PRD: Smart Chunking Pipeline for Vehicle Service Manual RAG

Revision History

Version	Date	Changes
1.0	2025-02-13	Initial PRD — single-manual (1999 XJ)
2.0	2025-02-13	Generalized to multi-manual architecture with manual profile system. Added CJ-5 Universal (1953–71) and M38A1 TM 9-8014 as target manuals.

---

1. Objective

Build a configurable chunking and indexing pipeline that transforms OCR'd vehicle service manuals — across manufacturers, eras, and document conventions — into a high-quality vector store optimized for a troubleshooting/repair chatbot. The pipeline must:

- Adapt to fundamentally different manual structures via a **manual profile** configuration system
- Preserve procedural integrity (never split a step sequence or separate a safety callout from its governed procedure)
- Enrich every chunk with structured metadata for filtered retrieval
- Handle varied OCR quality across different source scans
- Support a growing library of manuals within a single unified vector store

1.1 Target Manuals

Manual	Era	Publisher	Pages	Structure Type
1999 Jeep Cherokee (XJ)	Modern			

Factory Service Manual	(1999)	DaimlerChrysler	1,948	Chrysler Group/Section/Procedure
1953–71 Jeep Universal CJ Series Service Manual	Classic (1953–71)	Jeep Corporation	~400+	Lettered Section/Paragraph number
TM 9-8014: M38A1 / M170 Operator & Org. Maintenance	Military (1955)	Dept. of the Army / Air Force	~370	Military Chapter/Section/Paragraph

1.2 Success Criteria

Metric	Target
Chunk procedural integrity	100% — no step sequence is split across chunks
Safety callout attachment	100% — every WARNING/CAUTION/Note stays with its parent procedure
Vehicle/engine/drivetrain applicability tagging	≥90% of variant-specific chunks correctly tagged
Chunk size distribution	80% of chunks between 300–1500 tokens
Retrieval accuracy (manual spot-check)	Top-3 retrieval returns the correct procedure for ≥85% of a 50-query test set per manual
New manual onboarding time	<4 hours from PDF to indexed vector store using profile system

---

2. Comparative Structure Analysis

The three target manuals represent three fundamentally different document architectures. Understanding these differences drives the design of the profile system.

2.1 Structure Comparison

```
| Level 1: GROUP (numeric ID)          "0 Lubrication and Maintenance"|
|   Level 2: SECTION (ALL CAPS)        "SERVICE PROCEDURES"           |
|     Level 3: PROCEDURE (ALL CAPS)    "JUMP STARTING PROCEDURE"      |
|       Level 4: SUB-PROCEDURE         "TOWING-REAR END LIFTED"       |
|         Steps: (1), (2), (3)...                                      |
|                                                                      |
| Page format: {group_id} - {page_num}  e.g., "0 - 12"                |
| Variants: Domestic + International (suffix 'a')                      |
| Vehicles: Single (1999 Cherokee XJ)                                  |
| Engines: 2.5L I4, 4.0L I6, 2.5L Diesel                              |
├──────────────────────────────────────────────────────────────────────┤
| 1953-71 CJ Universal (Jeep Corp)                                     |
|                                                                      |
| Level 1: SECTION (letter ID)         "B Lubrication"                 |
|   Level 2: PARAGRAPH (Par. X-N)      "B-1. General"                  |
|     Level 3: SUB-PARAGRAPH            "B-4. Engine Lubrication..."   |
|       Steps: a., b., c. or (1), (2)                                  |
|                                                                      |
| Page format: Simple sequential page numbers                          |
| Variants: None (single market)                                       |
| Vehicles: Multiple (CJ-3B, CJ-5, CJ-5A, CJ-6, CJ-6A, DJ-5, DJ-6)|
| Engines: Hurricane F4-134 (I4), Dauntless V-6 225                    |
├──────────────────────────────────────────────────────────────────────┤
| TM 9-8014 M38A1 (US Army)                                            |
|                                                                      |
| Level 1: CHAPTER (numeric)           "Chapter 3"                     |
|   Level 2: SECTION (Roman numeral)   "Section II"                    |
|     Level 3: PARAGRAPH (integer)     "42. Starting the Engine"       |
|       Level 4: SUB-PARA              "a.", "b.", "(1)", "(2)"        |
|                                                                      |
| Page format: Simple sequential page numbers                          |
| Variants: None                                                       |
| Vehicles: Two (M38A1 utility, M170 ambulance)                        |
| Engine: Single (F-head I4)                                           |
└──────────────────────────────────────────────────────────────────────┘
```

