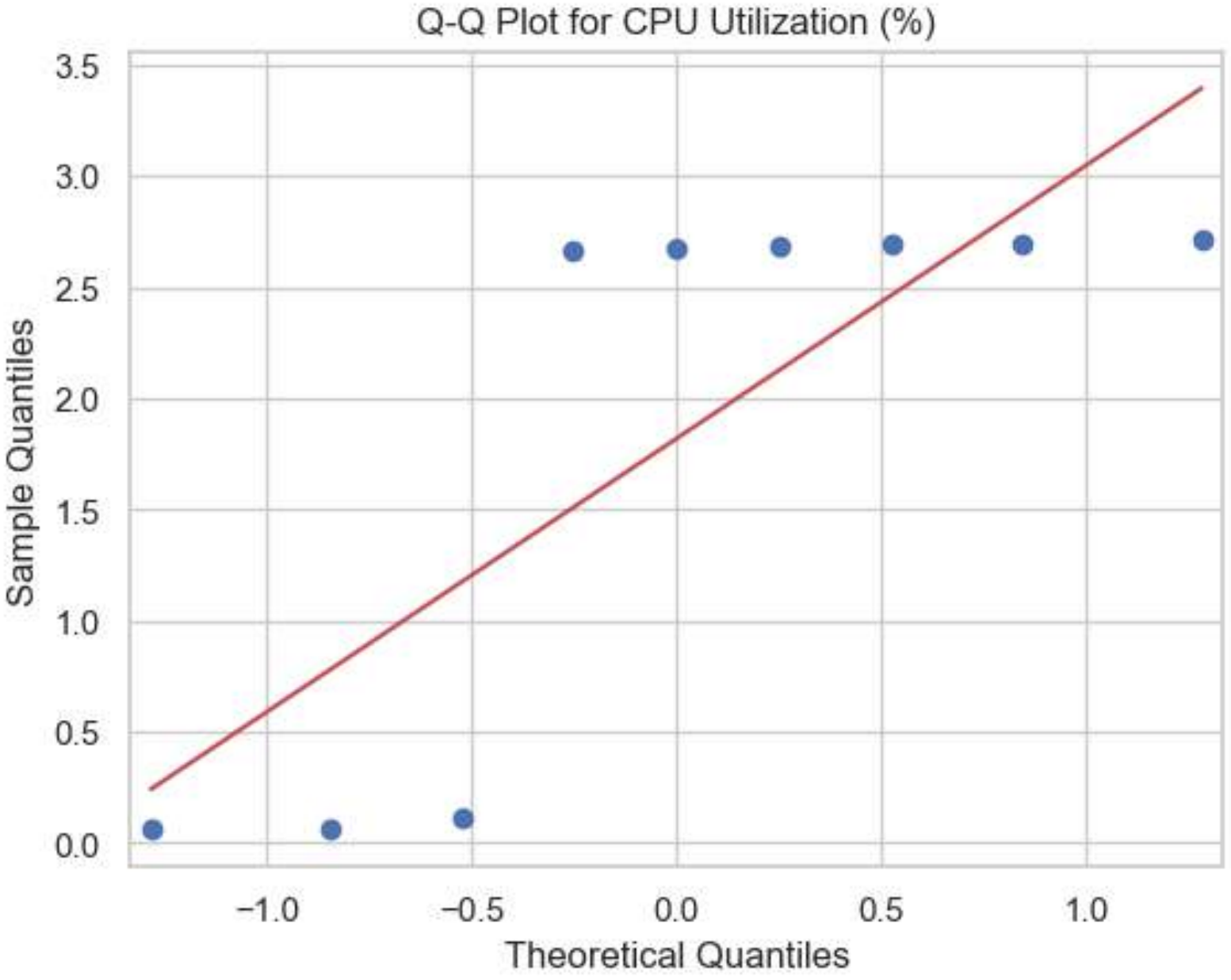

Inference Duration Across Repetitions for Each Task and Model
Task and Model
Task
classification
sentiment_analysis
question_answering
Model
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3
Inference Duration (Seconds)
Repetition (Run number)
5
10
15
20
25
0
10
20
30
40
50
60

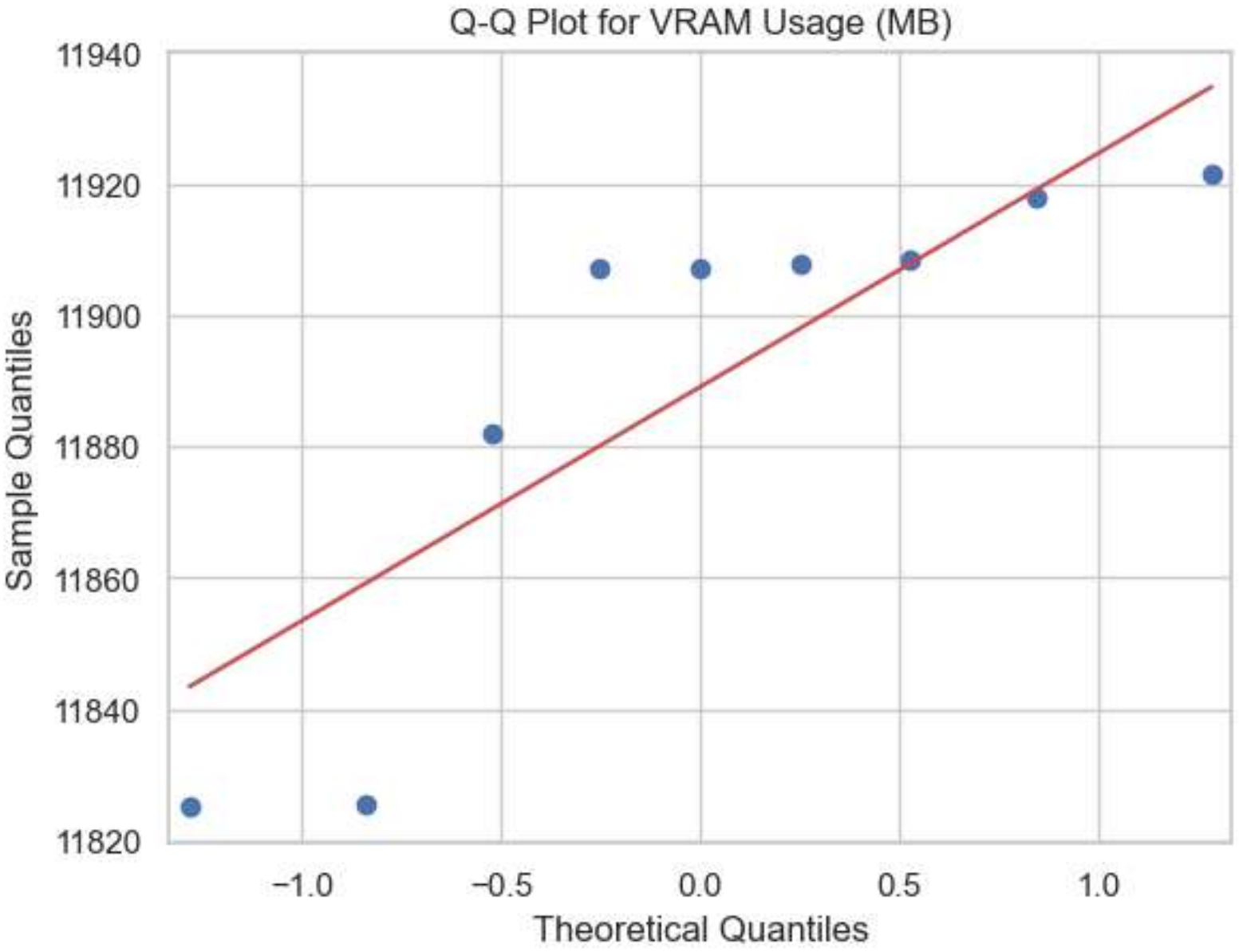

GPU Power vs. GPU Utilization Over 30 Runs for Each Task and Model
GPU Power (W)
GPU Utilization (%)
180
160
140
120
100
80
60
40
20
0
20
40
60
80
100
Task and Model
Task
classification
sentiment_analysis
question_answering
Model
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3

