```
vocab.json:   0%|          | 0.00/1.04M [00:00<?, ?B/s]
merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/1.36M [00:00<?, ?B/s]
/opt/conda/lib/python3.10/site-packages/transformers/tokenization_utils_base.py:1617: FutureWarning: deprecated in transformers v4.45, and will be then set to `False` by default. For more details check
  warnings.warn(
README.md:   0%|          | 0.00/10.5k [00:00<?, ?B/s]
test-00000-of-00001.parquet:   0%|          | 0.00/733k [00:00<?, ?B/s]
train-00000-of-00001.parquet:   0%|          | 0.00/6.36M [00:00<?, ?B/s]
validation-00000-of-00001.parquet:   0%|          | 0.00/657k [00:00<?, ?B/s]
Generating test split:   0%|          | 0/4358 [00:00<?, ? examples/s]
Generating train split:   0%|          | 0/36718 [00:00<?, ? examples/s]
Generating validation split:   0%|          | 0/3760 [00:00<?, ? examples/s]
Loading baseline model (FP32)...
model.safetensors:   0%|          | 0.00/548M [00:00<?, ?B/s]
generation_config.json:   0%|          | 0.00/124 [00:00<?, ?B/s]
Baseline model size: 486.70 MB
Baseline perplexity: 55.55
Baseline inference latency: 7.76 ms

Applying 8-bit quantization using bitsandbytes...
Model size after 8-bit quantization: 168.35 MB
Perplexity after 8-bit quantization: 54.75
Inference latency after 8-bit quantization: 46.60 ms

Changes after 8-bit quantization:
Perplexity change: -0.80
Inference latency change: 0.038835 seconds
Model size reduction: 318.35 MB

Applying 4-bit quantization using bitsandbytes...
Model size after 4-bit quantization: 127.85 MB
Perplexity after 4-bit quantization: 62.42
Inference latency after 4-bit quantization: 23.35 ms

Changes after 4-bit quantization:
Perplexity change: 6.87
Inference latency change: 0.015585 seconds
Model size reduction: 358.85 MB

Applying NF4 quantization using bitsandbytes...
Model size after NF4 quantization: 127.85 MB
Perplexity after NF4 quantization: 59.51
Inference latency after NF4 quantization: 22.95 ms

Changes after NF4 quantization:
Perplexity change: 3.96
Inference latency change: 0.015191 seconds
Model size reduction: 358.85 MB

--- Analysis ---
Baseline model size: 486.70 MB
8-bit model size: 168.35 MB
4-bit model size: 127.85 MB
NF4 model size: 127.85 MB

Perplexity Comparison:
Baseline perplexity: 55.55
8-bit perplexity: 54.75
4-bit perplexity: 62.42
NF4 perplexity: 59.51

Inference Latency Comparison:
Baseline latency: 7.76 ms
8-bit latency: 46.60 ms
4-bit latency: 23.35 ms
NF4 latency: 22.95 ms
```