Training Step Distribution: 2 GPUs

- GEMM: 49.2%
- All-to-All: 5.8%
- Dropout: 7.9%
- LayerNorm: 0.8%
- Softmax: 8.1%
- Data Transfers: 1.0%
- Elementwise: 17.7%
- Misc: 9.5%