1) Latency is normalized to the fastest data point of each model (i.e., CPU-GPU with batch 1024, the red box below)
2) Purpose of this graph is to show:
   a) Embedding-intensive RM1/2 has a smaller performance gap between CPU-only vs. CPU-GPU because the throughput-limited MLP portion is relatively small compared to MLP-intensive RM3/4
   b) End-to-end training time scales roughly proportional to batch size