Reference App – Updated Production Plan

Scope: Stage 1–3 (Classification → Extraction → Compliance)

Stack: Python, sklearn, deterministic ML

1. System Overview

- Three-stage pipeline: classification, extraction, compliance.
- Each stage independently testable and confidence-gated.
- Shadow-first deployment philosophy.

2. Stage 1 – Reference Type Classification (LOCKED)

- Input: raw reference string.
- Output: reference type (book, journal, website, conference, etc.).
- Gold dataset: 400 balanced references with adversarial variants.
- Model: TF-IDF (word + character n-grams) + Logistic Regression.
- Performance target: Macro F1 ≥ 0.90, Precision ≥ 0.95 at Tier-1.
- Model saved as stage1_reference_classifier.pkl.

3. Stage 2 – Structured Field Extraction (NEXT)

- Input: raw reference + reference type.
- Output: canonical structured fields (authors, year, title, DOI, etc.).
- Phase 1: deterministic rules and regex.
- Phase 2: optional ML-assisted extraction.
- Fail-closed on low-confidence extraction.

4. Stage 3 – Harvard Compliance & Violations

- Detect multi-label compliance issues.
- Implemented tags: missing_year, missing_access_date, missing_pages, missing_place, url_without_doi.
- Supports both rule-based and ML approaches.

5. Adversarial & Robustness Strategy

- Noise injection: punctuation removal, n.d. substitution, uppercase.
- Stress-tests robustness and calibration.
- Prevents brittle formatting assumptions.

6. Shadow Evaluation & Safety

- Shadow-only evaluation before user-facing rollout.
- Metrics: Precision, Recall, F1, calibration curves.
- Fail-closed behaviour on low confidence.

7. Immediate Next Actions

- Add confidence calibration (Platt scaling).
- Lock Stage 1 thresholds.
- Build Stage 2 gold extraction dataset.
- Extend Stage 3 violation taxonomy.
- Run full shadow evaluation.