2.2 Key Structural Differences

Feature	1999 XJ	CJ Universal	TM 9-8014
Top-level organizer	Numbered Groups (0, 2, 3, 5…)	Lettered Sections (A, B, C, D…)	Numbered Chapters (1, 2, 3, 4)
Procedure identifier	ALL CAPS heading	Paragraph number (B-7, D-4)	Sequential paragraph number (42, 43)

Step numbering	`(1)`, `(2)`, `(3)`	`a.`, `b.`, `(1)`, `(2)`	`a.`, `b.`, `(1)`, `(2)`
Figure references	`Fig. 1`, `Fig. 2`	`FIG. B-1`, `FIG. A-3`	`fig. 1`, `fig. 11`
Safety callout style	`WARNING:` / `CAUTION:` (ALL CAPS block)	Italicized cautions or inline notes	`Caution:` (sentence case) / `Note.`
Cross-references	"Refer to Group 9"	"Refer to Par. B-3" / "See Section D"	"par. 81b" / "fig. 11"
Page number format	`{group_id} - {page_num}`	Sequential integers	Sequential integers
Multi-vehicle handling	Single vehicle, engine variants	Model-specific callouts inline	Two vehicles, differences in §5
Table of contents	Per-group INDEX blocks + Group Tab Locator	Per-section CONTENTS with PAR. references	Master TOC with paragraph ranges
Spec tables	Inline within sections	End-of-section specification blocks	Inline tabulated data blocks
Maintenance schedules	Mileage-based Schedule A/B	Mileage-based Service Maintenance Schedule	Echelon-based PM services (A/B/C/D)

2.3 Common Patterns (Exploitable Across All Manuals)

Despite the structural differences, all three manuals share patterns that the core chunking engine can rely on:

1. **Hierarchical organization** — Every manual has a clear multi-level hierarchy, just with different naming conventions and identifiers.
2. **Self-contained procedures** — Repair/maintenance procedures are discrete units with a title, optional safety callouts, numbered steps, and figure references.
3. **Specification tables** — All manuals contain tabular data (capacities, torque values, dimensions) that must be kept atomic.

4. **Figure cross-references** — All use figure numbering (with different naming conventions) to reference diagrams.
5. **Internal cross-references** — All reference other sections/procedures within the manual.
6. **Vehicle variant callouts** — All distinguish between sub-models or configurations (engine, drivetrain, body style) at some level.
7. **Safety callout hierarchy** — All distinguish between critical safety warnings and informational notes, though the specific formatting varies.

---

3. Manual Profile System

The profile system is the abstraction layer that makes a single chunking engine work across different manual formats. Each manual gets a YAML profile that teaches the pipeline how to parse its structure.