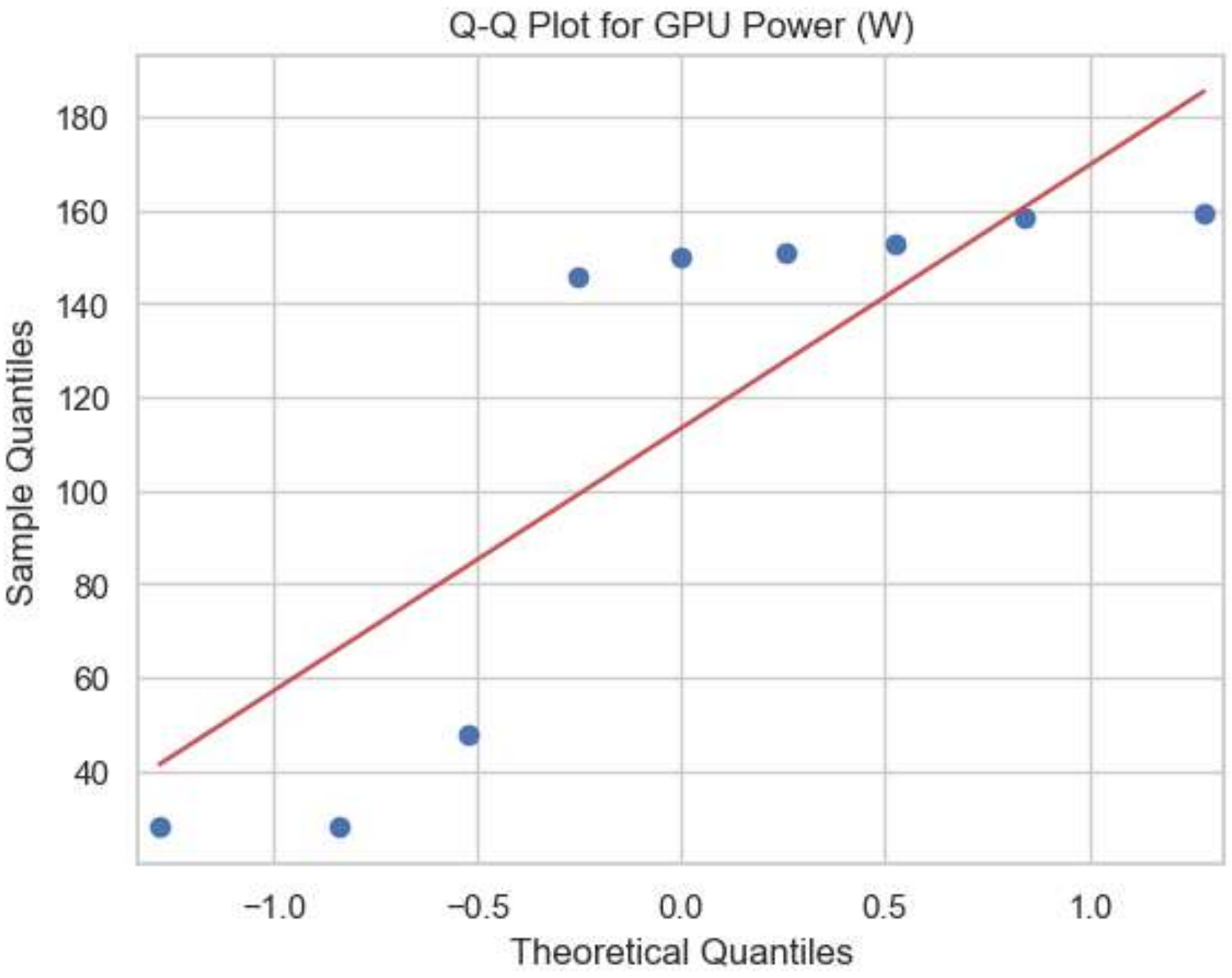

GPU Utilization Over 30 Runs for Each Task and Model
GPU Utilization (%)
Repetition (Run #)
0
20
40
60
80
100
0
10
20
30
40
50
60
Task and Model
Task
classification
sentiment_analysis
question_answering
Model
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3

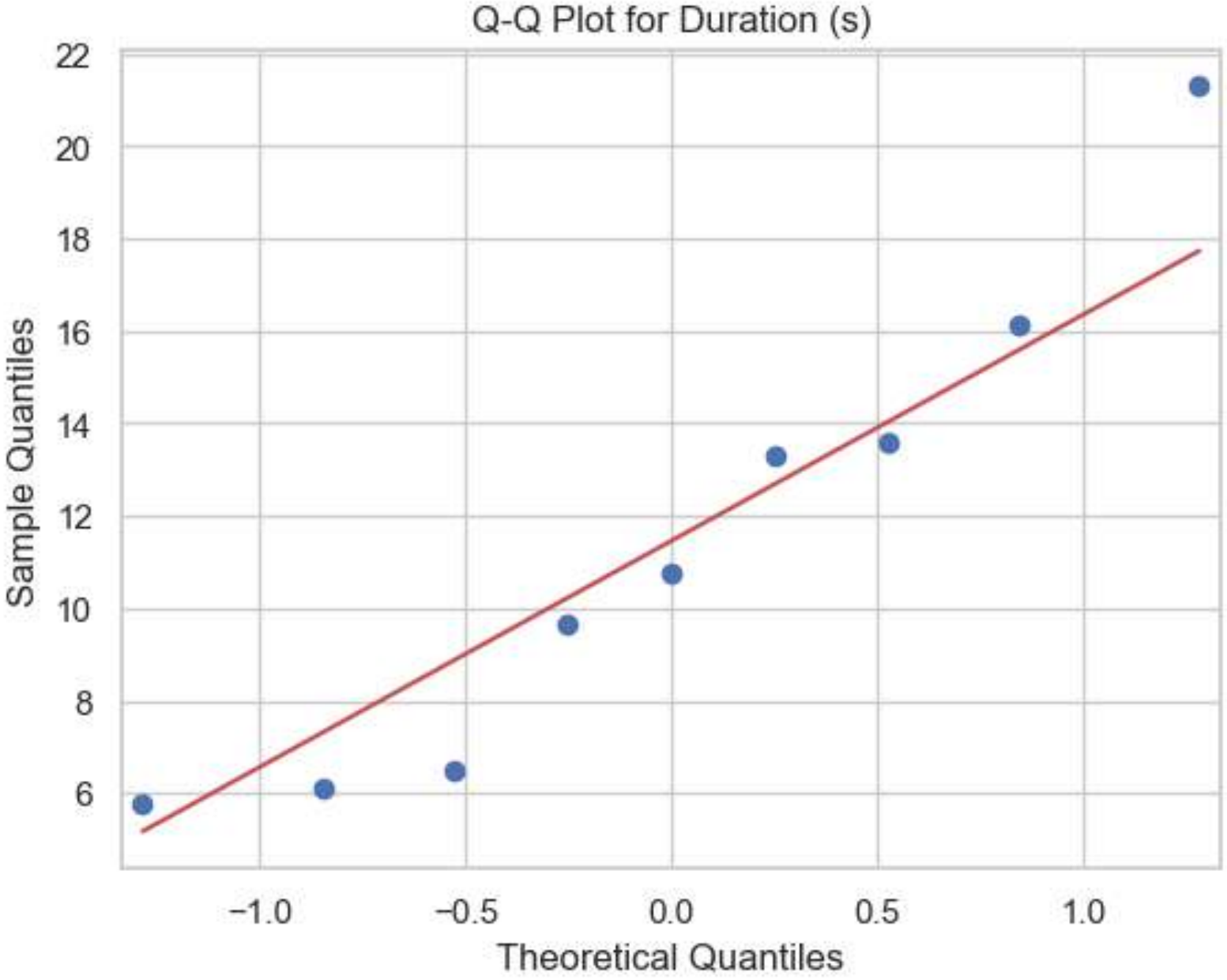

GPU Energy Consumption (J) per Model and Prompt Type
GPU Energy Consumption (Joules)
3500
3000
2500
2000
1500
1000
500
0
Model
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3
classification
question_answering
sentiment_analysis
Prompt Type

