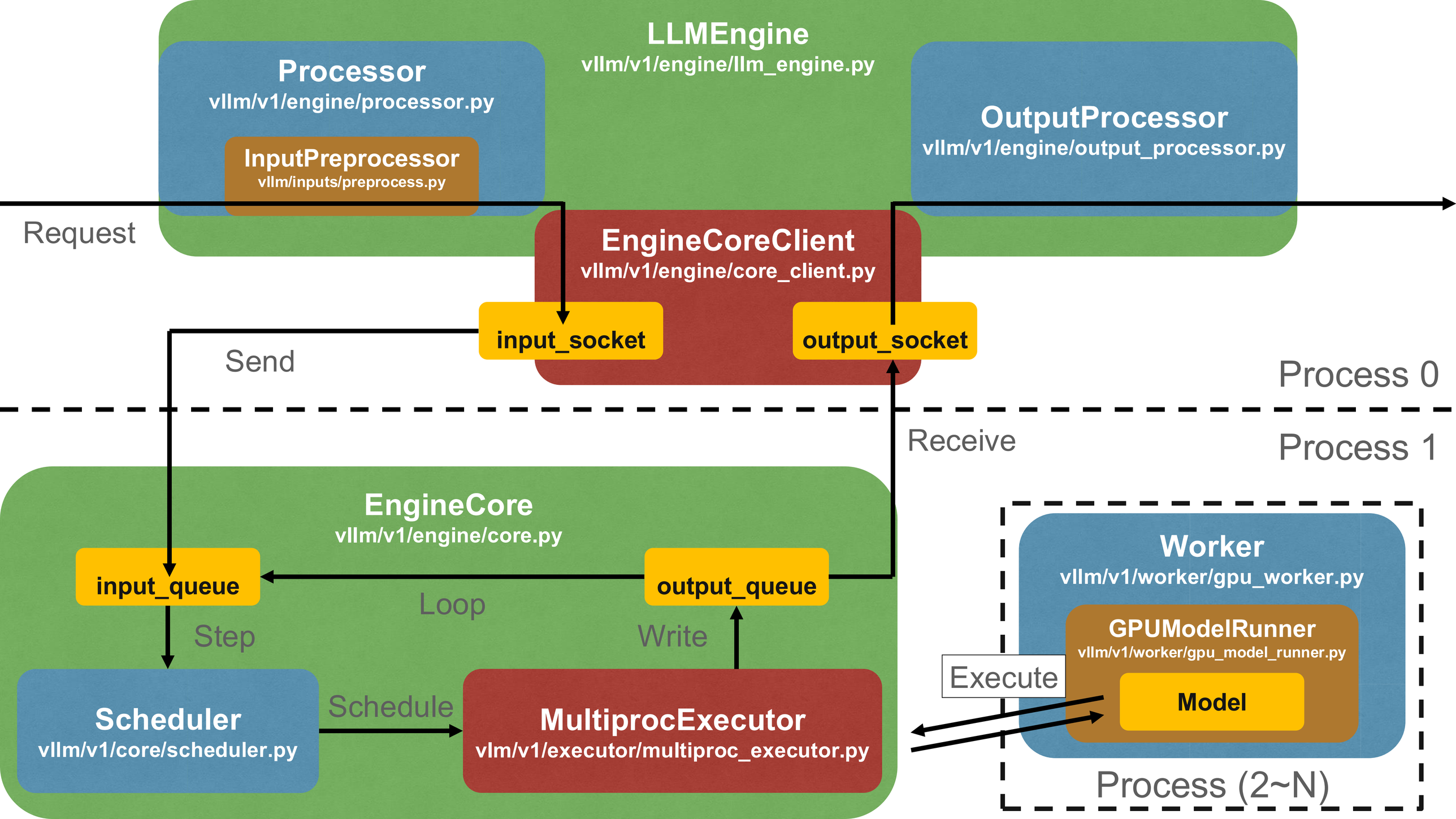
LLMEngine
vllm/v1/engine/llm_engine.py
Processor
vllm/v1/engine/processor.py
InputPreprocessor
vllm/inputs/preprocess.py
OutputProcessor
vllm/v1/engine/output_processor.py
Request
EngineCoreClient
vllm/v1/engine/core_client.py
input_socket
output_socket
Send
Process 0
Receive
Process 1
EngineCore
vllm/v1/engine/core.py
input_queue
output_queue
Loop
Step
Write
Scheduler
vllm/v1/core/scheduler.py
Schedule
MultiprocExecutor
vlm/v1/executor/multiproc_executor.py
Worker
vllm/v1/worker/gpu_worker.py
GPUModelRunner
vllm/v1/worker/gpu_model_runner.py
Execute
Model
Process (2~N)