ClassMind — AI Study Assistant (One-Page Recruiter Summary)

What it is

ClassMind is a lightweight, open-source tool that turns university lecture slides (PDF/PPTX) into **actionable study assets** in minutes: per-slide and deck summaries, a **Mermaid mind map** of key topics, **semantic search** across content, and **Anki-ready flashcards**. It runs entirely on **free, open-source models in Google Colab** (no paid API keys).

Why it matters

Students and new hires lose hours skimming dense decks. ClassMind automates the heavy lifting—**summarizing, organizing, and enabling recall**—so they can focus on understanding, not formatting. It also demonstrates practical, end-to-end ML engineering: ingestion → NLP → retrieval → UX-friendly outputs.

My role

I designed and built the end-to-end pipeline, from data extraction and model selection to evaluation and packaging for one-click use in Colab/GitHub.

Outcomes (MVP)

- **Input:** Any PDF/PPTX lecture deck (tested on 38-slide real course deck).
- **Outputs (saved to outputs/):**
    - deck_summary.txt — concise 200–300 word overview
    - slide_summaries.json — per-slide text + summaries
    - mindmap.mmd — Mermaid mind map (copy to mermaid.live for PNG/SVG)
    - flashcards.csv — 20–100+ Q/A pairs importable to Anki
    - search(query) — returns top matching slides with scores
- **Runtime:** Works on Colab free tier (CPU/GPU optional).

- **No secrets:** Runs without paid tokens; optional OCR for scanned decks.

How it works (architecture)

1. **Ingest** — PyMuPDF (PDF) / python-pptx (PPTX) with **OCR fallback** via Tesseract (eng+deu) for image-heavy slides.

2. **Summarize** — Multilingual mT5 (XLSum) for per-slide and deck summaries. A **grounding prompt + chunking** prevents hallucinations.

3. **Embed & Search** — sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 with cosine similarity for fast, language-agnostic semantic search.

4. **Mind Map** — Keyword extraction → **Mermaid** mind-map code linking topics to slide indices.

5. **Flashcards** — Heuristics convert summaries/bullets into **Anki CSV** (DE/EN detection for localized prompts).

6. **Packaging** — Colab notebook + GitHub badge → **one-click run**; artifacts downloadable for sharing.

Highlights & Engineering Decisions

- **Grounded summarization**: Token-aware chunking + "use-only-source" instructions + post-filtering reduce off-topic generations.

- **Multilingual by design**: OCR eng+deu, multilingual summarizer and embeddings; handles mixed DE/EN decks.

- **Resilient ingestion**: Automatically marks image_only/ocr_used slides for transparency.

- **User-centric outputs**: Mind map for structure, Anki cards for recall, JSON for downstream automation.

- **Sane defaults, simple UX**: "Upload → Run All → Download outputs" in Colab; no infra needed.

Tech Stack

Python • Transformers (mT5) • Sentence-Transformers (MiniLM) • PyMuPDF • python-pptx • Tesseract OCR • scikit-learn • pandas • Mermaid

Impact & Example Metrics

- **Time saved**: From ~1–2 hrs of manual reviewing to **minutes** for a 30–40 slide deck.
- **Coverage**: 100% of slides summarized; **20–100+** flashcards auto-generated depending on content density.
- **Search quality**: Retrieves the correct slide for queries like “Markov property” or “SVM margin” with top-5 hit rates observed in testing.

Roadmap (optional stretch)

- **RAG Q&A**: Retrieve top-k slide chunks, answer grounded questions.
- **Streamlit UI**: Drag-and-drop web app; export buttons for all artifacts.
- **Docker image**: Reproducible local runs and classroom deployment.
- **Quality knobs**: Domain glossaries, better keyphrase clustering for mind maps, spaced-repetition tuning for flashcards.

Takeaway: ClassMind shows real, production-minded ML/NLP skills—building a complete, multilingual study assistant that turns raw slides into summaries, structure, search, and spaced-repetition materials with a single click.