3.1 Profile Schema

```
# manual_profile.yaml — Schema Definition

manual_id: string          # Unique identifier, e.g., "xj-1999", "cj-universal-53
manual_title: string       # Human-readable title
source_url: string         # Source PDF location
source_format: string      # "pdf-ocr" | "pdf-native" | "pdf-scanned"

# ── Vehicle Coverage ──────────────────────────────────────
vehicles:
  - model: string          # e.g., "Cherokee XJ", "CJ-5", "M38A1"
    years: string          # e.g., "1999", "1953-1971", "1952-1955"
    drive_type: string[]   # ["2WD", "4WD"] or ["4x4"]
    engines:
      - name: string       # e.g., "4.0L I6 MPI"
        code: string       # e.g., "ERH", "V-6 225", "F-head I4"
        aliases: string[]  # ["4.0L", "4.0", "six cylinder", "inline 6", "4.0 lit
    transmissions:
      - name: string
        code: string

# ── Document Structure ────────────────────────────────────
structure:
  hierarchy:               # Ordered list of hierarchy levels, top to bottom
    - level: 1
```

```
      name: string         # "group" | "section" | "chapter"
      id_pattern: regex    # Pattern to extract the level ID from text
      title_pattern: regex # Pattern to extract the level title
      known_ids:           # Optional: ground-truth list for validation
        - id: string
          title: string
    - level: 2
      name: string
      id_pattern: regex
      title_pattern: regex
    # ... etc

  page_number:
    pattern: regex         # How page numbers appear in headers/footers
    group_prefixed: bool   # Whether pages are numbered within groups/sections

  step_patterns:           # Ordered patterns for numbered/lettered steps
    - regex               # e.g., "^\((\d+)\)" for (1), (2), (3)
    - regex               # e.g., "^([a-z])\." for a., b., c.

  figure_reference:
    pattern: regex         # e.g., "Fig\.\s+(\d+)", "FIG\.\s+([A-Z]-\d+)"
    scope: string          # "global" (sequential across manual) | "per-section"

  cross_reference:
    patterns: regex[]      # Patterns that identify internal cross-refs
    # e.g., ["Refer to Group (\d+)", "Par\. ([A-Z]-\d+)", "par\.\s+(\d+[a-z]?)"]

# ── Safety Callouts ──────────────────────────────────────
safety_callouts:
  - level: "warning"       # Most severe
    pattern: regex         # e.g., "^WARNING:" or "^WARNING:"
    style: string          # "block" (multi-line ALL CAPS) | "inline" (sentence)
  - level: "caution"
    pattern: regex
    style: string
  - level: "note"
    pattern: regex
    style: string

# ── Content Types ────────────────────────────────────────
content_types:
  maintenance_schedule:
    present: bool
    structure: string      # "mileage-bands" | "echelon-based" | "interval-table"
    description: string    # How to chunk this specific schedule format
  wiring_diagrams:
```

```
    present: bool
    section_id: string     # Which top-level section contains them
  specification_tables:
    location: string       # "inline" | "end-of-section" | "dedicated-section"

# ── OCR Cleanup ──────────────────────────────────────────────
ocr_cleanup:
  quality_estimate: string # "good" | "fair" | "poor"
  known_substitutions:     # Manual-specific OCR errors observed
    - from: string
      to: string
  header_footer_patterns: regex[]  # Patterns to strip from page headers/footers
  garbage_detection:
    enabled: bool          # Whether to run non-English character detection
    threshold: float       # % non-ASCII chars to flag a line as garbage

# ── Market Variants ──────────────────────────────────────────
variants:
  has_market_variants: bool
  variant_indicator: string  # e.g., "suffix_a" for Group 0a, or "none"
  markets: string[]          # ["domestic", "international"] or ["US", "export"]
```

