Plot 4: Model Overall Layer - FuseMax (Varying Seq Length, SRAM Constraint)

Legend:
- 2D PE Array Roof (Peak: 123.2 TFLOPs/s)
- BERT Overall (N_seq=256 (Short)) (2D)
- BERT Overall (N_seq=4096 (Long)) (2D)
- OPT Overall (N_seq=256 (Short)) (2D)
- OPT Overall (N_seq=4096 (Long)) (2D)
- Pythia Overall (N_seq=256 (Short)) (2D)
- Pythia Overall (N_seq=4096 (Long)) (2D)

X-axis: Operational Intensity (FLOPs/Byte)
Y-axis: Performance (GFLOPs/s)