Kani in the LM Stack

Where does it fit in the ecosystem?

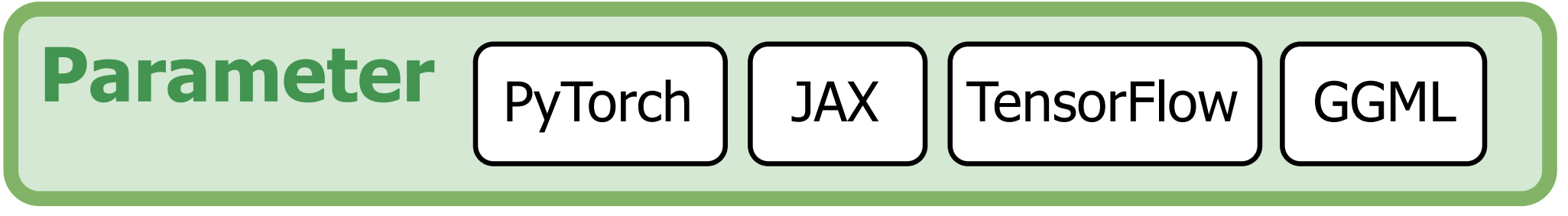
- Four main layers of LM software

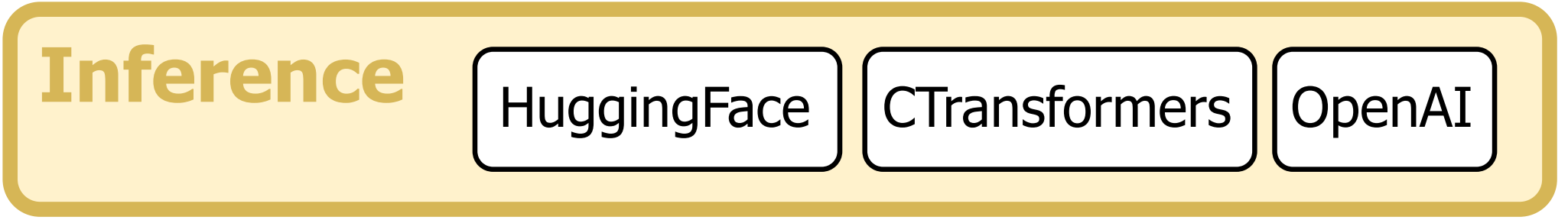
- **Parameter**: low-level weights, physical devices

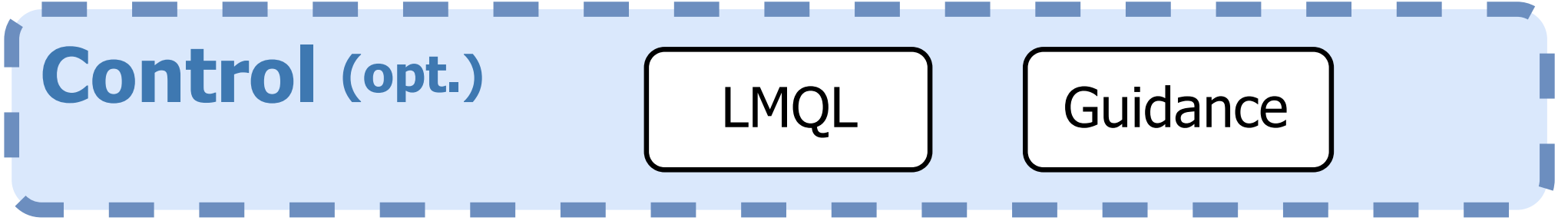
- **Inference**: sampling strategies, tokenization

- **Control** (optional): guided decoding, dynamic prompt branching

- **Application**: chat history management, tool usage


Parameter
PyTorch
JAX
TensorFlow
GGML


Inference
HuggingFace
CTransformers
OpenAI

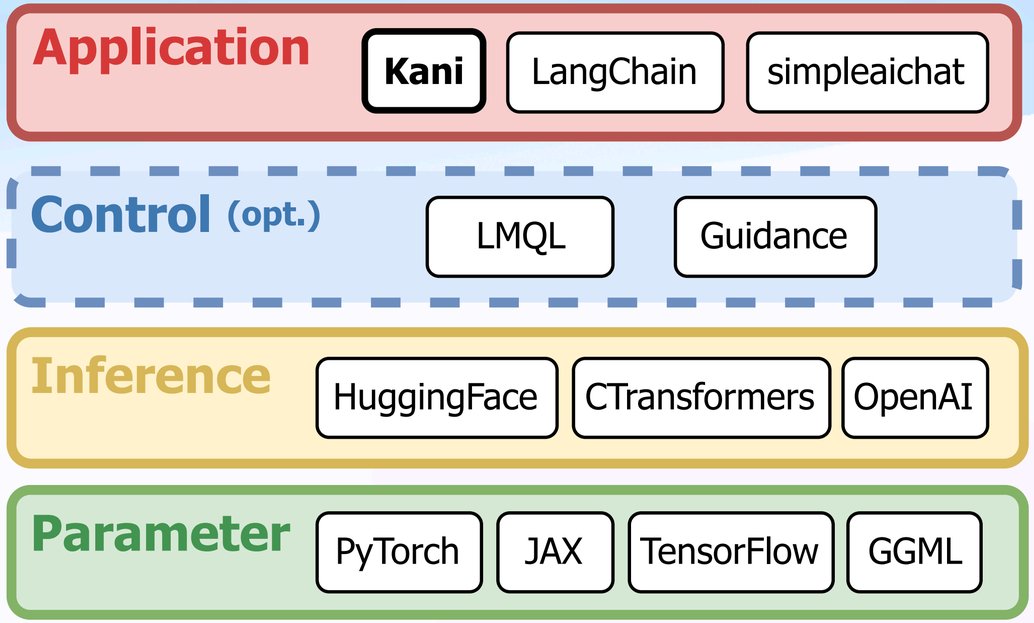

Control (opt.)
LMQL
Guidance


Application
Kani
LangChain
simpleaichat

Kani in the LM Stack

Where does it fit in the ecosystem?

- Four main layers of LM software
- **Parameter**: low-level weights, physical devices
- **Inference**: sampling strategies, tokenization
- **Control** (optional): guided decoding, dynamic prompt branching
- **Application**: chat history management, tool usage



Features & Framework Comparison