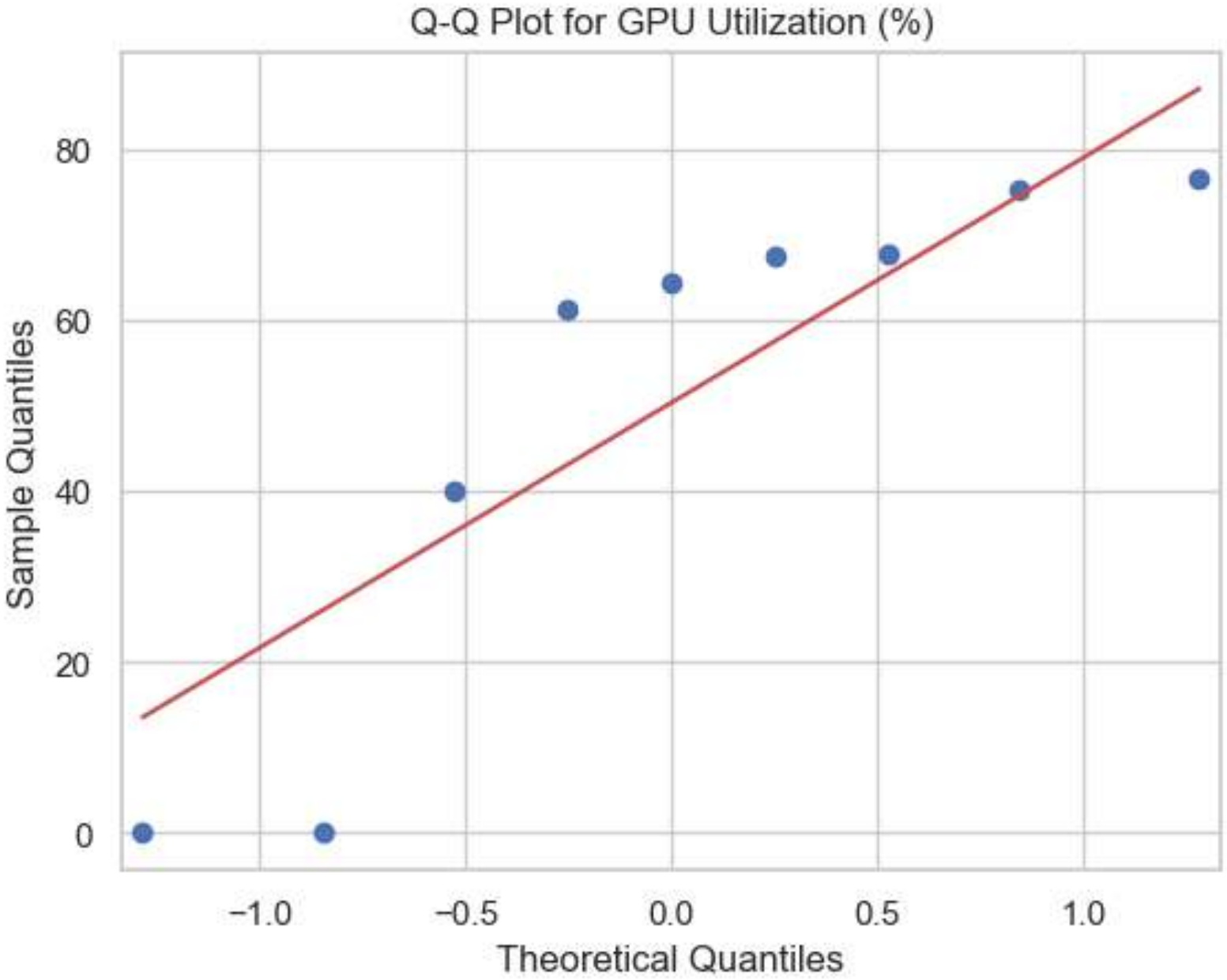

Correlation Between Inference Time and GPU Power
Duration (s)
GPU Power (W)
1.00
0.05
0.05
1.00
Duration (s)
GPU Power (W)
1.0
0.8
0.6
0.4
0.2

