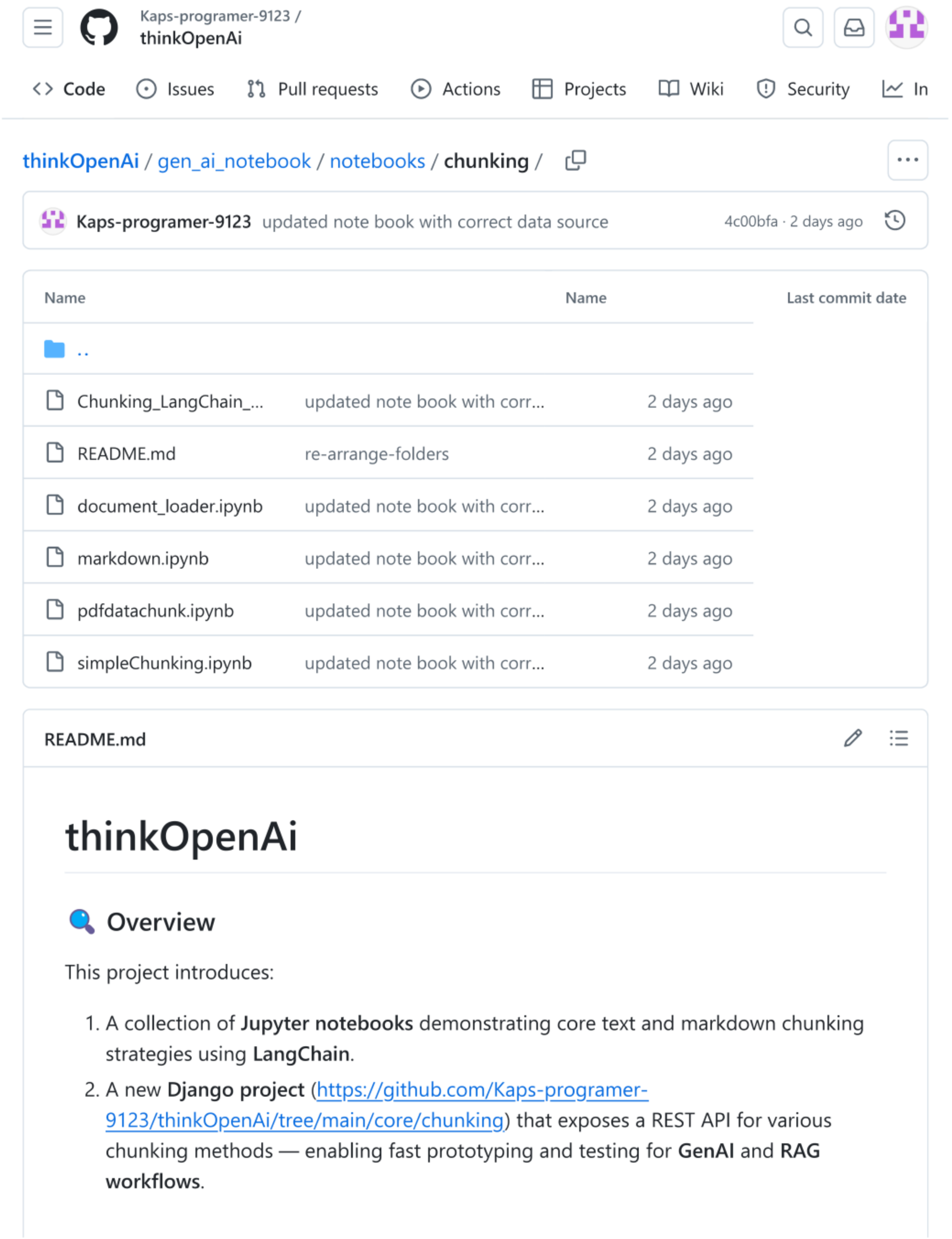
Chunking-data

Thursday, May 8, 2025 11:21 AM



thinkOpenAi / gen_ai_notebook / notebooks / chunking /

Kaps-programer-9123 updated note book with correct data source 4c00bfa · 2 days ago

Name	Name	Last commit date
..		
Chunking_LangChain_...	updated note book with corr...	2 days ago
README.md	re-arrange-folders	2 days ago
document_loader.ipynb	updated note book with corr...	2 days ago
markdown.ipynb	updated note book with corr...	2 days ago
pdfdatachunk.ipynb	updated note book with corr...	2 days ago
simpleChunking.ipynb	updated note book with corr...	2 days ago

