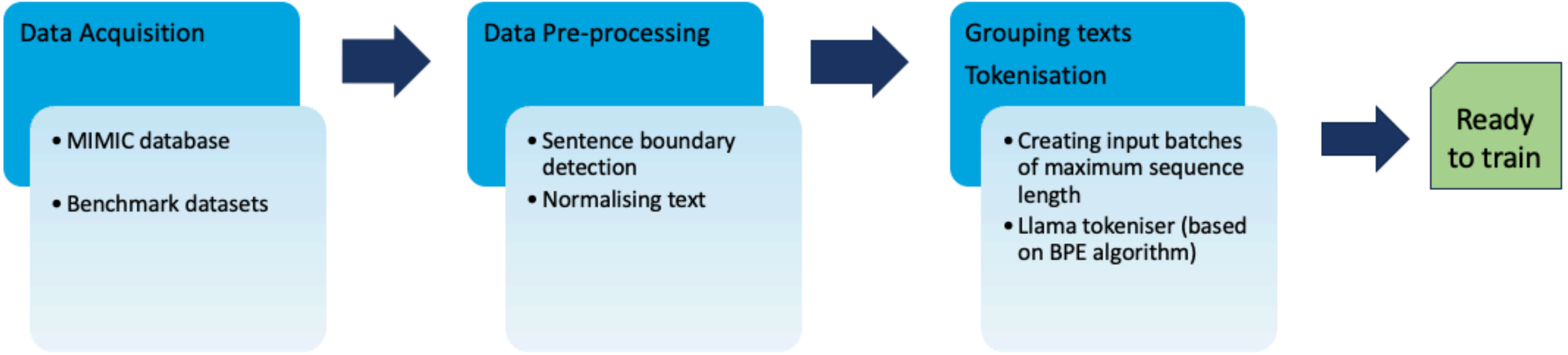
Data Acquisition
• MIMIC database
• Benchmark datasets
Data Pre-processing
• Sentence boundary detection
• Normalising text
Grouping texts
Tokenisation
• Creating input batches of maximum sequence length
• Llama tokeniser (based on BPE algorithm)
Ready to train