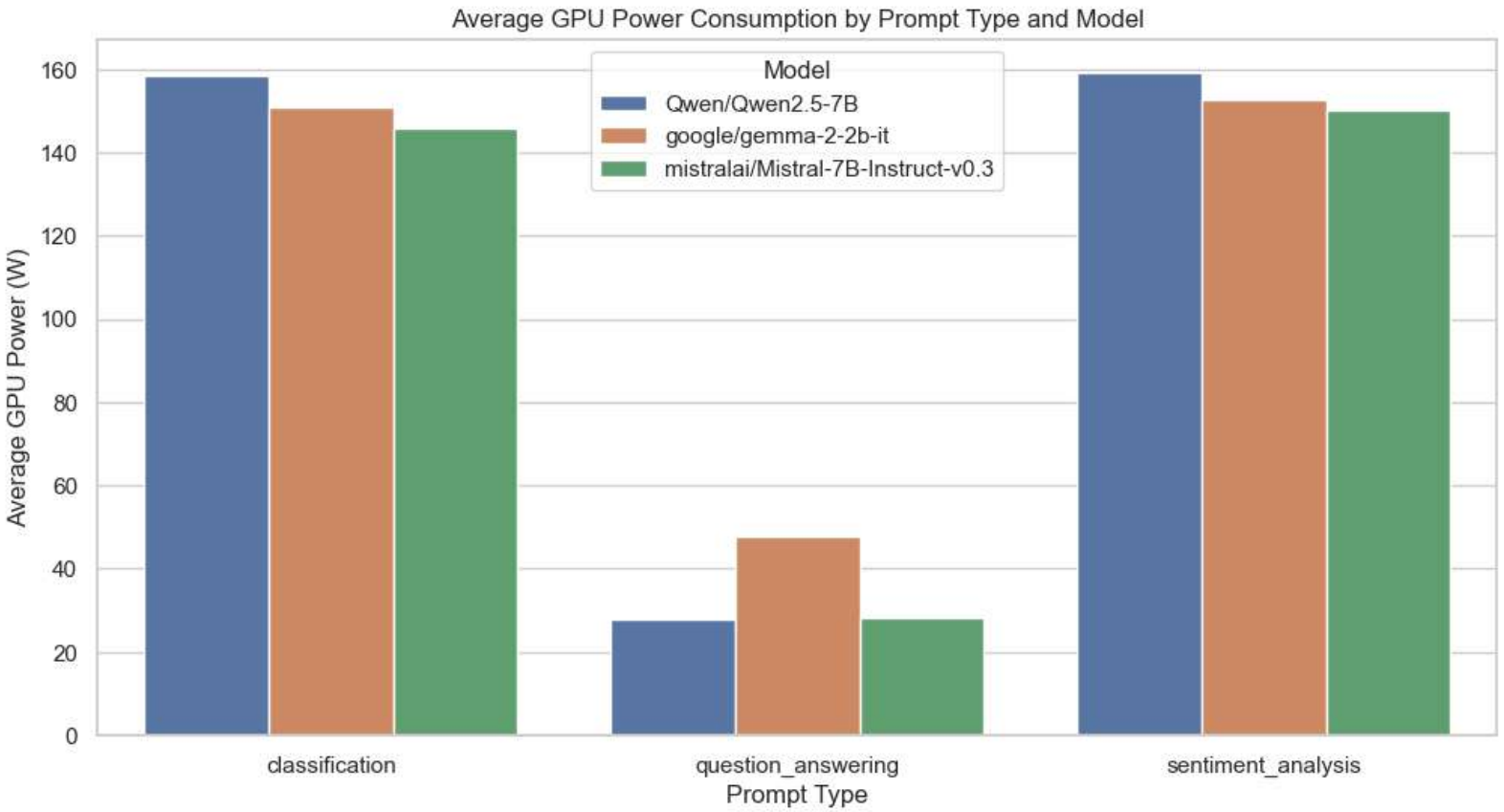

Q-Q Plot for CPU Utilization (%)
Sample Quantiles
Theoretical Quantiles
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0

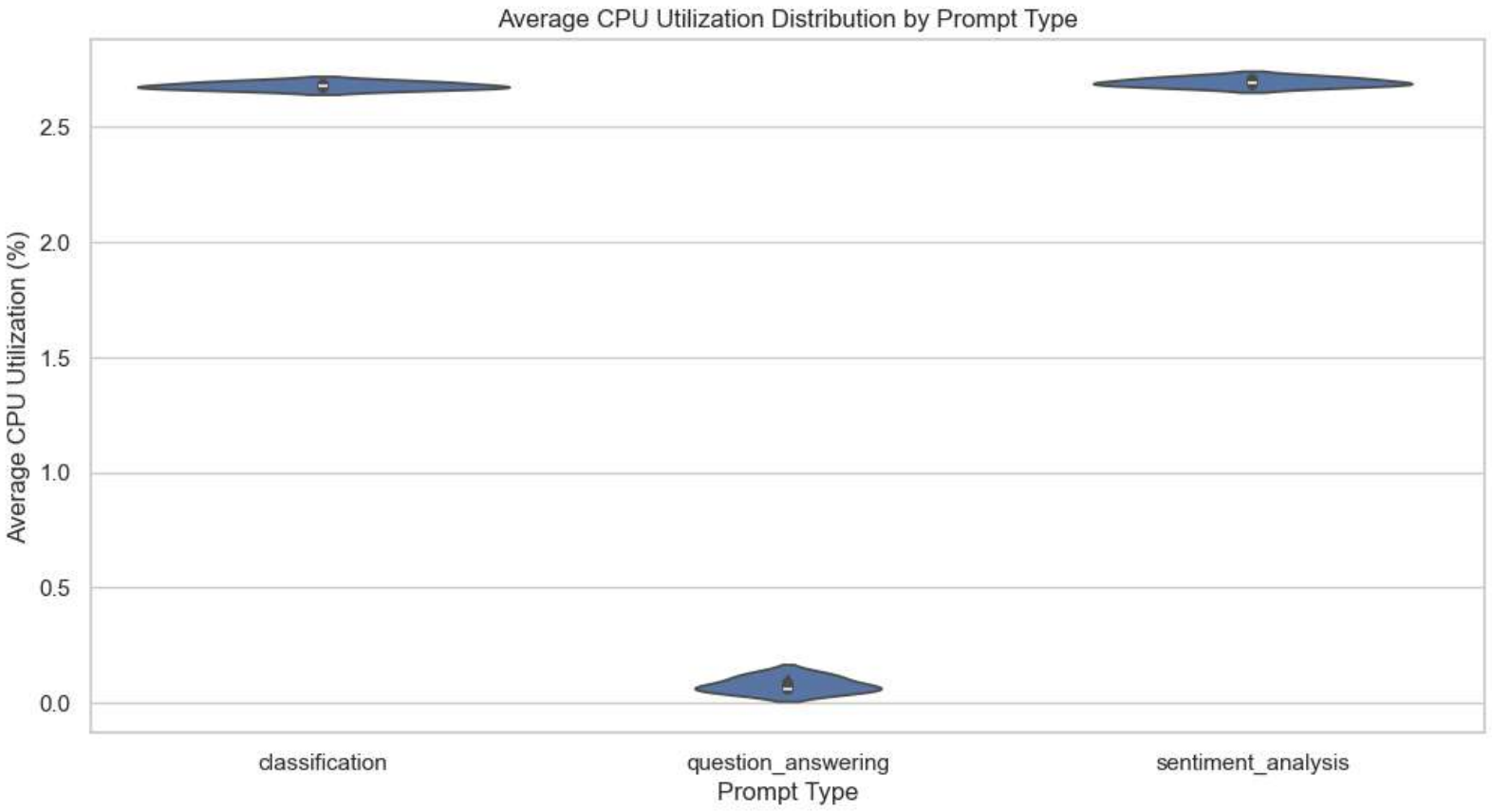

Q-Q Plot for VRAM Usage (MB)
Sample Quantiles
Theoretical Quantiles
11940
11920
11900
11880
11860
11840
11820
−1.0
−0.5
0.0
0.5
1.0

