```
(base) ➜  A4 /opt/anaconda3/bin/python q1.py
/opt/anaconda3/lib/python3.11/site-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_
download` is deprecated and will be removed in version 1.0.0. Downloads always resume when possible. If yo
u want to force a new download, use `force_download=True`.
  warnings.warn(
Model size before quantization: 486.70 MB
README.md: 100%|████████████████████████████| 10.5k/10.5k [00:00<00:00, 13.3MB/s]
test-00000-of-00001.parquet: 100%|███████████| 733k/733k [00:00<00:00, 13.3MB/s]
train-00000-of-00001.parquet: 100%|████████| 6.36M/6.36M [00:00<00:00, 21.6MB/s]
validation-00000-of-00001.parquet: 100%|█████| 657k/657k [00:00<00:00, 30.8MB/s]
Generating test split: 100%|█████| 4358/4358 [00:00<00:00, 344863.06 examples/s]
Generating train split: 100%|█| 36718/36718 [00:00<00:00, 1079020.61 examples/s]
Generating validation split: 100%|█| 3760/3760 [00:00<00:00, 930692.42 examples/
[Perplexity before quantization: 55.55                                         ]
Inference latency before quantization: 49.96 ms

Performing whole-model quantization...
Model size after whole-model quantization: 155.48 MB
Perplexity after whole-model quantization: 51.67
Inference latency after whole-model quantization: 52.10 ms

Changes after whole-model quantization:
Perplexity increase: -3.87
Inference latency change: 0.002139 seconds
Model size reduction: 331.22 MB
/opt/anaconda3/lib/python3.11/site-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_
download` is deprecated and will be removed in version 1.0.0. Downloads always resume when possible. If yo
u want to force a new download, use `force_download=True`.
  warnings.warn(

Performing selective quantization on MLP layers...
Model size after MLP quantization: 459.81 MB
Perplexity after MLP quantization: 50.80
Inference latency after MLP quantization: 52.65 ms

Changes after MLP quantization:
Perplexity increase: -4.75
Inference latency change: 0.002683 seconds
Model size reduction: 26.89 MB

Performing selective quantization on decoder layers...
Model size after decoder quantization: 378.59 MB
Perplexity after decoder quantization: 51.69
Inference latency after decoder quantization: 50.80 ms

Changes after decoder quantization:
Perplexity increase: -3.85
Inference latency change: 0.000833 seconds
Model size reduction: 108.11 MB
(base) ➜  A4
```