TEXTPRALINE - APP SUMMARY

WHAT IT IS
TextPraline is a deterministic text-cleaning layer for post-extraction content.
It normalizes noisy text from PDF/OCR/HTML pipelines into ingestion-ready text without rewriting meaning.

WHO IT'S FOR
Primary persona: RAG/ML engineers and data platform developers preparing extracted documents before chunking and embedding.

WHAT IT DOES (KEY FEATURES)
- Provides one-shot API: praline(text, ...), plus clean_text(...) and clean_lines(...).
- Auto-detects text profile: clean_web, pdf_like, ocr_like, or unknown.
- Removes extraction artifacts: glyph<...>, (cid:NNN), PUA chars, zero-width chars, BOM, soft hyphen.
- Normalizes Unicode/punctuation (NFKC, quotes/dashes, NBSP) and list bullets.
- Reduces layout noise using heuristics for axis garbage and vertical/single-character runs.
- Supports profiles: safe, markdown_safe, strict.
- Optional report mode returns PralineReport metrics (removed ToC/layout/repeated/boilerplate lines).

HOW IT WORKS (REPO-EVIDENCED ARCHITECTURE)
Data flow:
Raw text -> detect_text_profile -> extraction normalization -> invariant guardrails
-> line passes (boilerplate, layout noise, repeated lines opt-in, ToC/bullets)
-> whitespace + blank-line collapse -> final guardrails -> cleaned text (+ optional report).

Components/services:
- Public package export: textpraline.__init__ exposes praline and clean_text.
- Core pipeline: textpraline/cleaner/clean.py.
- Mapping tables: textpraline/cleaner/mappings.py.
- Tests: tests/basic_test.py and tests/test_pdf.py validate invariants/idempotence/PDF path.
- External services: Not found in repo.

HOW TO RUN (MINIMAL GETTING STARTED)
1) Use Python 3.10+ (declared in pyproject.toml).
2) Install project dependencies from pyproject.toml (exact install command: Not found in repo).
3) Quick validation: run tests, e.g. pytest tests/basic_test.py.
4) Minimal usage:
   from textpraline import praline
   cleaned = praline(raw_text)
5) CLI note: clean.py documents "textpraline < infile.txt > outfile.txt"; console-script wiring in pyproject.toml is Not found in repo.