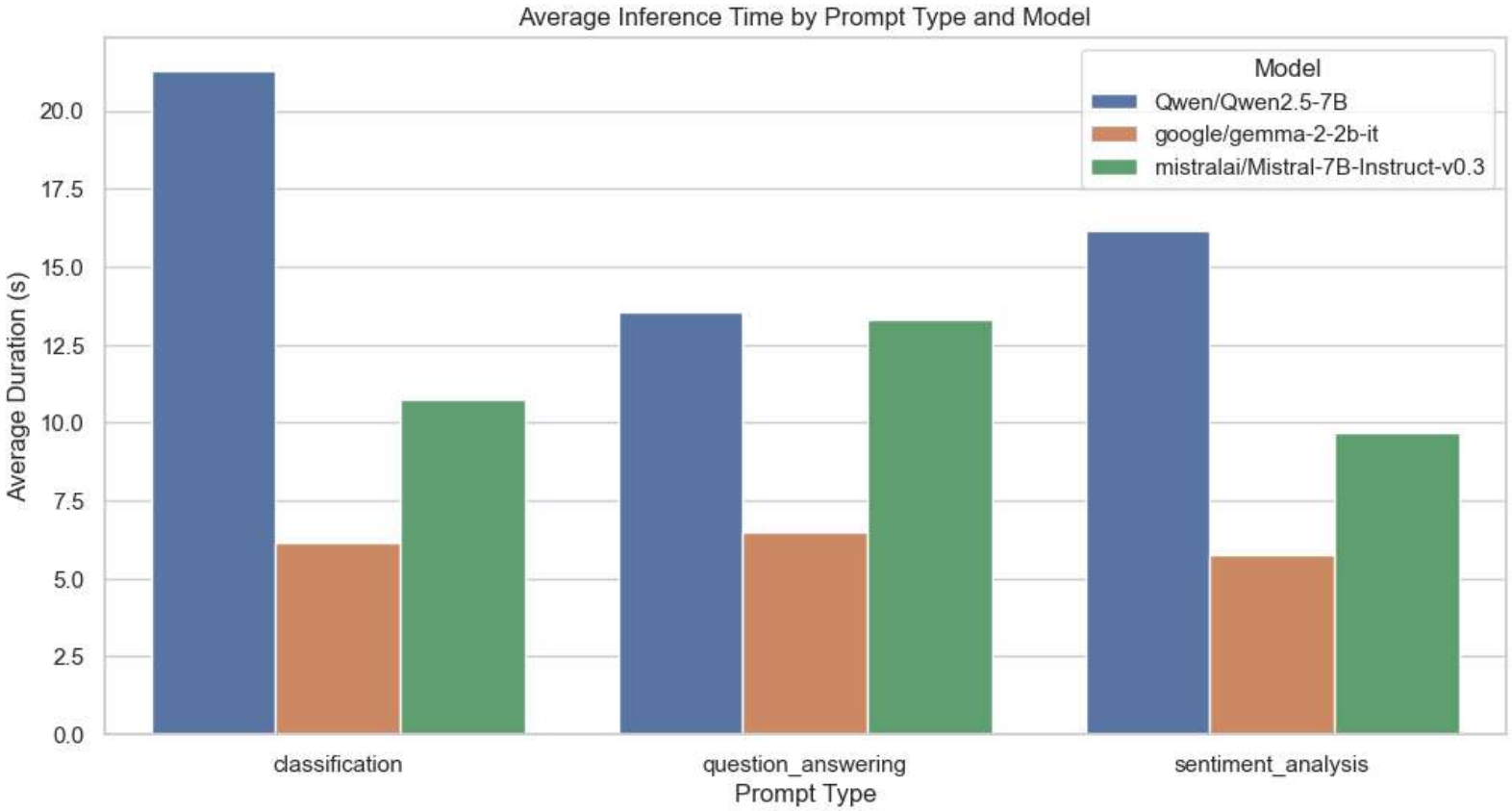

Q-Q Plot for GPU Power (W)
Sample Quantiles
180
160
140
120
100
80
60
40
−1.0
−0.5
0.0
0.5
1.0
Theoretical Quantiles

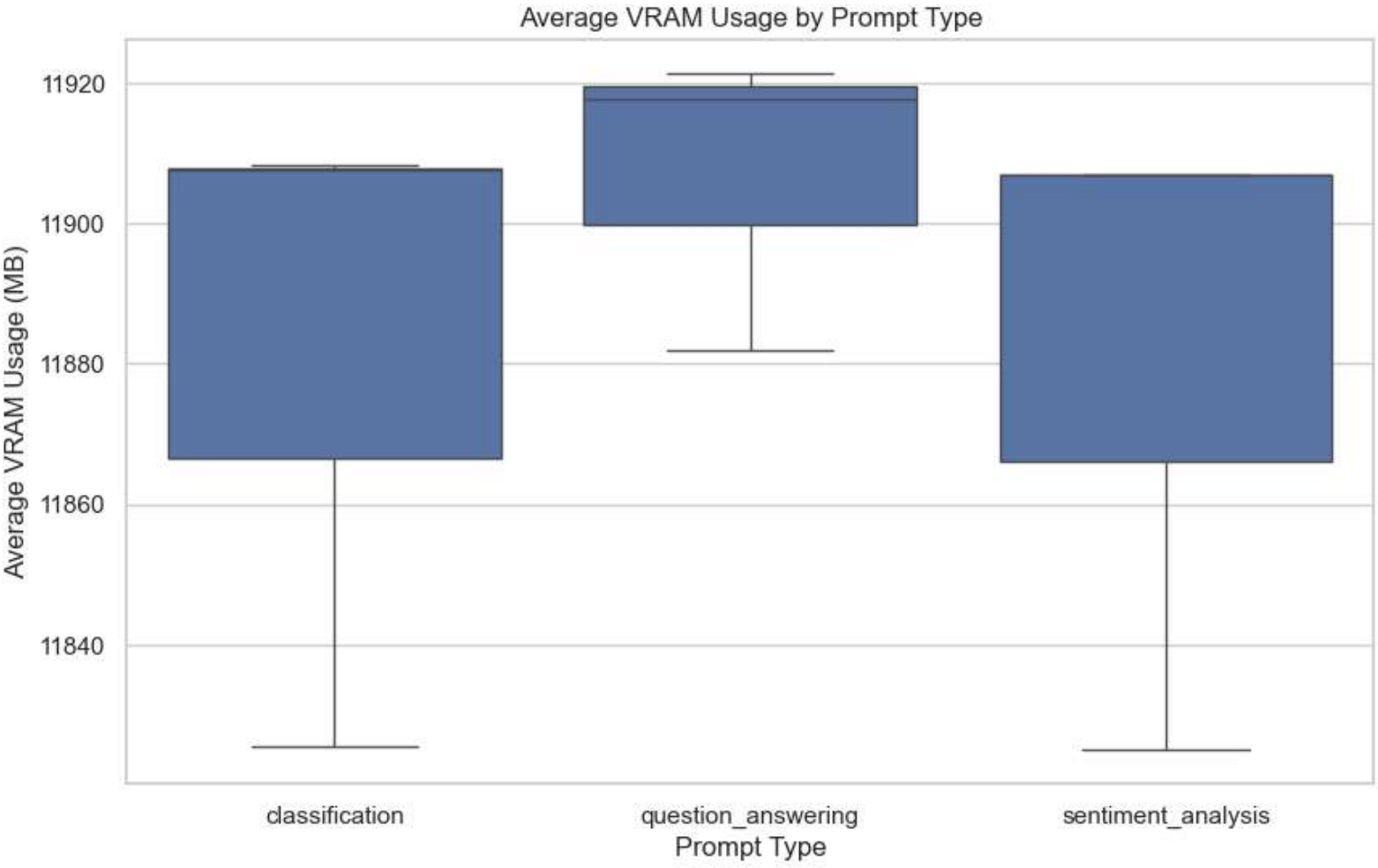

Q-Q Plot for Duration (s)
Sample Quantiles
Theoretical Quantiles
22
20
18
16
14
12
10
8
6
−1.0
−0.5
0.0
0.5
1.0

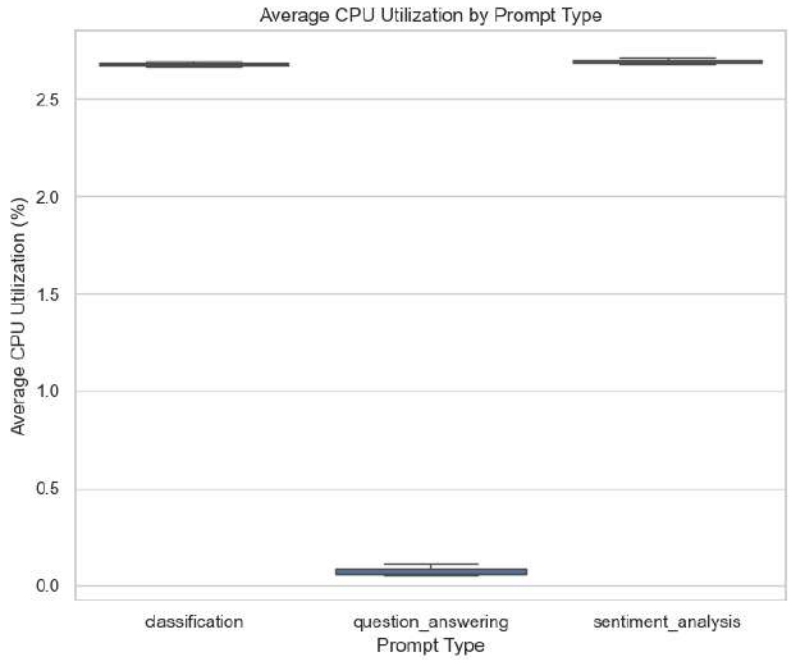

Q-Q Plot for GPU Utilization (%)
Sample Quantiles
Theoretical Quantiles
80
60
40
20
0
−1.0
−0.5
0.0
0.5
1.0

