

SatuChat

Panduan Operasional Chatbot RAG
Kementerian Keuangan RI — BATII

Apa itu SatuChat?

Chatbot internal berbasis **Retrieval-Augmented Generation (RAG)** untuk menjawab pertanyaan berdasarkan dokumen resmi unit kerja.

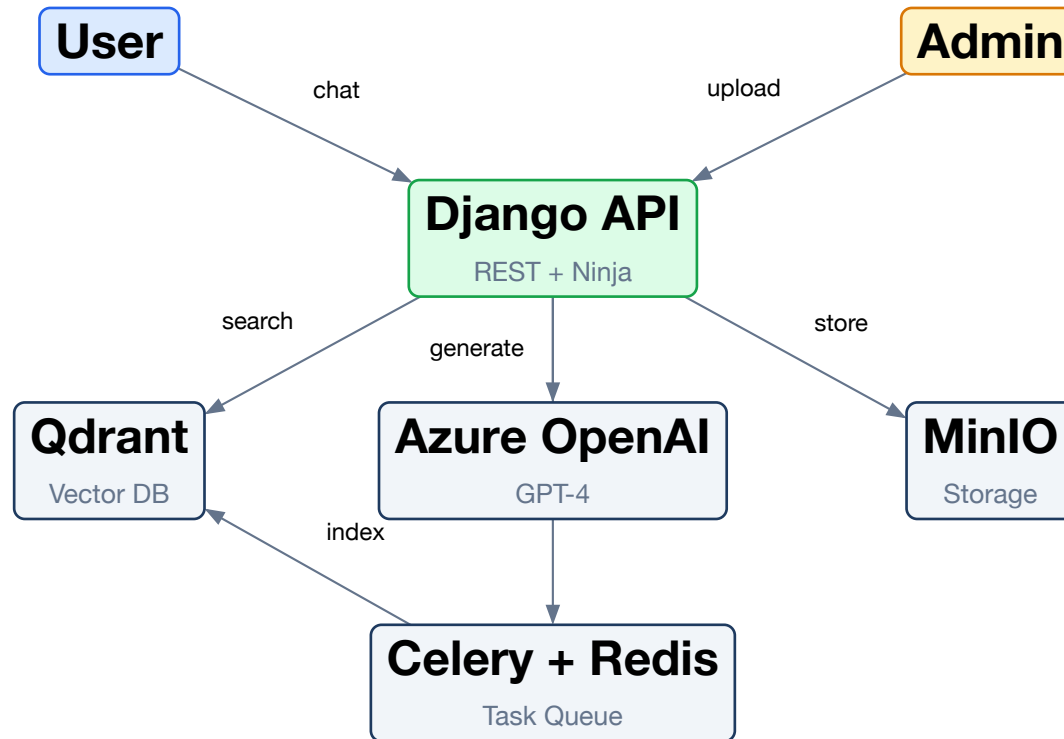
Domain: `satuchat.kemenkeu.go.id`

Stack: Django · PostgreSQL · Qdrant · Azure OpenAI

Fitur Utama

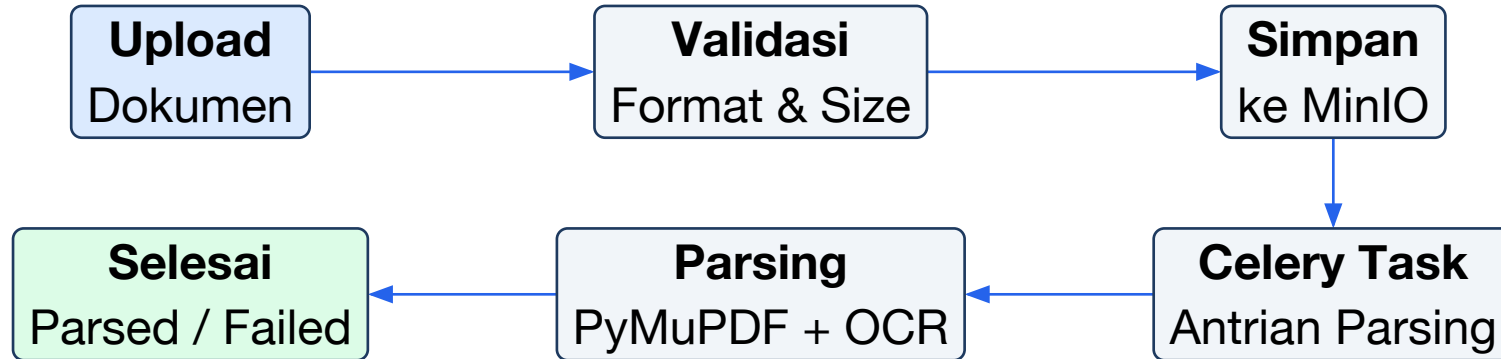
- Chat RAG (sync + streaming)
- Multi-tenant per unit kerja
- Manajemen dokumen & approval
- Semantic & exact cache
- Role-based access control

