PocketMCP – Text-based PDF (Phase 1)

PocketMCP – Text-based PDF (Phase 1)

This PDF is intentionally generated with selectable text so that your extractor can read it page-by-page. It describes the same ingestion flow used by PocketMCP: watch → extract → normalize → chunk → embed → index. The embedding model is Xenova/all-MiniLM-L6-v2, vectors are stored in SQLite + sqlite-vec.

What to Test

- Page-wise extraction (expect 2 pages)
- Basic punctuation and Unicode: en–dashes, “quotes”, emojis ▯
- Searches for keywords like MiniLM-L6-v2, sqlite-vec, chunk size 1000

PocketMCP – Text-based PDF (Phase 1): Page 2

Operational Guardrails

- PDF_MAX_PAGES should cap very large files.
- Encrypted PDFs should be skipped with a clear status.
- Scanned/image-only PDFs are marked needs_ocr (not processed in Phase 1).

Acceptance Checklist

1) Status=ok 2) Segments (pages) == 2 3) Queries for “sqlite-vec” and “MiniLM-L6-v2” return hits.