3.2 Profile: 1999 Jeep Cherokee XJ

```
manual_id: "xj-1999"
manual_title: "1999 Jeep Cherokee (XJ) Factory Service Manual"
source_url: "https://ia601507.us.archive.org/10/items/..."
source_format: "pdf-ocr"

vehicles:
  - model: "Cherokee XJ"
    years: "1999"
    drive_type: ["2WD", "4WD"]
    engines:
      - name: "2.5L I4 MPI Gasoline"
        code: "EPE"
        aliases: ["2.5L", "2.5", "four cylinder", "I4", "2.5 liter"]
      - name: "4.0L I6 MPI Gasoline"
        code: "ERH"
        aliases: ["4.0L", "4.0", "six cylinder", "inline 6", "I6", "4.0 liter"]
      - name: "2.5L I4 Diesel"
        code: "diesel"
        aliases: ["2.5L diesel", "diesel", "turbo diesel"]
    transmissions:
      - name: "AX5 5-Speed Manual"
        code: "DDQ"
```

```
      - name: "AW4 4-Speed Automatic"
        code: "DGS"

structure:
  hierarchy:
    - level: 1
      name: "group"
      id_pattern: "^(\\d+[A-Z]?[a-z]?)\\s"
      title_pattern: "^\\d+[A-Z]?[a-z]?\\s+(.+)"
      known_ids:
        - { id: "IN", title: "Introduction" }
        - { id: "0", title: "Lubrication and Maintenance" }
        - { id: "2", title: "Suspension" }
        - { id: "3", title: "Differential and Driveline" }
        - { id: "5", title: "Brakes" }
        - { id: "6", title: "Clutch" }
        - { id: "7", title: "Cooling System" }
        - { id: "8A", title: "Battery" }
        - { id: "8B", title: "Starting System" }
        - { id: "8C", title: "Charging System" }
        - { id: "8D", title: "Ignition System" }
        - { id: "8E", title: "Instrument Panel Systems" }
        - { id: "8F", title: "Audio Systems" }
        - { id: "8G", title: "Horn Systems" }
        - { id: "8H", title: "Vehicle Speed Control System" }
        - { id: "8J", title: "Turn Signal and Hazard Warning Systems" }
        - { id: "8K", title: "Wiper and Washer Systems" }
        - { id: "8L", title: "Lamps" }
        - { id: "8M", title: "Passive Restraint Systems" }
        - { id: "8N", title: "Electrically Heated Systems" }
        - { id: "8O", title: "Power Distribution Systems" }
        - { id: "8P", title: "Power Lock Systems" }
        - { id: "8Q", title: "Vehicle Theft/Security Systems" }
        - { id: "8R", title: "Power Seats Systems" }
        - { id: "8S", title: "Power Window Systems" }
        - { id: "8T", title: "Power Mirror Systems" }
        - { id: "8U", title: "Chime/Buzzer Warning Systems" }
        - { id: "8V", title: "Overhead Console Systems" }
        - { id: "8W", title: "Wiring Diagrams" }
        - { id: "9", title: "Engine" }
        - { id: "11", title: "Exhaust System and Intake Manifold" }
        - { id: "13", title: "Frame and Bumpers" }
        - { id: "14", title: "Fuel System" }
        - { id: "19", title: "Steering" }
        - { id: "21", title: "Transmission and Transfer Case" }
        - { id: "22", title: "Tires and Wheels" }
        - { id: "23", title: "Body" }
```

```
        - { id: "24", title: "Heating and Air Conditioning" }
        - { id: "25", title: "Emission Control Systems" }
    - level: 2
      name: "section"
      id_pattern: "^([A-Z][A-Z ]{3,})$"
      title_pattern: "^([A-Z][A-Z ]{3,})$"
    - level: 3
      name: "procedure"
      id_pattern: null
      title_pattern: "^([A-Z][A-Z \\-\\/\\(\\)]{5,})$"
    - level: 4
      name: "sub-procedure"
      id_pattern: null
      title_pattern: "^([A-Z][A-Z \\-]+\\([A-Z ]+\\))$"

  page_number:
    pattern: "(?:XJ\\s+)?(?:[A-Z ]+)\\s+(\\d+[A-Z]?)\\s*-\\s*(\\d+)"
    group_prefixed: true

  step_patterns:
    - "^\\((\\d+)\\)\\s"

  figure_reference:
    pattern: "\\(Fig\\.\\s+(\\d+)\\)"
    scope: "per-section"

  cross_reference:
    patterns:
      - "Refer to Group (\\d+[A-Z]?)"
      - "Refer to (Section \\d+)"

safety_callouts:
  - level: "warning"
    pattern: "^WARNING:"
    style: "block"
  - level: "caution"
    pattern: "^CAUTION:"
    style: "block"
  - level: "note"
    pattern: "^NOTE:"
    style: "inline"

content_types:
  maintenance_schedule:
    present: true
    structure: "mileage-bands"
    description: "Two schedules (A and B) with mileage intervals. Chunk by bands:
```

```
  wiring_diagrams:
    present: true
    section_id: "8W"
  specification_tables:
    location: "inline"

ocr_cleanup:
  quality_estimate: "fair"
  known_substitutions:
    - { from: "IJURY", to: "INJURY" }
    - { from: "Mopart", to: "Mopar" }
  header_footer_patterns:
    - "^XJ\\s+[A-Z ]+\\d+[A-Z]?\\s*-\\s*\\d+"
    - "^\\d+[A-Z]?\\s*-\\s*\\d+\\s+[A-Z ]+XJ$"
    - "\\(Continued\\)$"
  garbage_detection:
    enabled: true
    threshold: 0.5

variants:
  has_market_variants: true
  variant_indicator: "suffix_a"
  markets: ["domestic", "international"]
```

