Goal: AI systems

- Agent
  - learning
  - reasoning
  - acting
  - Autonomous

Available

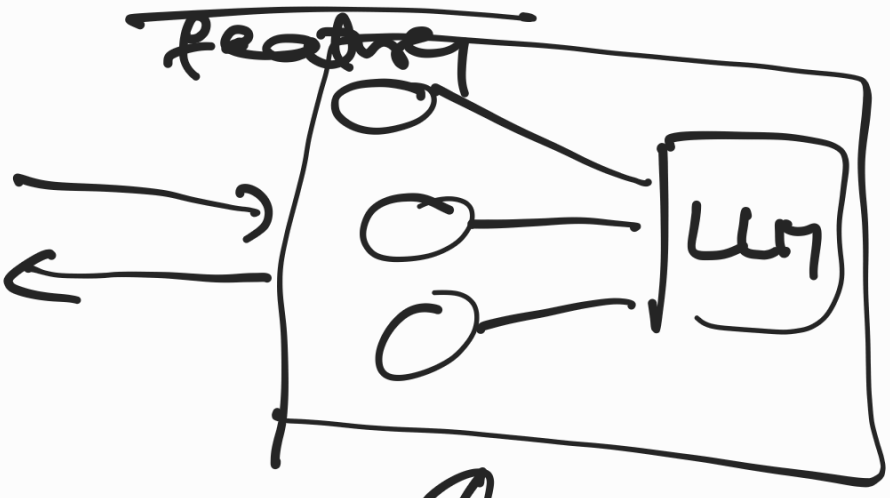
LLM
LMM
LRM → limitations

- ...
- memory
- tools
- kb
- hallucination

~~LRM~~

GenAI engs

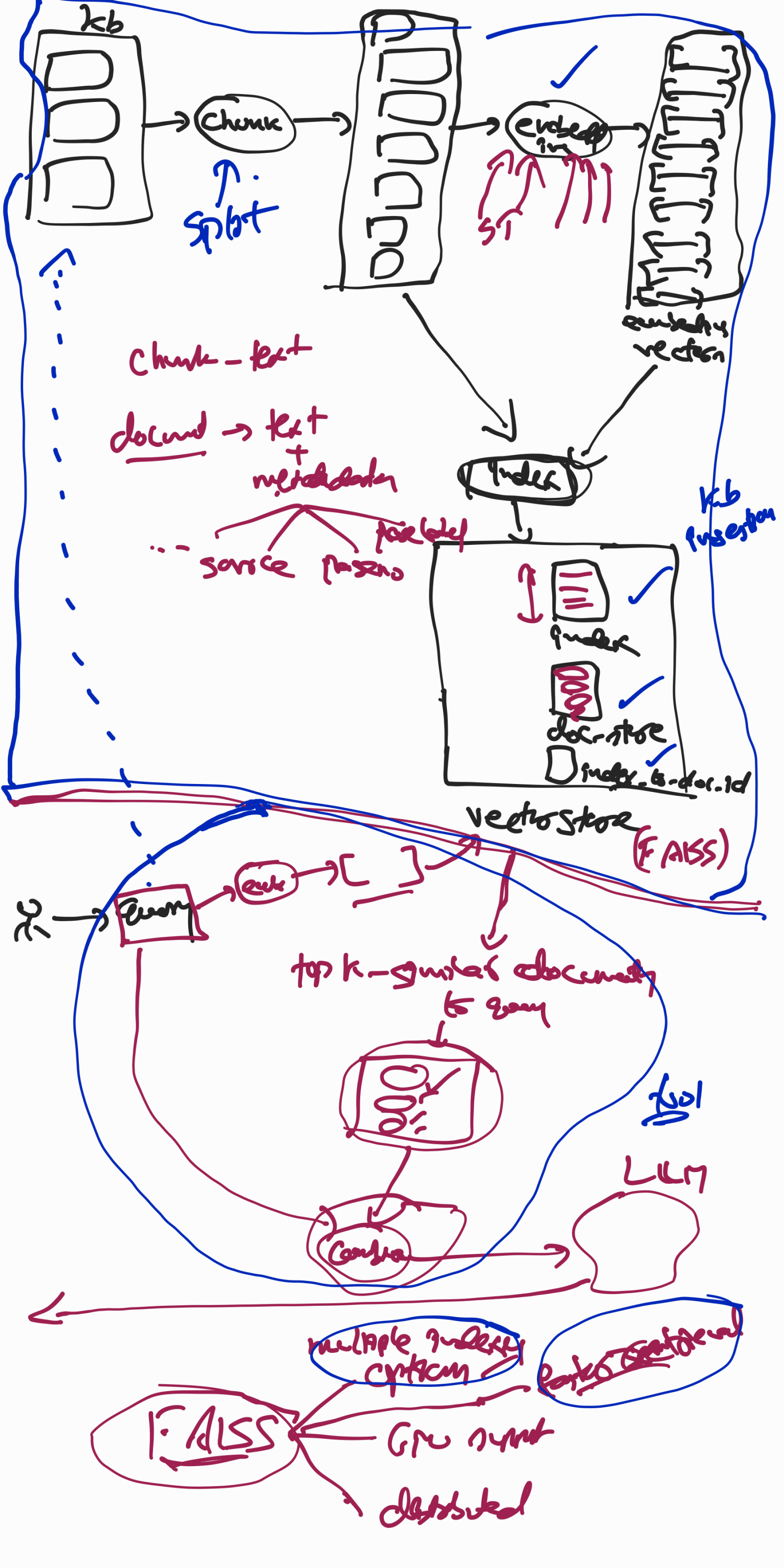
feature

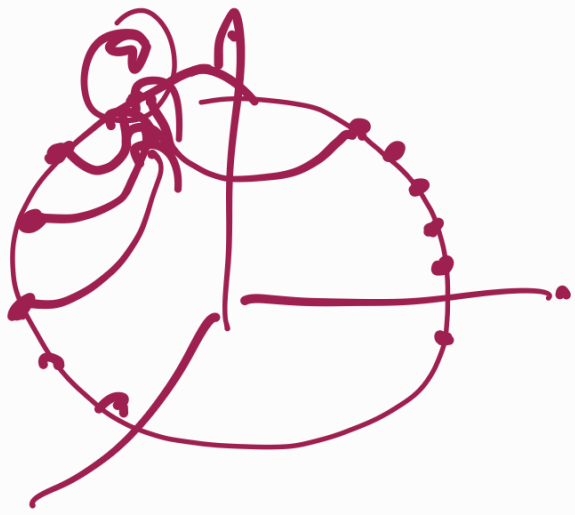


temperature — degree of creativity

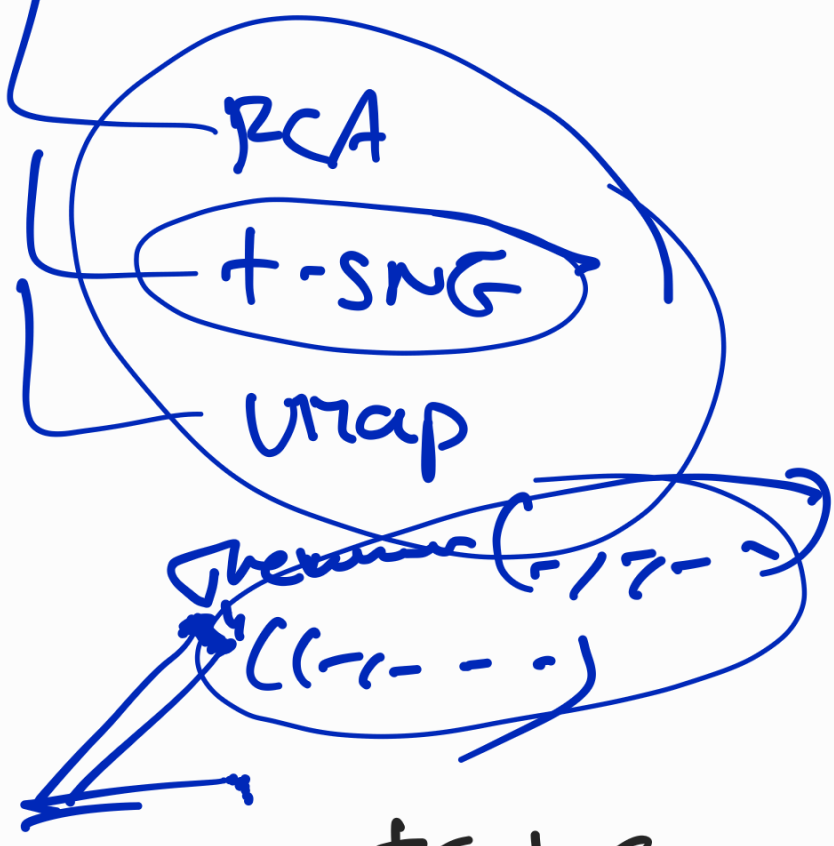
RAG Name

Issue: entire knowledge base needs to be passed, which is impractical


kb
chunk
split
embedding
ST
embedding vectors
chunk-text
document → text + metadata
source
page label
index
kb insertion
index
doc-store
index-to-doc-id
vectorstore (FAISS)
query
top k-similar documents to query
tool
LLM
multiple index options
FAISS
GPU support
distributed



Dimensionality Reduction



vectorize