Promp Length=2048, Prefill Phase CPU: Op Distribution Over Model Run

- MatMulNBits - 80.86 %
- MatMul - 13.47 %
- Add - 1.82 %
- Softmax - 1.19 %
- Mul - 0.78 %
- Transpose - 0.63 %
- Concat - 0.34 %
- Slice - 0.29 %
- QuickGelu - 0.21 %
- SimplifiedLayerNormalization - 0.18 %
- Neg - 0.07 %
- UNKNOWN--Start/Stop Overhead? - 0.05 %
- Cast - 0.03 %
- Where - 0.03 %
- Gather - 0.01 %
- Unsqueeze - 0.00 %
- Reshape - 0.00 %
- Shape - 0.00 %
- Expand - 0.00 %
- Div - 0.00 %
- Sqrt - 0.00 %
- Sub - 0.00 %
- ConstantOfShape - 0.00 %
- Less - 0.00 %
- Equal - 0.00 %
- Range - 0.00 %
- Squeeze - 0.00 %