Promp Length=2048, Subsequent Pass NPU: Op Distribution Over Model Run

- vitisaiep (MatMulNBits) - 36.83 %
- Transpose - 25.75 %
- Concat - 16.46 %
- Mul - 8.26 %
- MatMul - 6.01 %
- UNKNOWN--Start/Stop Overhead? - 2.96 %
- Slice - 0.96 %
- Gather - 0.54 %
- Add - 0.52 %
- Unsqueeze - 0.38 %
- Reshape - 0.25 %
- Shape - 0.21 %
- SimplifiedLayerNormalization - 0.19 %
- Div - 0.17 %
- Softmax - 0.12 %
- Neg - 0.11 %
- Sqrt - 0.10 %
- QuickGelu - 0.09 %
- Cast - 0.07 %
- Where - 0.01 %
- Expand - 0.00 %
- Sub - 0.00 %
- ConstantOfShape - 0.00 %
- Equal - 0.00 %
- Range - 0.00 %
- Less - 0.00 %
- Squeeze - 0.00 %