Quality vs cold latency (top) and peak memory (bottom); all model pairs pooled

