LangChain with Offline LLMs

Setup:

1. Download Llama-2-7b-hf model (~7 GB)
2. Use mlc_llm.MLCEngine(model="Llama-2-7b-hf")

Create chains:

- from_llm() works with local models
- Chain: Retriever -> LLM -> Formatter

Performance:

- Batch processing faster
- Memory: ~16 GB for 7B model + context

Example working code available

Ready to integrate with pipeline