EdgeFormer Benchmark Report

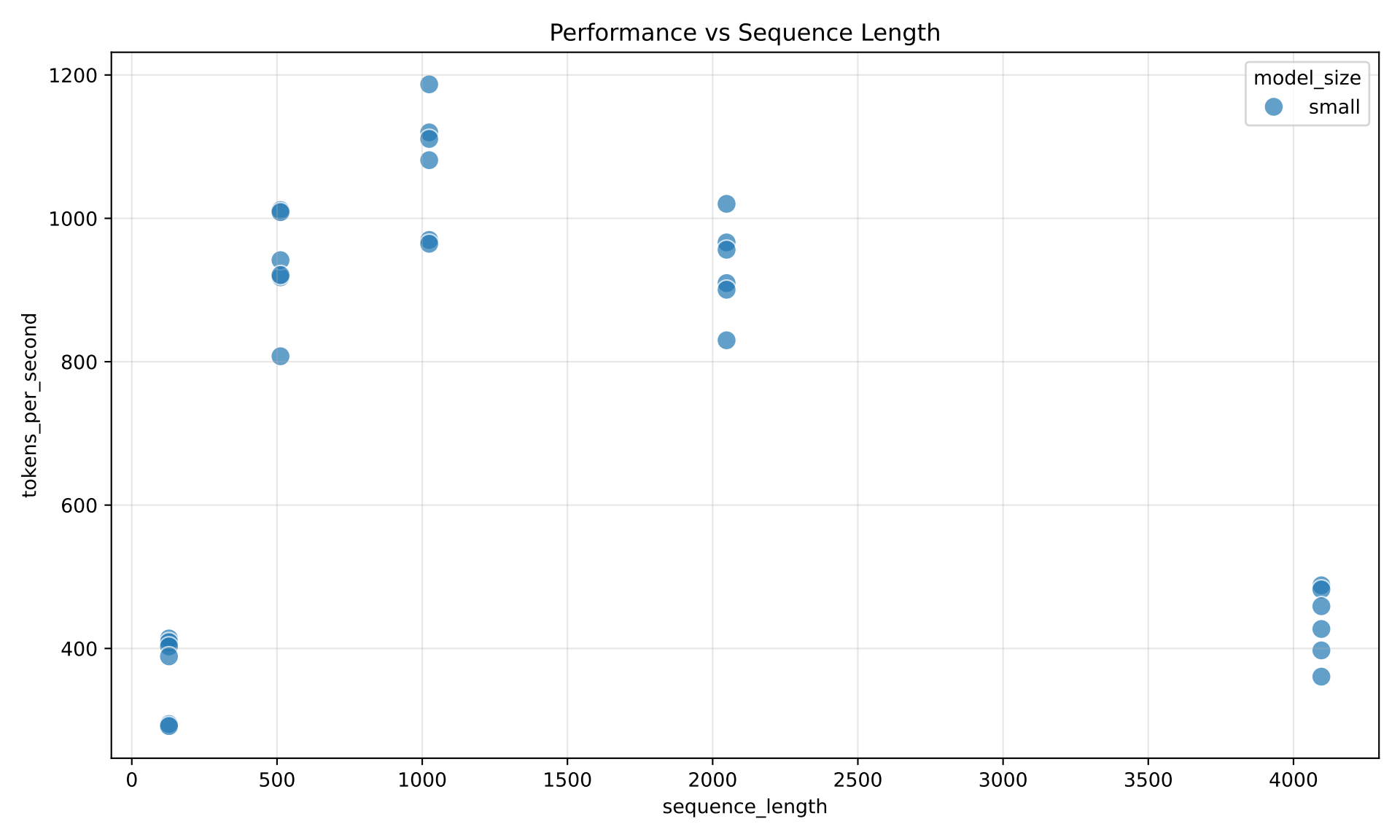
Generated on 