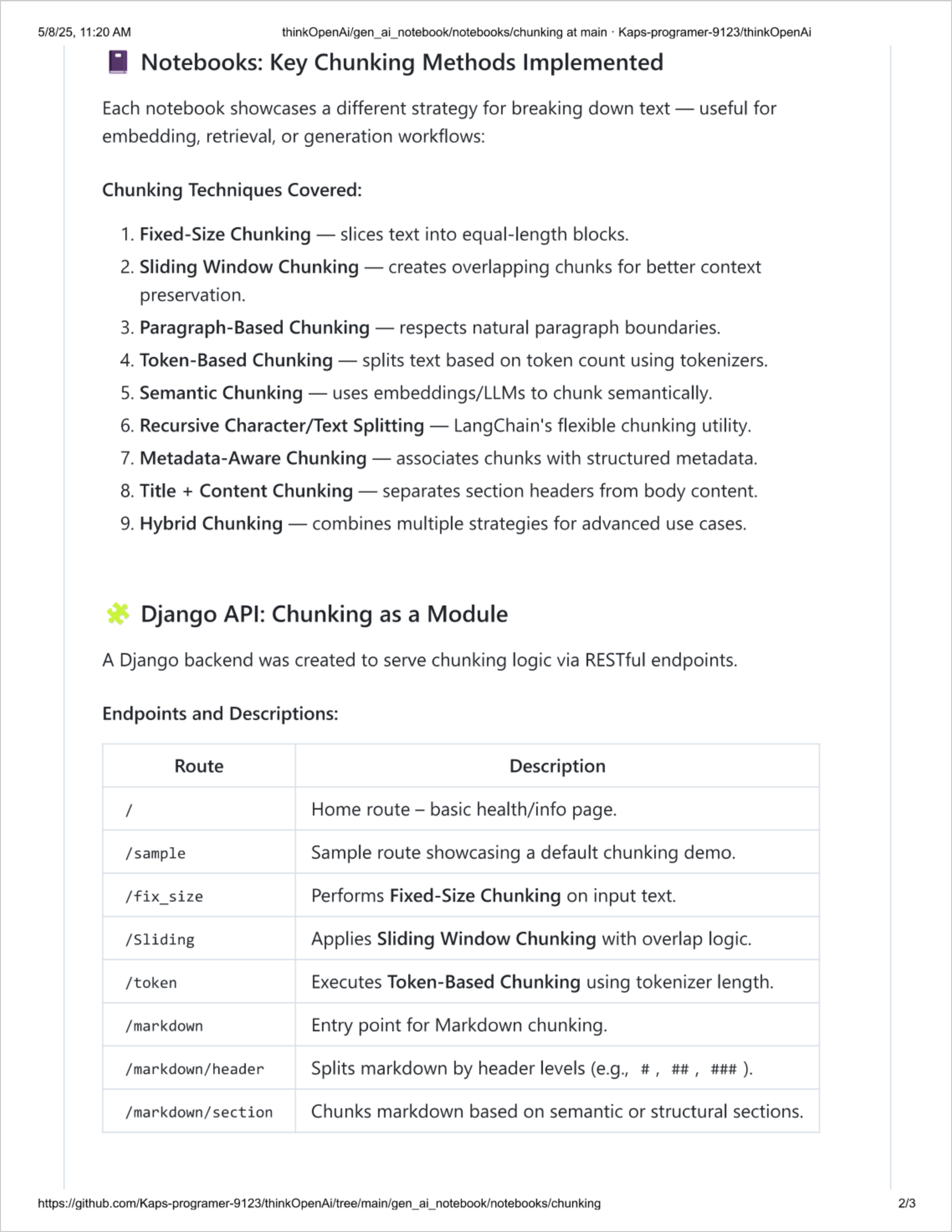
README.md

thinkOpenAi

🔍 Overview

This project introduces:

1. A collection of **Jupyter notebooks** demonstrating core text and markdown chunking strategies using **LangChain**.
2. A new **Django project** (https://github.com/Kaps-programer-9123/thinkOpenAi/tree/main/core/chunking) that exposes a REST API for various chunking methods — enabling fast prototyping and testing for **GenAI** and **RAG workflows**.



📕 Notebooks: Key Chunking Methods Implemented

Each notebook showcases a different strategy for breaking down text — useful for embedding, retrieval, or generation workflows:

Chunking Techniques Covered:

