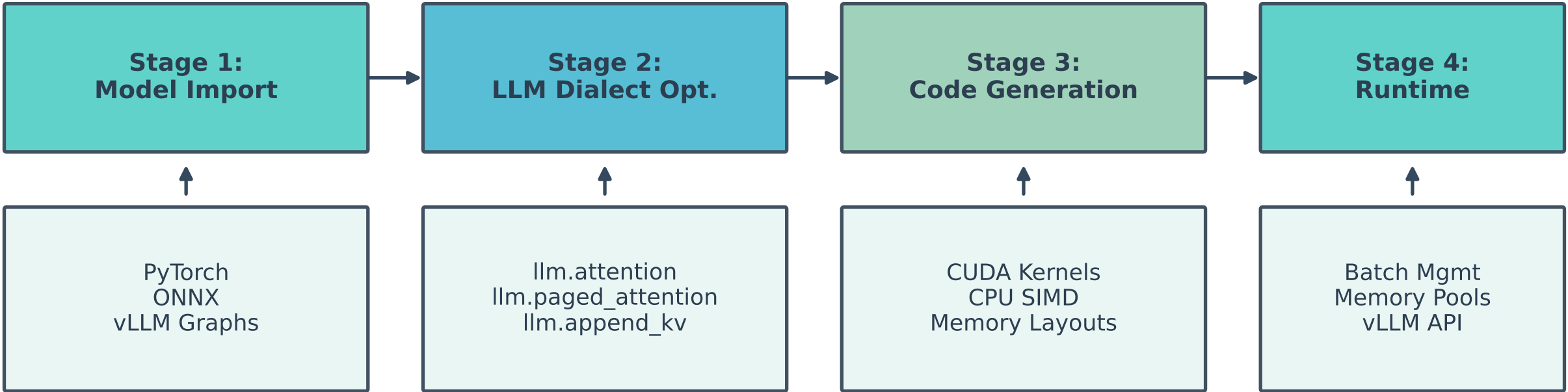


LLMIR Compilation Pipeline



**Compile-Time: Block allocation,
Kernel selection, Sharing detection**

**Runtime: Dynamic growth,
Parameter binding, Communication**

Performance: 22.4% over vLLM | 37.8% over SGLang | Perplexity within 0.1%