Memory Scalability Across Model Sizes

Legend:
- FP16
- INT8
- INT4
- 3.5-bit

80 GB (A100)
24 GB (RTX 4090)

Y-axis: Memory Footprint (GB)
X-axis: Model Size (Billion Parameters)