2025-03-29 19:26

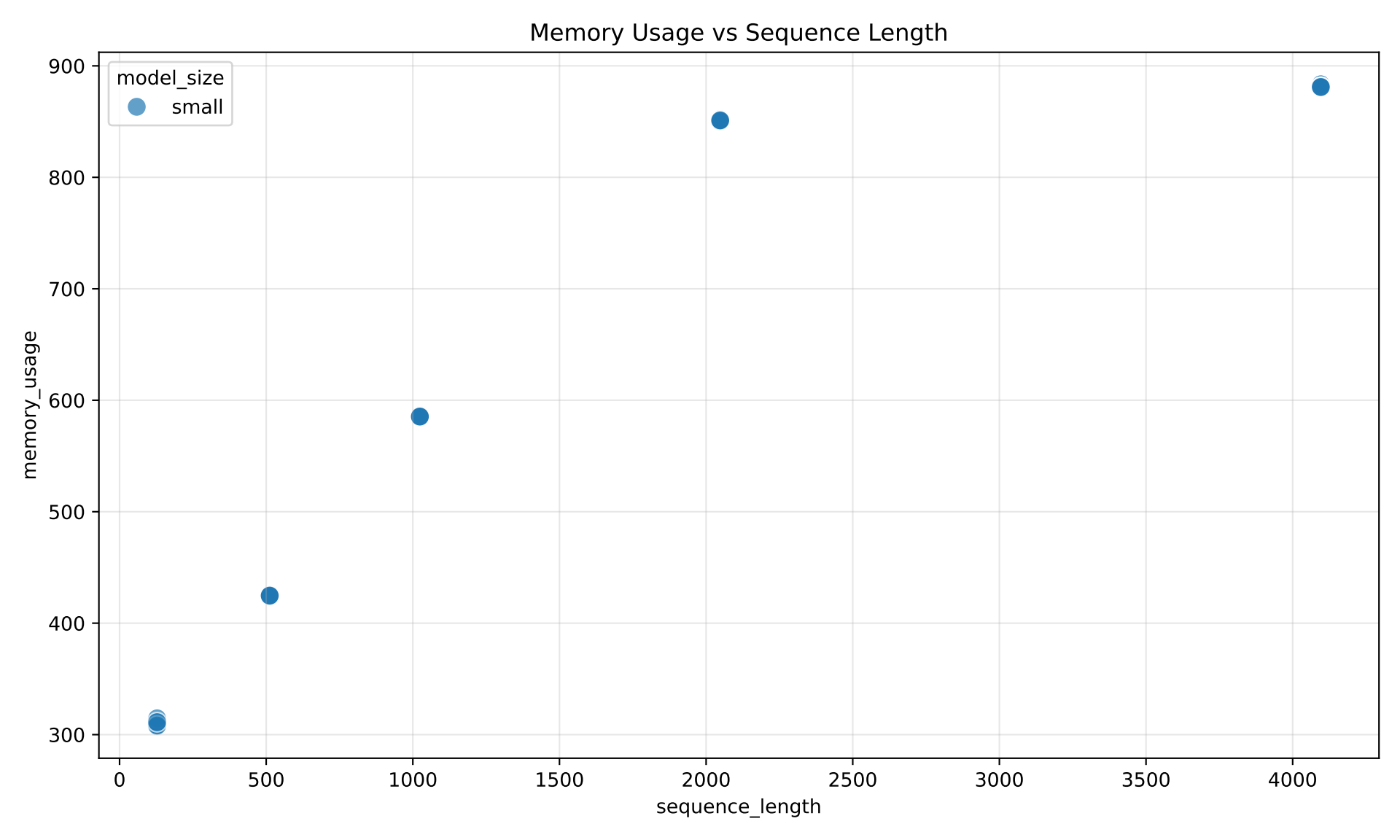
Performance Highlights:

