Task2 Report

Evaluation Metrics

Metric	Baseline GPT-2	INT8 – BitsAndBytes	NF4 – BitsAndBytes	INT8 – Custom
Memory (MB)	486.7119598388672	168.3559799194336	127.8559799194336	486.7119598388672
Latency (ms/sample)	4.62472749383826	7.582223854566876	6.251330908976103	4.794717901631405
Accuracy	0.939078947368421	0.939078947368421	0.9384210526315789	0.9380263157894737
Precision	0.9391756508348497	0.9391618030224902	0.9385485081472786	0.9379530834464211
Recall	0.939078947368421	0.939078947368421	0.9384210526315789	0.9380263157894737
F1 Score	0.9390681378448096	0.9390594030818255	0.9384185468208894	0.9379470964298836

Table 1: Evaluation metrics comparison across quantized GPT-2 models on the AG News dataset.

Observation: Quantized models maintain near baseline accuracy while reducing memory usage. The NF4 variant achieves the best balance between accuracy, latency, and model size, while the custom quantized model performs relatively poor due to micro optimization that are missing.

Model Metrics Visualization

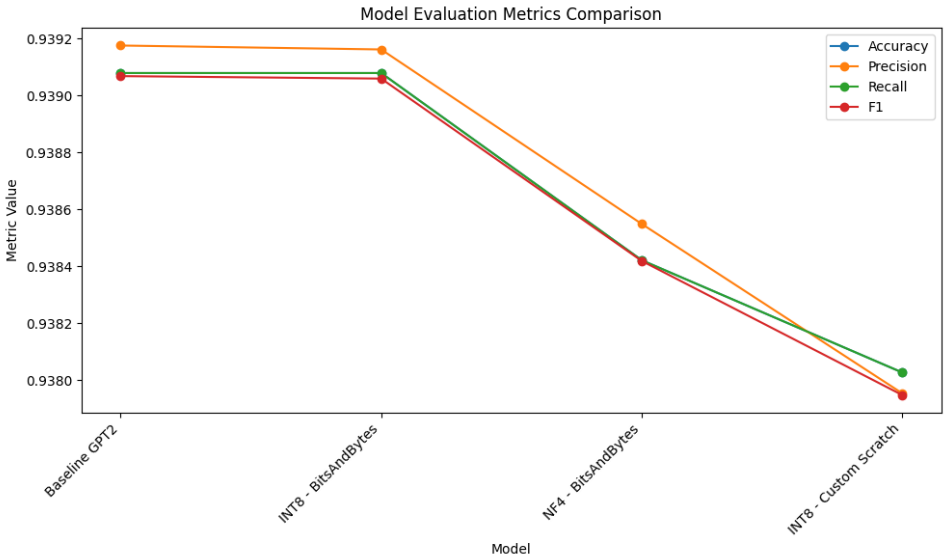
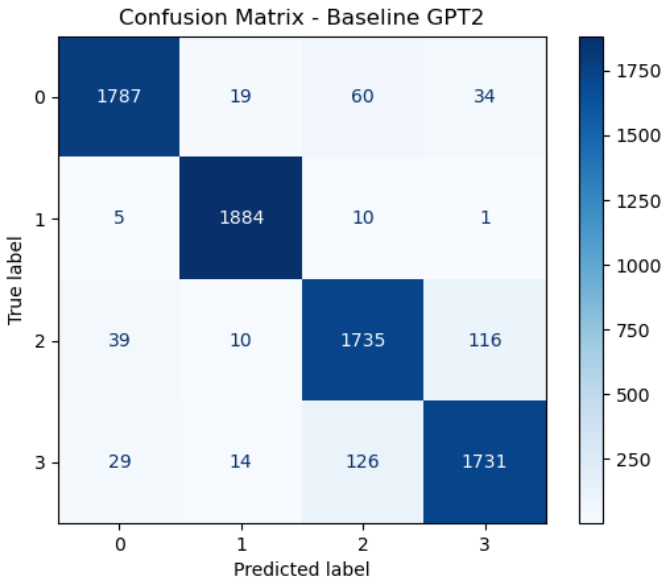