3.3 Profile: 1953–71 Jeep Universal CJ Series

```
manual_id: "cj-universal-53-71"
manual_title: "1953-71 Jeep Universal Series Service Manual (SM-1046)"
source_url: "https://cjclub.co.il/files/53-71_CJ5.pdf"
source_format: "pdf-ocr"

vehicles:
  - model: "CJ-3B"
    years: "1953-1971"
    drive_type: ["4WD"]
    engines:
      - name: "Hurricane F4-134"
        code: "F4"
        aliases: ["F4-134", "Hurricane", "four cylinder", "F4", "134 cubic inch",
  - model: "CJ-5"
    years: "1955-1971"
    drive_type: ["4WD"]
    engines:
      - name: "Hurricane F4-134"
        code: "F4"
        aliases: ["F4-134", "Hurricane", "four cylinder", "F4"]
```

```
    - name: "Dauntless V-6 225"
      code: "V6"
      aliases: ["V-6", "V6", "Dauntless", "V-6 225", "225", "V6-225"]
- model: "CJ-5A"
  years: "1964-1971"
  drive_type: ["4WD"]
  engines:
    - name: "Hurricane F4-134"
      code: "F4"
      aliases: ["F4-134", "Hurricane"]
    - name: "Dauntless V-6 225"
      code: "V6"
      aliases: ["V-6", "Dauntless"]
- model: "CJ-6"
  years: "1956-1971"
  drive_type: ["4WD"]
  engines:
    - name: "Hurricane F4-134"
      code: "F4"
      aliases: ["F4-134", "Hurricane"]
    - name: "Dauntless V-6 225"
      code: "V6"
      aliases: ["V-6", "Dauntless"]
- model: "CJ-6A"
  years: "1964-1971"
  drive_type: ["4WD"]
  engines:
    - name: "Hurricane F4-134"
      code: "F4"
      aliases: ["F4-134", "Hurricane"]
    - name: "Dauntless V-6 225"
      code: "V6"
      aliases: ["V-6", "Dauntless"]
- model: "DJ-5"
  years: "1965-1971"
  drive_type: ["2WD"]
  engines:
    - name: "Hurricane F4-134"
      code: "F4"
      aliases: ["F4-134", "Hurricane"]
- model: "DJ-6"
  years: "1965-1971"
  drive_type: ["2WD"]
  engines:
    - name: "Hurricane F4-134"
      code: "F4"
      aliases: ["F4-134", "Hurricane"]
```

```yaml
structure:
  hierarchy:
    - level: 1
      name: "section"
      id_pattern: "^([A-Z])\\s"
      title_pattern: "^[A-Z]\\s+(.+)"
      known_ids:
        - { id: "A", title: "General Data" }
        - { id: "B", title: "Lubrication and Periodic Services" }
        - { id: "C", title: "Tune-Up" }
        - { id: "D", title: "Hurricane F4 Engine" }
        - { id: "D1", title: "Dauntless V-6 Engine" }
        - { id: "E", title: "Fuel System" }
        - { id: "F", title: "Exhaust System" }
        - { id: "F1", title: "Exhaust Emission Control System F4 Engine" }
        - { id: "F2", title: "Exhaust Emission Control System V6-225 Engine" }
        - { id: "G", title: "Cooling System" }