Arsitektur Sistem



Pengelolaan Dokumen

Alur Upload & Parsing

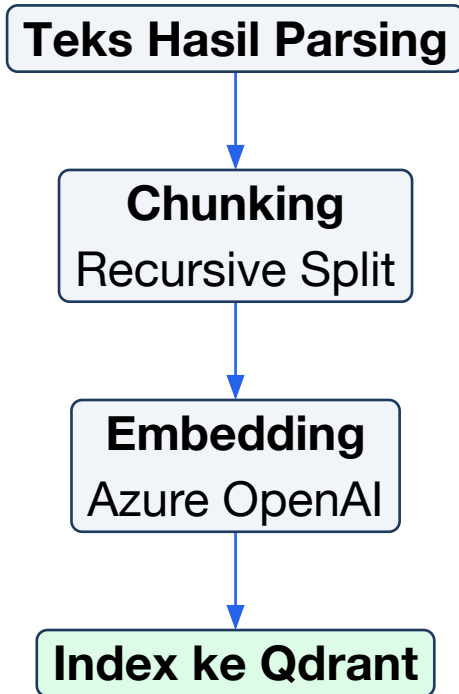


Format: **PDF** **DOCX** **PPTX** **XLSX**

Parser: PyMuPDF (teks) + Tesseract (OCR)

Deduplikasi via hash

Proses Chunking



Strategi Chunking

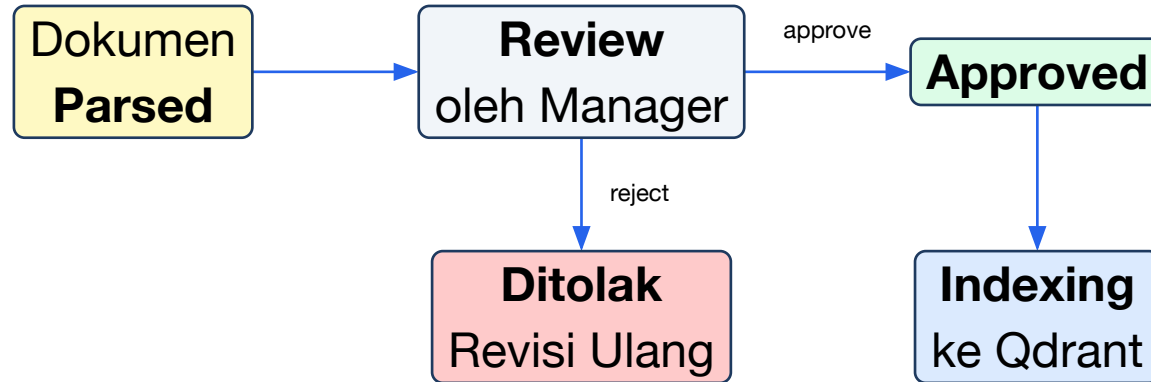
- Recursive text splitting
- Overlap antar chunk
- Metadata per chunk:

<code>tenant_id</code>	<code>collection_id</code>
<code>classification</code>	<code>group_id</code>

Isolasi Multi-Tenant

Setiap chunk dilabeli `tenant_id` sehingga data antar unit kerja **tidak bisa saling akses**.

