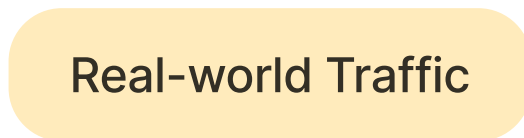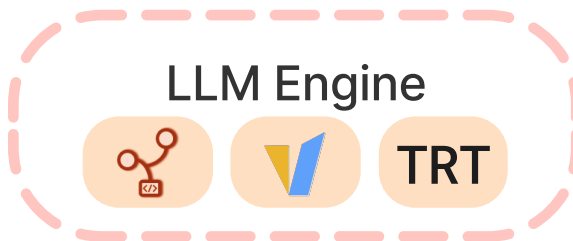Open-source Models + LLM Engine

Real-world Traffic + Real-world Serving Config

Kernels
- Attention
- GEMM
- Normalization
- . . .

Workloads
- Axes
- Input Tensors