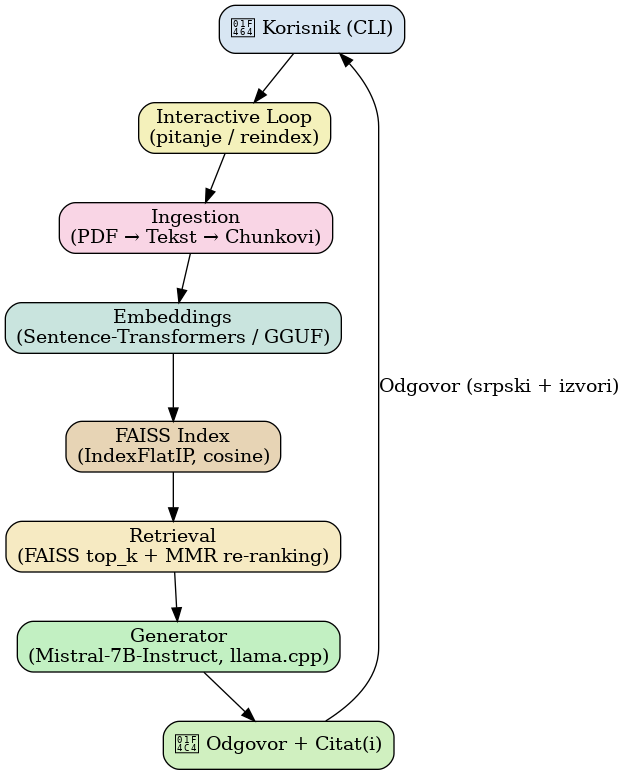
# RAG Baseline – Arhitektura, Objašnjenje i Plan Evaluacije



Slika 1. Arhitektura sistema zasnovanog na RAG pristupu.

## 1. Ingestion

PDF dokumenti se učitavaju koristeći biblioteku pypdf. Tekst se čisti uklanjanjem nevidljivih znakova, mekih crtica i višestrukih praznina. Rečenice se segmentiraju pomoću jednostavnog regularnog izraza, a zatim se spajaju u chunkove od oko 900 karaktera sa preklapanjem od 120 karaktera. Svaki chunk dobija metapodatke (naziv fajla, broj strane, indeks chunk-a) koji se kasnije koriste za citiranje i analizu relevantnosti.

## 2. Embeddings

Za vektorizaciju teksta koristi se Sentence-Transformers model ('all-MiniLM-L6-v2'), dimenzije 384. Alternativno, može se koristiti lokalni GGUF embedder kroz 'llama\_cpp' za potpuno offline rad. Embeddings se čuvaju u FAISS indeksu koristeći normalizovane vektore (cosine similarity).

## 3. Retrieval

Na upit korisnika, sistem prvo kreira embedding pitanja i pretražuje FAISS indeks radi pronalaženja top\_k (8) sličnih chunkova. Zatim se primenjuje MMR re-ranking (k=4, λ=0.55) radi postizanja ravnoteže između relevantnosti i raznolikosti. Konteksti se ubacuju u prompt kao ulaz u model.

## 4. Generator

Za generaciju odgovora koristi se lokalni LLM u GGUF formatu (npr. 'Mistral-7B-Instruct-v0.2.Q4\_K\_M.gguf') preko 'llama\_cpp'. Prompt koristi [INST] format i forsira odgovore na srpskom jeziku u formi sažetih objašnjenja uz navođenje izvora. Parametri generacije uključuju temperature=0.15, top\_p=0.9 i repeat\_penalty=1.05.

## 5. Plan Evaluacije

Evaluacija sistema se sprovodi kroz tri grupe metrika:  
• Retrieval metrike: Recall@k, MRR@k, nDCG@k  
• Generacione metrike: Faithfulness (groundedness), Exact Match, broj citata  
• Performanse: prosečno vreme generacije, tokens/sec, zauzeće memorije.  
  
Test set bi trebao da sadrži 20-30-ak primera pitanja sa očekivanim odgovorima i stranicama iz zakona (citati).

## 6. Preporuke i Dalji Rad

• Obezbediti konzistentnost između embeddera i modela (isti model za index i query fazu).  
• Testirati različite veličine chunkova (700–1200) i preklapanja.  
• Dodati automatski evaluacionu skriptu sa batch testom i metrikama.  
• Po potrebi integrisati Qwen2.5 ili LLaMA 3 GGUF za poređenje.