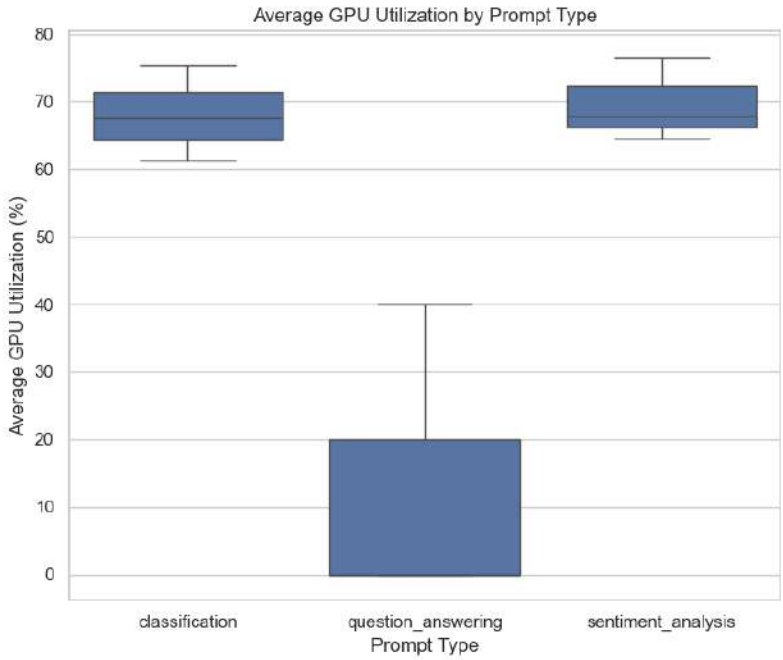

Average GPU Power Consumption by Prompt Type and Model
Model
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3
Average GPU Power (W)
0
20
40
60
80
100
120
140
160
classification
question_answering
sentiment_analysis
Prompt Type

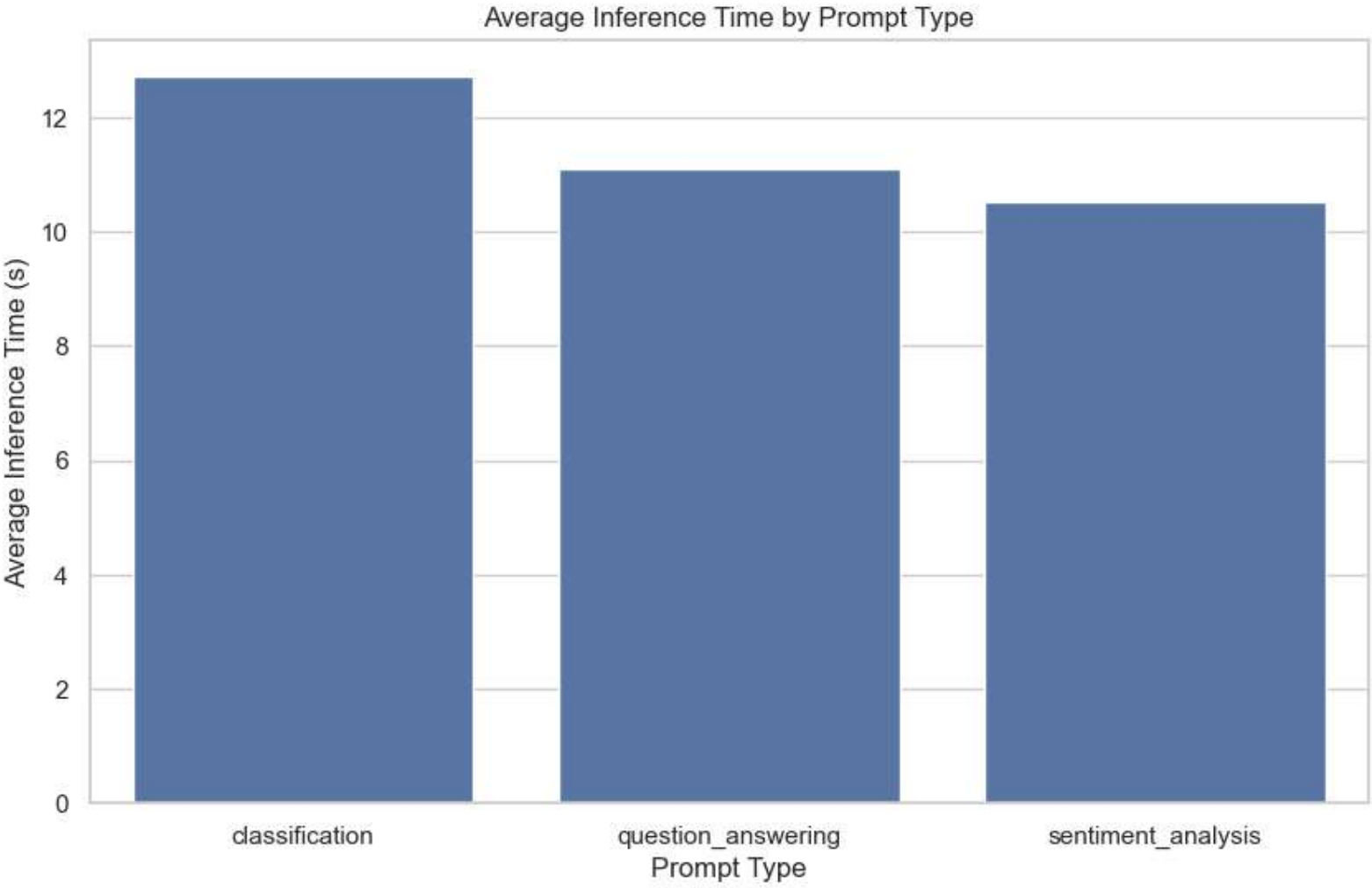

Average CPU Utilization Distribution by Prompt Type
Average CPU Utilization (%)
2.5
2.0
1.5
1.0
0.5
0.0
classification
question_answering
sentiment_analysis
Prompt Type