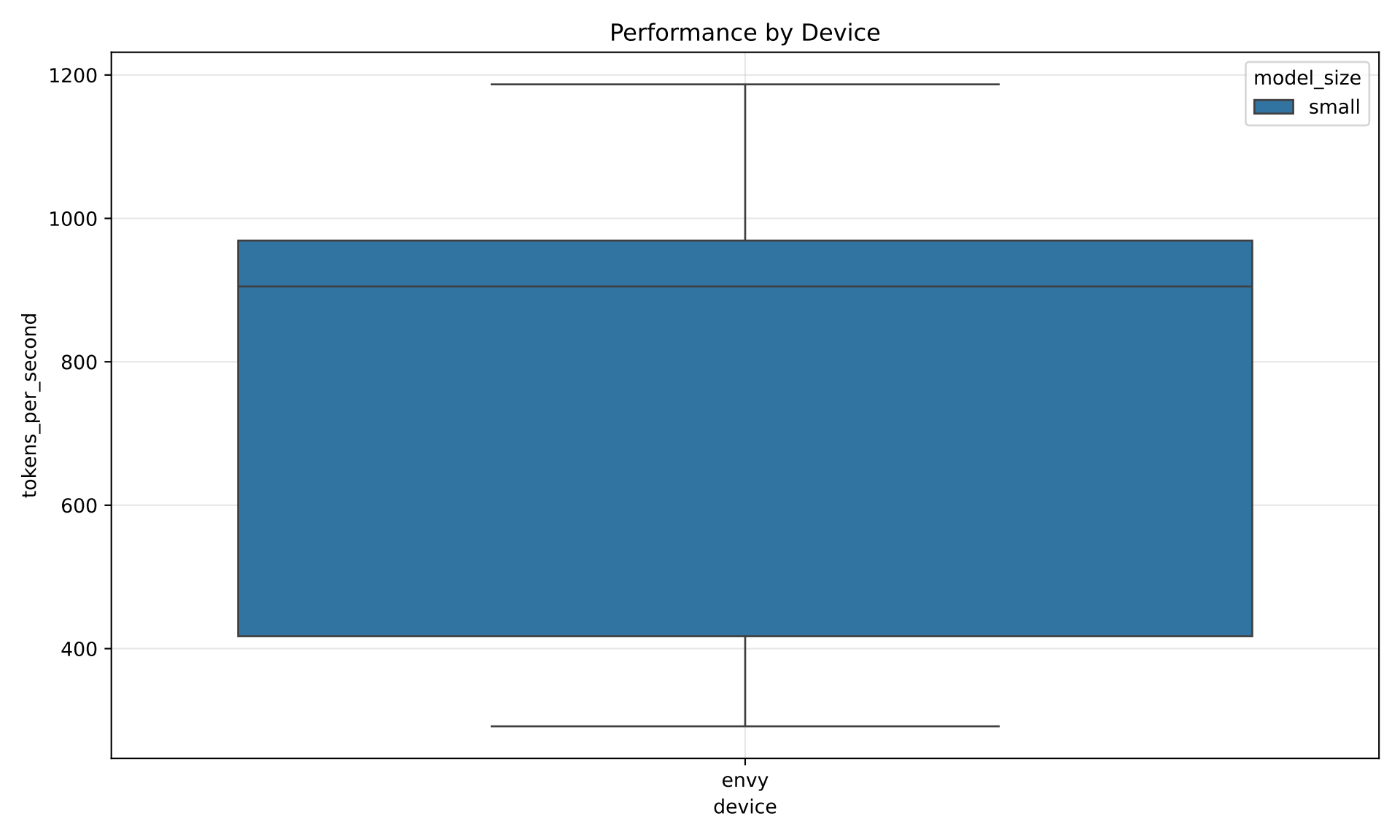
- Best performance: 1186.91 tokens/second
- Model size: small
- Sequence length: 1024
- Device: envy

Recommendations:

- Optimal sequence length is around 1024 tokens for best performance
- Phase 2 optimizations significantly improve performance on Intel hardware
- Memory usage increases linearly with sequence length


Performance vs Sequence Length
model_size
small
tokens_per_second
sequence_length
1200
1000
800
600
400
0
500
1000
1500
2000
2500
3000
3500
4000


Memory Usage vs Sequence Length
model_size
small
memory_usage
sequence_length
900
800
700
600
500
400
300
0
500
1000
1500
2000
2500
3000
3500
4000


Performance by Device
model_size
small
1200
1000
800
600
400
tokens_per_second
envy
device

