Latency is normalized to the fastest data point of each model (i.e., CPU-GPU with batch 1024, the red box below) Purpose of this graph is to show: 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 End-to-end training time scales roughly proportional to batch size 32 16 8 4 0.5 CPU-only CPU-GPU b1024 | b2048 | b4096 | b8192 | b1024 | b2048 | b4096 | b8192 | b1024 | b2048 | b4096 | b8192 | b1024 | b2048 | b4096 | b8192 |

RM3

RM4

MLP intensive

RM2

Latency (normalized)

RM1

Embedding intensive