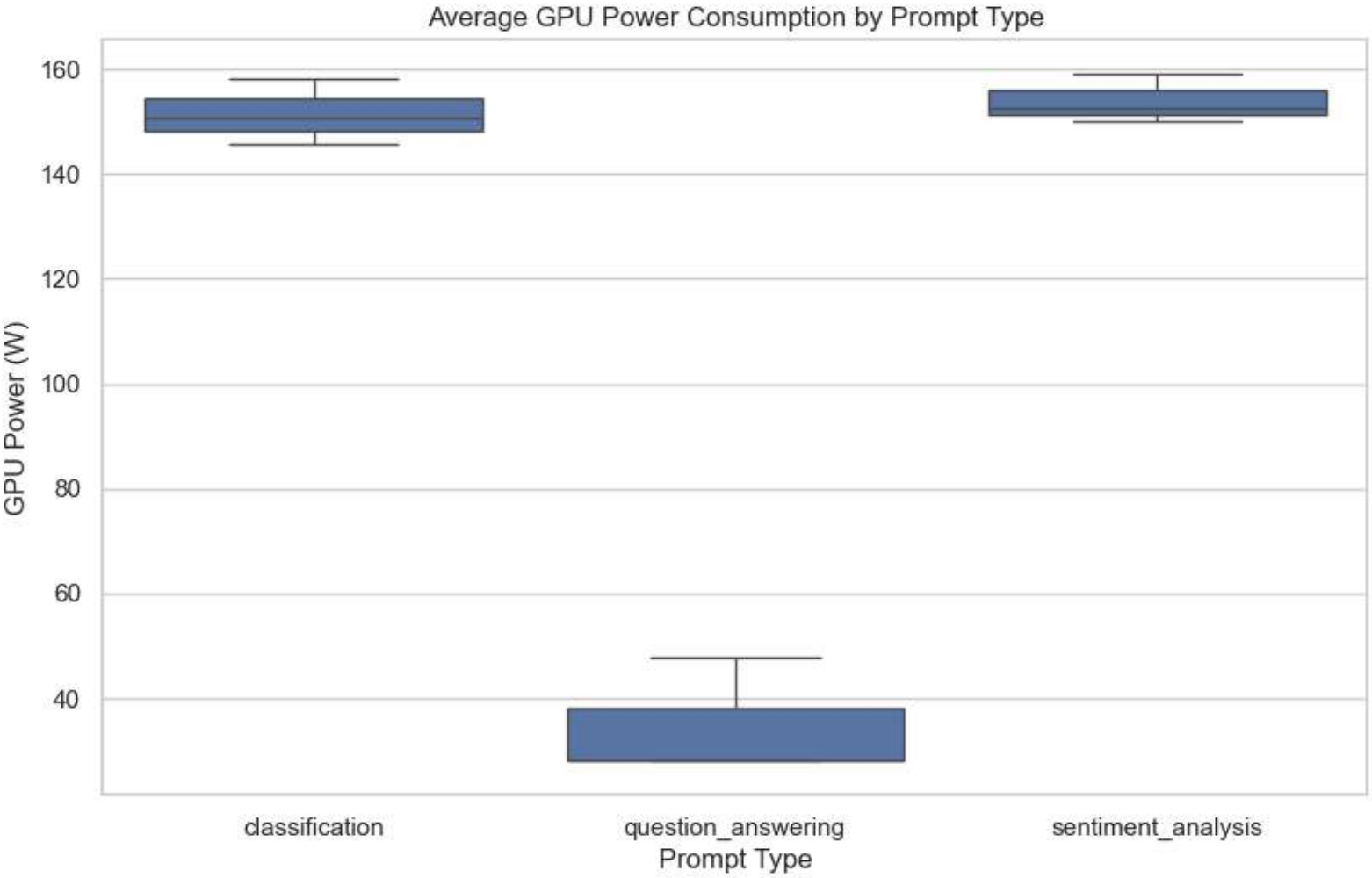

Average Inference Time by Prompt Type and Model
Model
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3
Average Duration (s)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
classification
question_answering
sentiment_analysis
Prompt Type

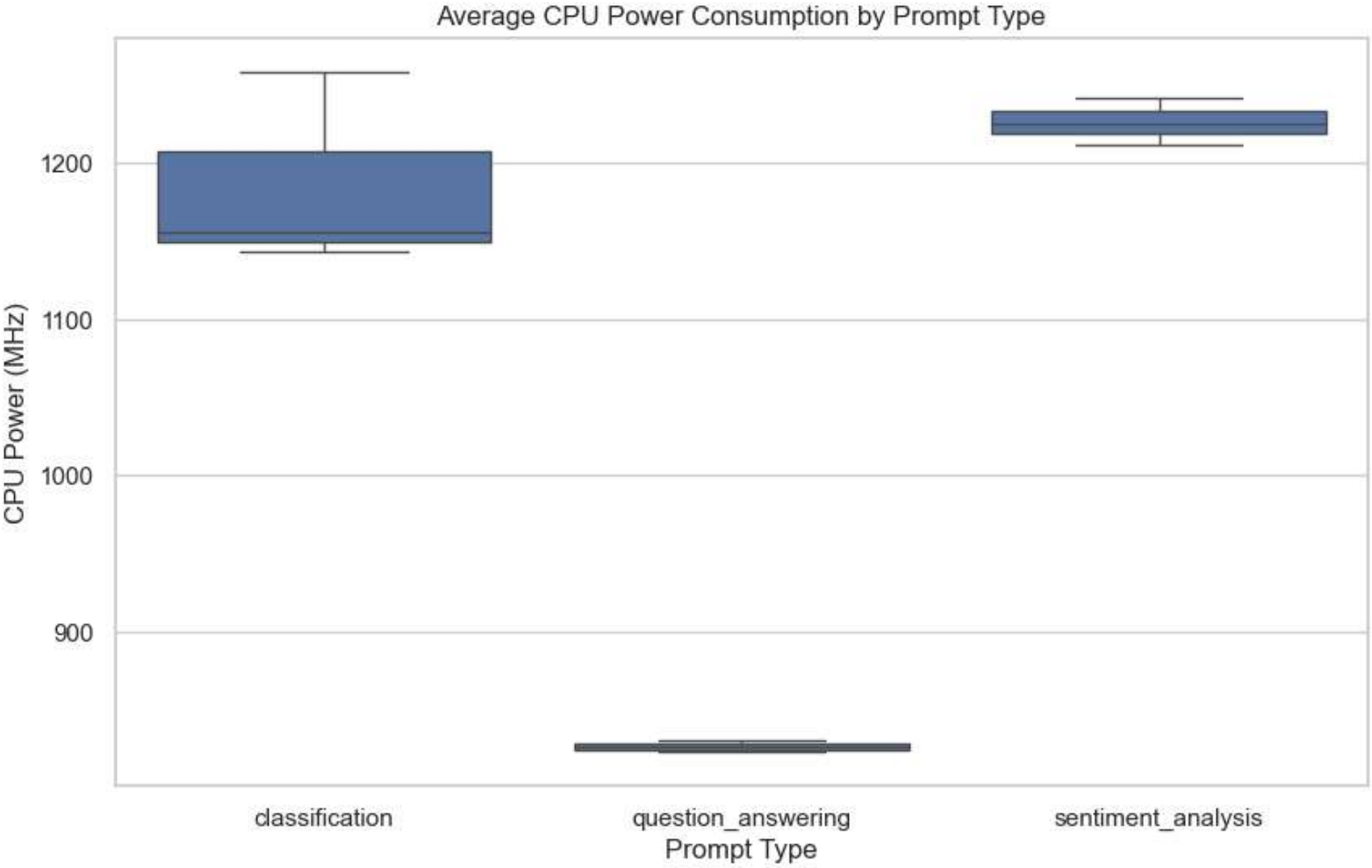

Average VRAM Usage by Prompt Type
Average VRAM Usage (MB)
11920
11900
11880
11860
11840
classification
question_answering
sentiment_analysis
Prompt Type