        - { id: "H", title: "Electrical" }
        - { id: "I", title: "Clutch" }
        - { id: "J", title: "3-Speed Transmission" }
        - { id: "J1", title: "4-Speed Transmission" }
        - { id: "K", title: "Transfer Case" }
        - { id: "L", title: "Propeller Shafts" }
        - { id: "M", title: "Front Axle" }
        - { id: "N", title: "Rear Axle" }
        - { id: "O", title: "Steering" }
        - { id: "P", title: "Brakes" }
        - { id: "Q", title: "Wheels" }
        - { id: "R", title: "Frame" }
        - { id: "S", title: "Springs and Shock Absorbers" }
        - { id: "T", title: "Body" }
        - { id: "U", title: "Miscellaneous" }
    - level: 2
      name: "paragraph"
      id_pattern: "^([A-Z]\\d?-\\d+)\\.\\s"
      title_pattern: "^[A-Z]\\d?-\\d+\\.\\s+(.+)"
    - level: 3
      name: "sub-paragraph"
      id_pattern: null
      title_pattern: null  # Sub-paragraphs are not consistently titled

  page_number:
    pattern: "^(\\d+)$"
    group_prefixed: false

  step_patterns:
```

```
    - "^([a-z])\\.\\s"
    - "^\\((\\d+)\\)\\s"

  figure_reference:
    pattern: "FIG\\.\\s+([A-Z]\\d?-\\d+)"
    scope: "per-section"

  cross_reference:
    patterns:
      - "Refer to Par\\.\\s+([A-Z]\\d?-\\d+)"
      - "(?:See|Refer to)\\s+Section\\s+([A-Z]\\d?)"
      - "Refer to the Lubrication Chart"
      - "Fig\\.\\s+([A-Z]\\d?-\\d+)"

safety_callouts:
  - level: "caution"
    pattern: "^Caution:"
    style: "inline"
  - level: "note"
    pattern: "^Note:"
    style: "inline"
  # This era manual has fewer formal WARNING blocks;
  # safety info is often embedded in procedure prose

content_types:
  maintenance_schedule:
    present: true
    structure: "interval-table"
    description: "Single service maintenance schedule as a table with mileage col
  wiring_diagrams:
    present: true
    section_id: "H"
  specification_tables:
    location: "end-of-section"

ocr_cleanup:
  quality_estimate: "fair"
  known_substitutions:
    - { from: "'Jeep'", to: "Jeep" }
    - { from: "Jeep?", to: "Jeep" }
    - { from: "UNIVERSA L", to: "UNIVERSAL" }
    - { from: "SERIE S", to: "SERIES" }
    - { from: "LUBRICATIO N", to: "LUBRICATION" }
    - { from: "GENERA L", to: "GENERAL" }
    - { from: "SERVIC E", to: "SERVICE" }
  header_footer_patterns:
    - "^'?Jeep[*']?\\s+UNIVERSA?L\\s+SERIE?S\\s+SERVIC?E\\s+MANUAL"
```

```
    - "^[A-Z]\\s+[A-Z ]+$"
  garbage_detection:
    enabled: true
    threshold: 0.4

variants:
  has_market_variants: false
  variant_indicator: "none"
  markets: ["US"]
```