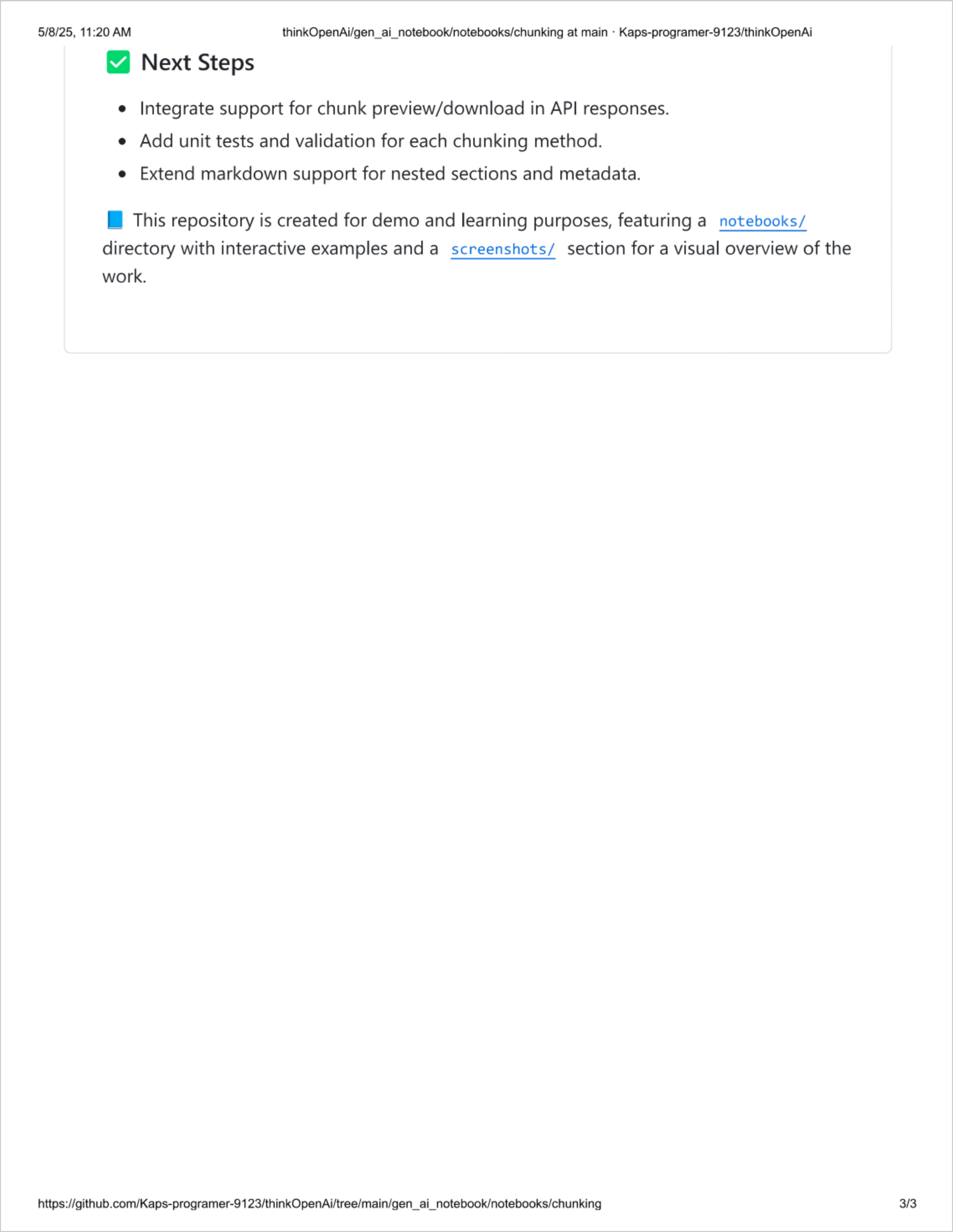
1. **Fixed-Size Chunking** — slices text into equal-length blocks.
2. **Sliding Window Chunking** — creates overlapping chunks for better context preservation.
3. **Paragraph-Based Chunking** — respects natural paragraph boundaries.
4. **Token-Based Chunking** — splits text based on token count using tokenizers.
5. **Semantic Chunking** — uses embeddings/LLMs to chunk semantically.
6. **Recursive Character/Text Splitting** — LangChain's flexible chunking utility.
7. **Metadata-Aware Chunking** — associates chunks with structured metadata.
8. **Title + Content Chunking** — separates section headers from body content.
9. **Hybrid Chunking** — combines multiple strategies for advanced use cases.

🧩 Django API: Chunking as a Module

A Django backend was created to serve chunking logic via RESTful endpoints.

Endpoints and Descriptions:

Route	Description
`/`	Home route – basic health/info page.
`/sample`	Sample route showcasing a default chunking demo.
`/fix_size`	Performs **Fixed-Size Chunking** on input text.
`/Sliding`	Applies **Sliding Window Chunking** with overlap logic.
`/token`	Executes **Token-Based Chunking** using tokenizer length.
`/markdown`	Entry point for Markdown chunking.
`/markdown/header`	Splits markdown by header levels (e.g., `#`, `##`, `###`).
`/markdown/section`	Chunks markdown based on semantic or structural sections.



✅ Next Steps

- Integrate support for chunk preview/download in API responses.
- Add unit tests and validation for each chunking method.
- Extend markdown support for nested sections and metadata.

📘 This repository is created for demo and learning purposes, featuring a `notebooks/` directory with interactive examples and a `screenshots/` section for a visual overview of the work.