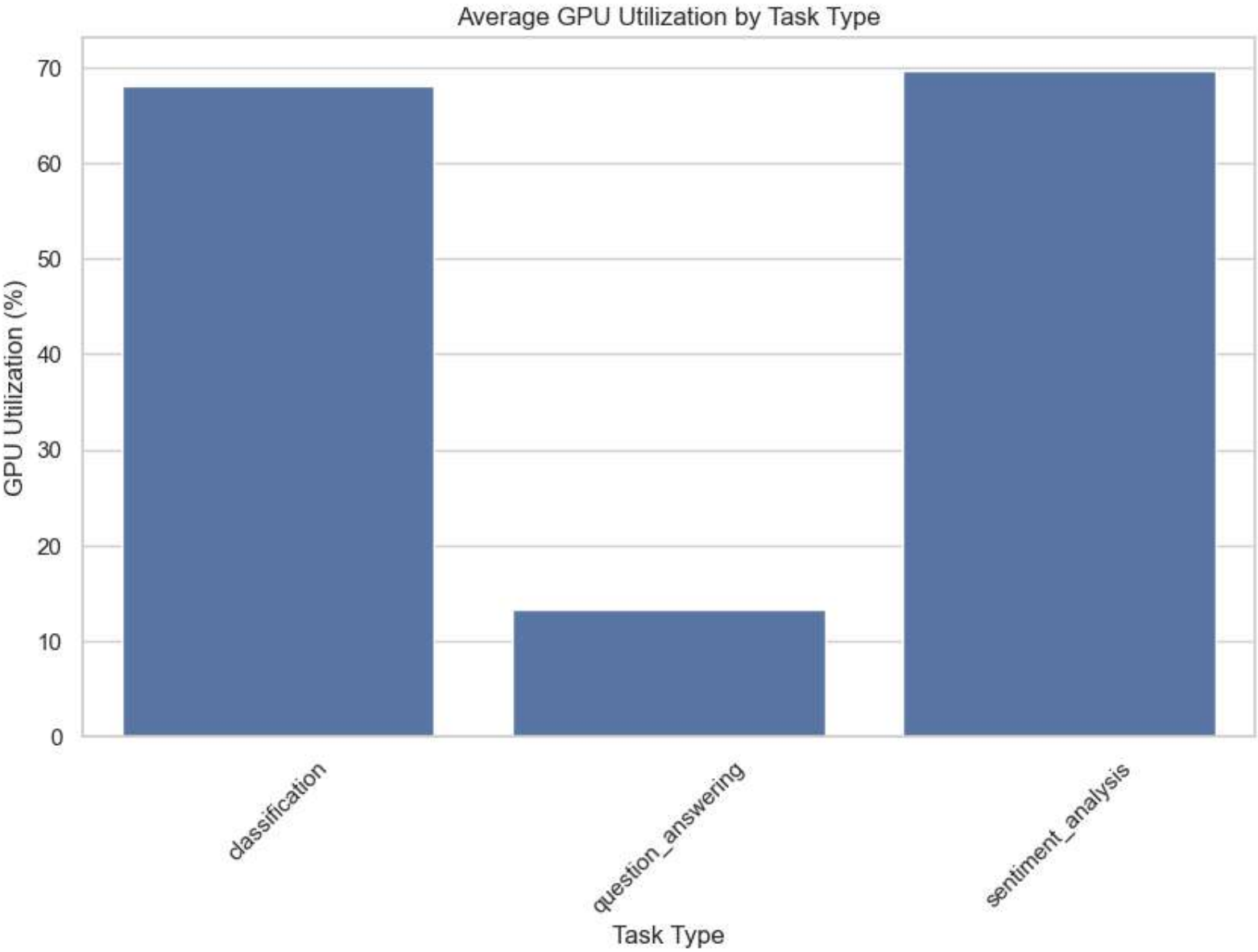

Average CPU Utilization by Prompt Type
Average CPU Utilization (%)
0.0
0.5
1.0
1.5
2.0
2.5
classification
question_answering
sentiment_analysis
Prompt Type

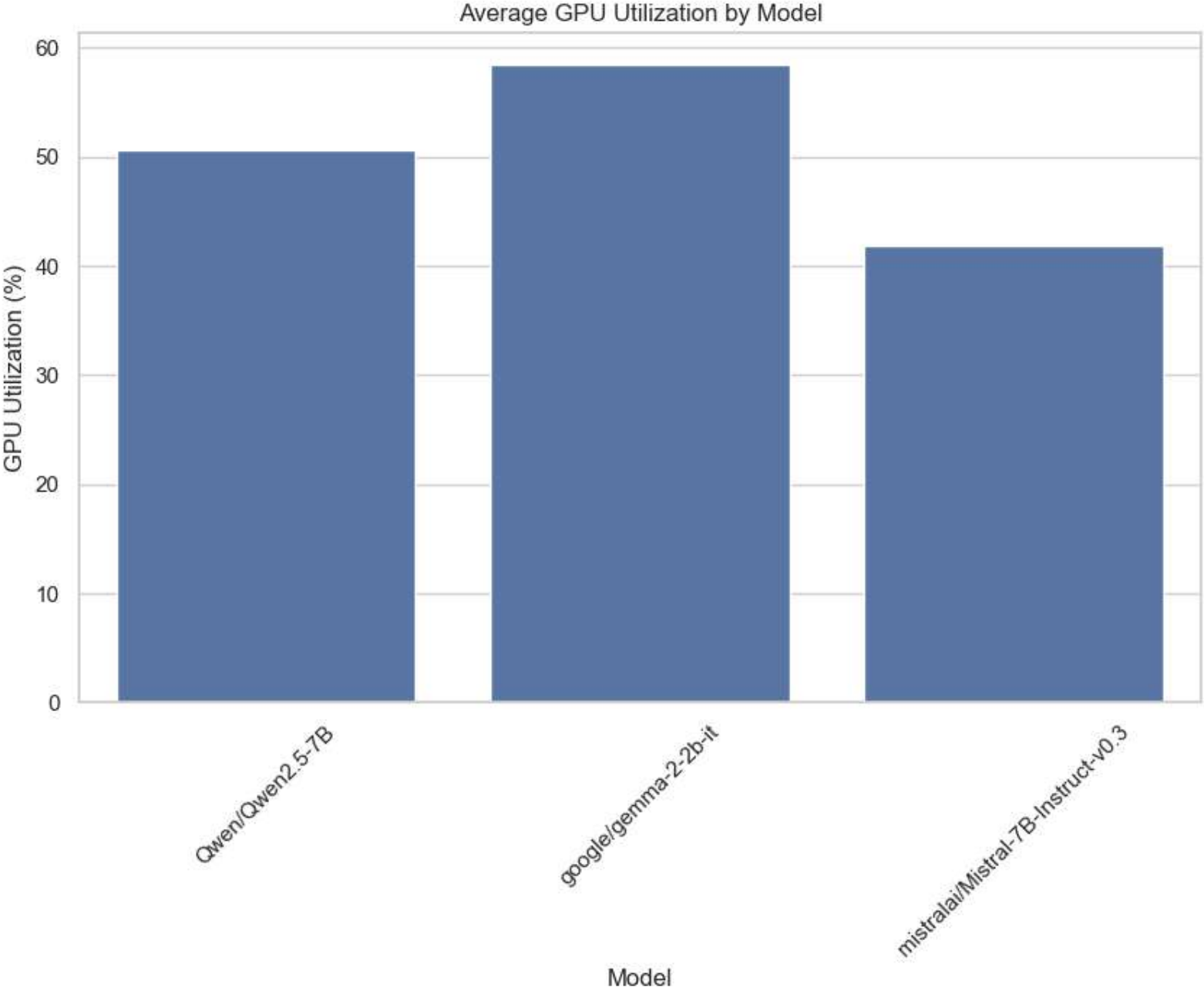

Average GPU Utilization by Prompt Type
Average GPU Utilization (%)
0
10
20
30
40
50
60
70
80
classification
question_answering
sentiment_analysis
Prompt Type


Average Inference Time by Prompt Type
Average Inference Time (s)
0
2
4
6
8
10
12
classification
question_answering
sentiment_analysis
Prompt Type


Average GPU Power Consumption by Prompt Type
GPU Power (W)
160
140
120
100
80
60
40
classification
question_answering
sentiment_analysis
Prompt Type


Average CPU Power Consumption by Prompt Type
CPU Power (MHz)
1200
1100
1000
900
classification
question_answering
sentiment_analysis
Prompt Type


Average GPU Utilization by Task Type
GPU Utilization (%)
0
10
20
30
40
50
60
70
classification
question_answering
sentiment_analysis
Task Type


Average GPU Utilization by Model
GPU Utilization (%)
60
50
40
30
20
10
0
Qwen/Qwen2.5-7B
google/gemma-2-2b-it
mistralai/Mistral-7B-Instruct-v0.3
Model