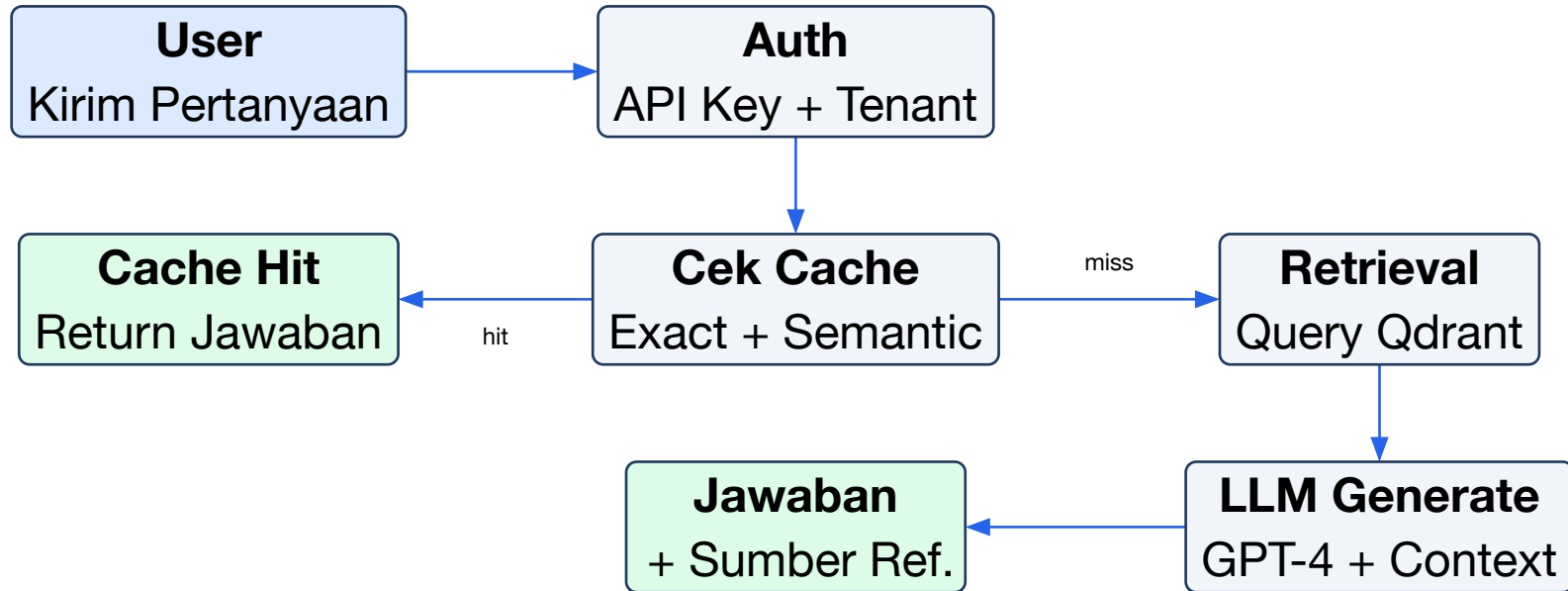
Review & Approval



Manager wajib mereview konten sebelum dokumen di-index ke vector database.

Alur Chat RAG

Alur Chat Pengguna



Detail Retrieval

Vector Search

1. Query di-embed via Azure OpenAI
2. Search Qdrant dengan filter:
 - `tenant_id = tenant user`
 - `classification = public`
atau `group_id` cocok
3. Top-K chunk relevan diambil
4. Chunk menjadi context LLM

Keamanan Data

- Tenant A **tidak bisa** akses data Tenant B
- Filter `tenant_id` wajib di setiap query
- Dokumen private hanya untuk grup tertentu
- Dokumen public diakses semua user dalam tenant

Isolasi data per unit kerja terjamin.

Strategi Caching

Exact Cache — Redis

- Key: hash dari query + tenant
- Hit jika query **persis sama**
- Response time: < **50ms**

Semantic Cache — Qdrant

- Embed query lalu cari kemiripan
- Hit jika similarity > threshold
- Hemat token LLM **hingga 70%**



Role & Hak Akses

Fitur	Admin	Manager	User
Kelola tenant & settings	Ya	–	–
Upload dokumen	Ya	Ya	–
Review & approve	Ya	Ya	–
Kelola koleksi	Ya	Ya	–
Chat dengan AI	Ya	Ya	Ya
Lihat riwayat chat	Ya	Ya	Ya
Feedback jawaban	Ya	Ya	Ya
Monitoring & dashboard	Ya	–	–

Monitoring & Proteksi

Rate Limiting

Batasi jumlah request per user per periode waktu untuk mencegah abuse.

Token Quota

Setiap tenant memiliki kuota token. Mencegah penggunaan LLM berlebihan.

Circuit Breaker

Jika Azure OpenAI down, circuit breaker aktif dan fallback ke cache.

Dashboard Monitoring

Total chat / hari

Token usage / tenant

Cache hit rate

Dokumen status

Ringkasan

Dokumen

Upload → Parse → Review → Index

Chunking

Split → Embed → Store di Qdrant

Chat

Query → Cache → Retrieval → LLM

Keamanan

Multi-tenant · RBAC · Isolasi Data