HippoLLM: scrutable and robust memory for LLMs with hybrid graph-vector databases

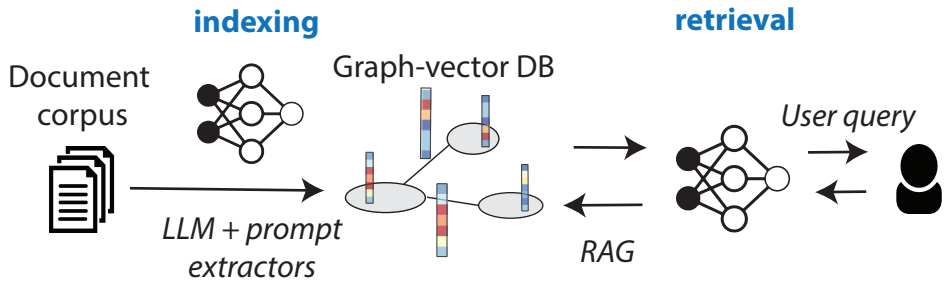
Adrian Valente - adrian.valente@ens.fr

Introduction

* LLMs and KG offer complemetary profiles:
  LLMs: flexible but not interpretable and unreliable knowledge
  KGs: interpretable & controllable source of knowledge but too rigid

* Current solutions for knowledge management in LLMs:
  fine-tuning: no idea if knowledge acquired, hard to modify
  RAG: expensive queries and still lacks interpretability

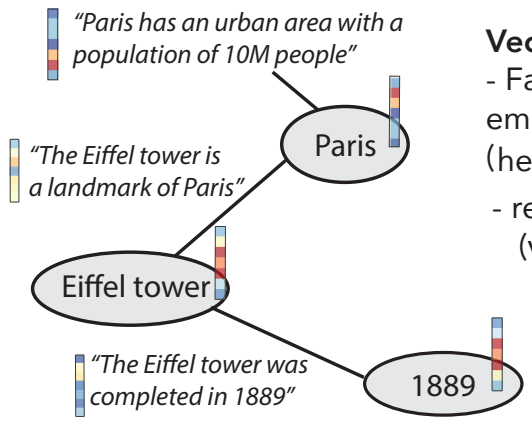
* Project idea: let the LLM manage its own memory freely!
  - memory as a KG (for human interpretability)
  - no-schema KG (only setences and embeddings)
  - minimalistic zero-shot approach (relies on LLM's emergent capacities)
  - use max 7B model



Memory: hybrid graph-vector DB

Description:
The graph composed of Entities (nodes) and Facts (n-ary relations). Facts are only *natural language sentences*. They are linked to their relevant entities upon creation, but do not belong to classes.

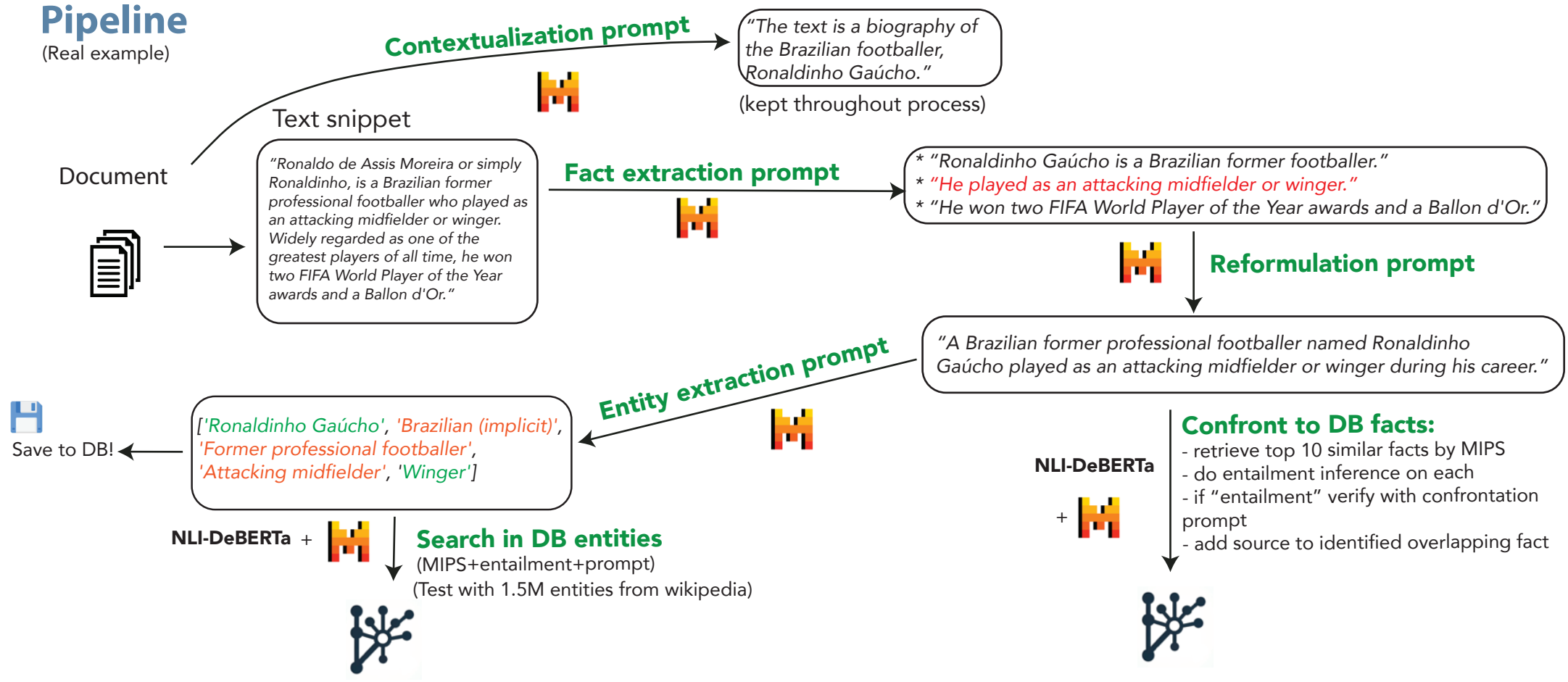
Data model:
- Entity: (id (str), description (str), embedding (vec), list[Fact])
- Fact: (sentence (str), embedding (vec), list[Entity], list[Source])



Vector component
- Facts and entities (with descriptions) embedded with SBERT model (here "all-MiniLM-L6-v2" (384d))
- retrieval managed by ChromaDB (vector database)

Pipeline

(Real example)



Results and further notes

Current tradeoffs
- Extraction of facts:
  - little hallucinations, often longer than atomic facts
  - reformulation necessary to create standalone facts
  - reformulation tends to add unnecessary information, although no hallucination has been observed.
- Confrontation to DB facts:
  - NLI alone or prompt alone not sufficient.
  - prompt with guidance or fine-tuned classifier could improve result
- Entity extraction (tested with 1.5B entities with low-quality abstracts):
  - Works well, rarely related entities are added, as well as spurious comments
  - Need to test NER model instead
  - Arcane synonyms still very difficult (could be added to descriptions)

Ideas to test:
- Mix with already existing schema-structure KGs by embedding their facts
- Running "bookeeping" operations to eventually merge entities
- Possibility to detect contradictory sources

Limitations:
- use of 7B model leads to inaccuracies (ex: too long sentences, or not completely standalone) but surprisingly good for amount of compute.
- difficulties in managing arcane synonymous entities (like surnames)
- tends to propose too many entities (positive phenomenon)

Resources:
ChromaDB, langchain, sentence-transformers, Mistral (7B)