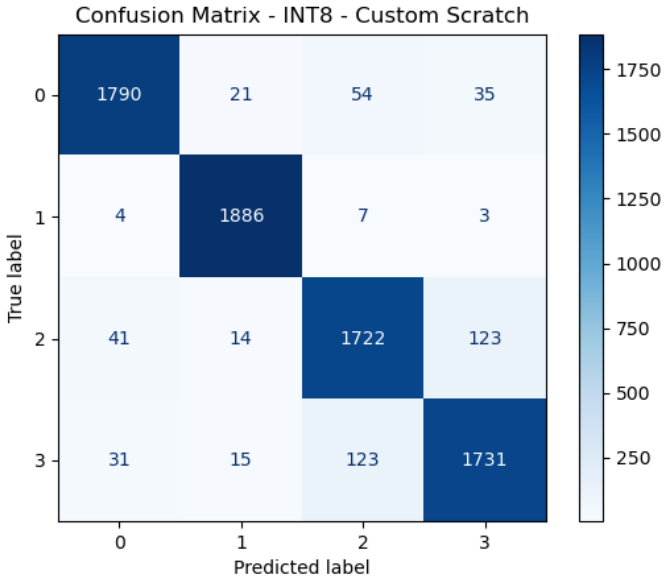
Figure 1: Comparison of Accuracy, Precision, Recall, and F1 across Models

Confusion Matrix Analysis

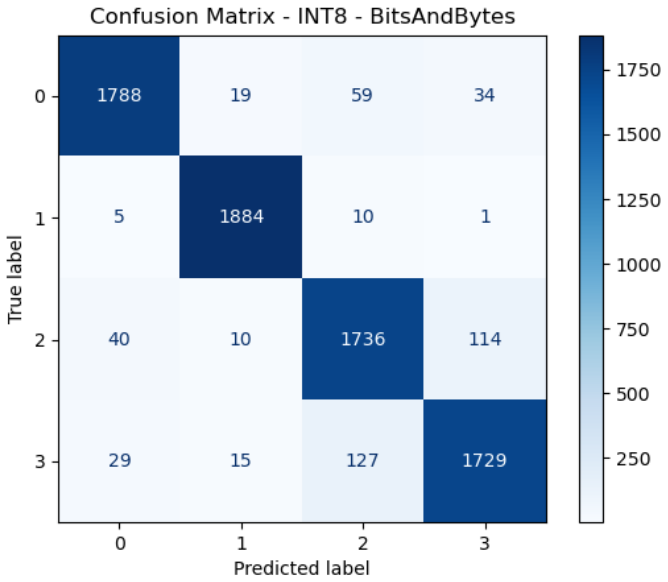
Baseline GPT-2



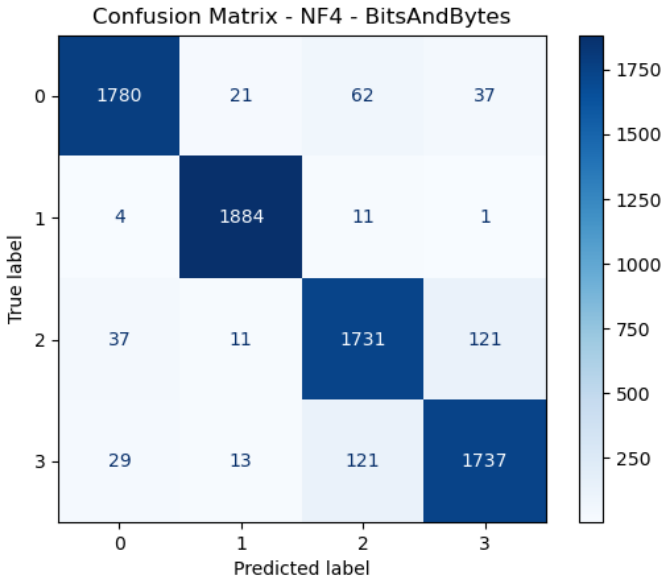
INT8 – Custom



INT8 – BitsAndBytes



NF4 – BitsAndBytes



Analysis

- The baseline model performs relatively well with very few misclassifications. But it consumes a lot of memory in comparison to quantized models.
- INT8 and NF4 reduce model memory by 4 times while having acceptable level of drop in accuracy and other metrics.

- Custom quantization functions lack the micro optimizations that are done in library quantization functions resulting in relaively poor performance.