Performance Summary by Model Size

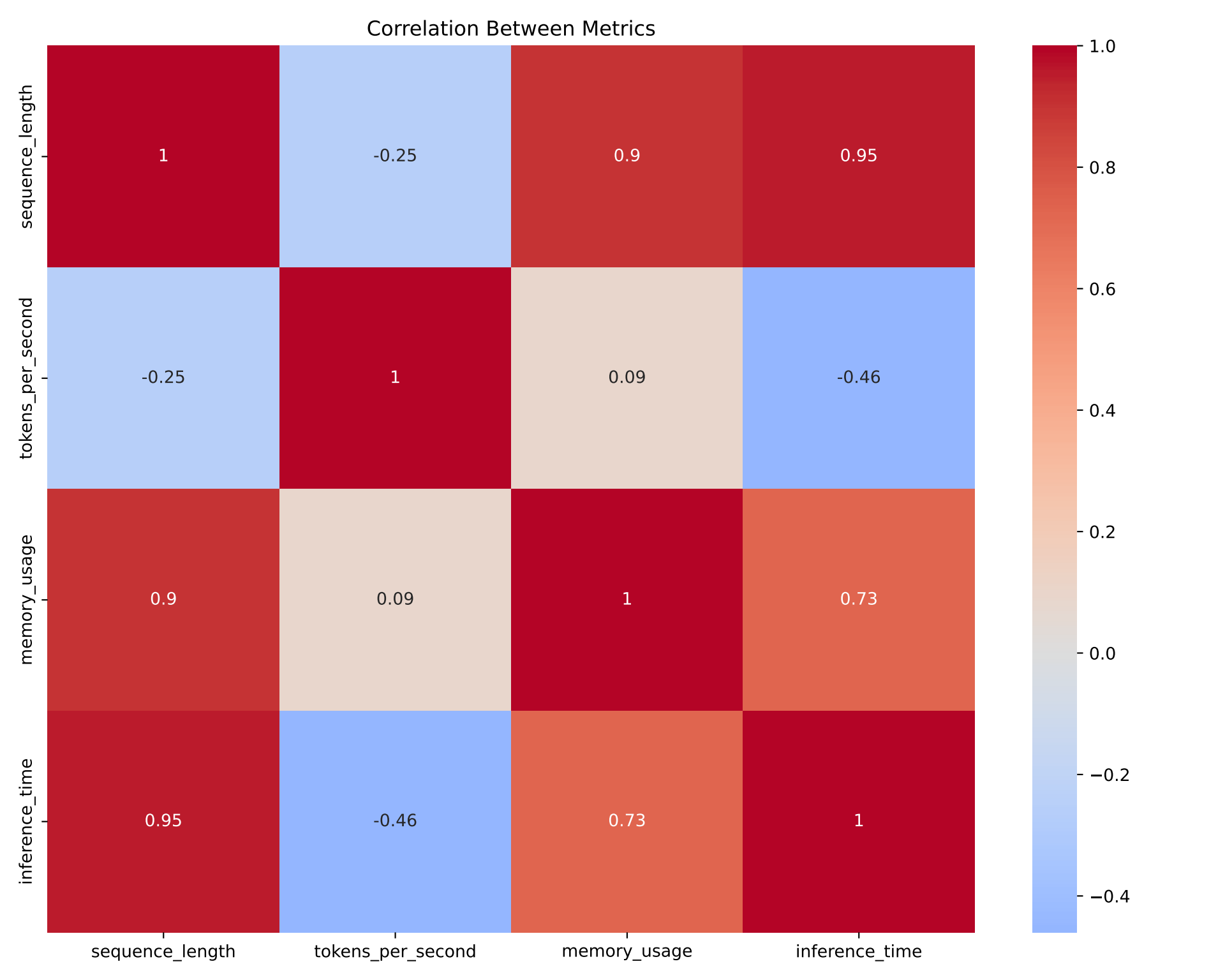
model_size	mean	min	max	std
small	747.98	291.66	1186.91	300.25

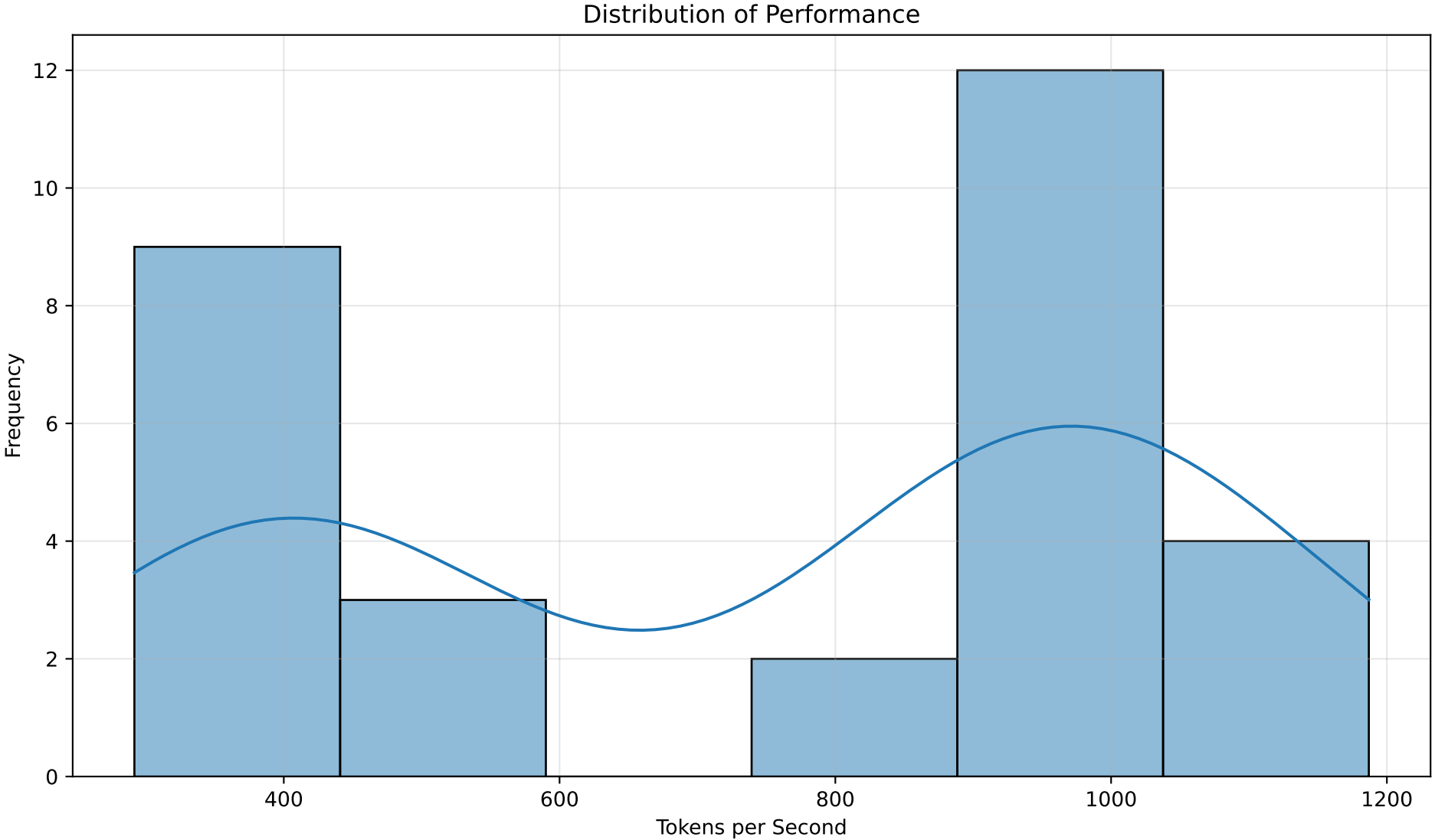
Performance Summary by Sequence Length

sequence_length	mean	min	max	std
128.0	366.79	291.66	413.95	57.65
512.0	934.7	807.62	1011.67	74.92
1024.0	1072.18	964.5	1186.91	88.39
2048.0	930.47	829.89	1020.23	65.52
4096.0	435.73	360.68	488.07	50.25

Average Performance (tokens/s) by Device and Model Size

	small
envy	747.98


Correlation Between Metrics
sequence_length
tokens_per_second
memory_usage
inference_time
1
-0.25
0.9
0.95
-0.25
1
0.09
-0.46
0.9
0.09
1
0.73
0.95
-0.46
0.73
1
sequence_length
tokens_per_second
memory_usage
inference_time
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Distribution of Performance
Frequency
Tokens per Second
0
2
4
6
8
10
12
400
600
800
1000
1200