3.4 Profile: TM 9-8014 M38A1 / M170

```
manual_id: "tm9-8014-m38a1"
manual_title: "TM 9-8014: Operation and Organizational Maintenance — 1/4-Ton 4x4
source_url: "http://www.pedros.cz/M38A1/TM9-8014.pdf"
source_format: "pdf-ocr"

vehicles:
  - model: "M38A1"
    years: "1952-1957"
    drive_type: ["4WD"]
    engines:
      - name: "F-head I4 (Hurricane)"
        code: "F-head"
        aliases: ["F-head", "four-cylinder", "Hurricane"]
    transmissions:
      - name: "T-90 3-Speed Synchromesh"
        code: "T-90"
  - model: "M170"
    years: "1953-1957"
    drive_type: ["4WD"]
    engines:
      - name: "F-head I4 (Hurricane)"
        code: "F-head"
        aliases: ["F-head", "four-cylinder", "Hurricane"]
    transmissions:
      - name: "T-90 3-Speed Synchromesh"
        code: "T-90"

structure:
  hierarchy:
    - level: 1
      name: "chapter"
      id_pattern: "^CHAPTER\\s+(\\d+)"
      title_pattern: "^CHAPTER\\s+\\d+\\.\\s+(.+)"
      known_ids:
```

```
        - { id: "1", title: "Introduction" }
        - { id: "2", title: "Operating Instructions" }
        - { id: "3", title: "Organizational Maintenance Instructions" }
        - { id: "4", title: "Shipment and Limited Storage and Destruction of Mate
    - level: 2
      name: "section"
      id_pattern: "^Section\\s+([IVXLC]+)\\.?"
      title_pattern: "^Section\\s+[IVXLC]+\\.?\\s+(.+)"
    - level: 3
      name: "paragraph"
      id_pattern: "^(\\d+)\\.\\s"
      title_pattern: "^\\d+\\.\\s+(.+)"
    - level: 4
      name: "sub-paragraph"
      id_pattern: "^([a-z])\\.\\s"
      title_pattern: null

  page_number:
    pattern: "^(\\d+)$"
    group_prefixed: false

  step_patterns:
    - "^([a-z])\\.\\s"
    - "^\\((\\d+)\\)\\s"

  figure_reference:
    pattern: "fig\\.?\\s+(\\d+)"
    scope: "global"

  cross_reference:
    patterns:
      - "par\\.?\\s+(\\d+[a-z]?)"
      - "pars?\\.?\\s+(\\d+)\\s+(?:through|and)\\s+(\\d+)"
      - "fig\\.?\\s+(\\d+)"
      - "table\\s+([IVXLC]+|\\d+)"

safety_callouts:
  - level: "warning"
    pattern: "^Warning:"
    style: "inline"
  - level: "caution"
    pattern: "^Caution:"
    style: "inline"
  - level: "note"
    pattern: "^Note[\\.:]"
    style: "inline"
```

```
content_types:
  maintenance_schedule:
    present: true
    structure: "echelon-based"
    description: "Military PM services organized by echelon (A=daily, B=weekly, C
  wiring_diagrams:
    present: true
    section_id: "Chapter 3, Section XVI-XIX"
  specification_tables:
    location: "inline"

ocr_cleanup:
  quality_estimate: "poor"
  known_substitutions:
    - { from: "TECHNIG~MANUAL", to: "TECHNICAL MANUAL" }
    - { from: "CHAPTEIR", to: "CHAPTER" }
    - { from: "Generd", to: "General" }
    - { from: "persomlel", to: "personnel" }
    - { from: "requirelnents", to: "requirements" }
    - { from: "materiak", to: "materials" }
    - { from: "inst.rument", to: "instrument" }
    - { from: "t.he", to: "the" }
    - { from: "operat.o", to: "operate" }
    - { from: "e.ngine", to: "engine" }
  header_footer_patterns:
    - "^\\*?TM\\s+9-8014"
    - "^\\d{5,}-\\d{2}-\\d+"
  garbage_detection:
    enabled: true
    threshold: 0.3

variants:
  has_market_variants: false
  variant_indicator: "none"
  markets: ["US military"]
```

3.5 LLM-Bootstrapped Profile Generation

For new manuals not yet profiled, an LLM can bootstrap a draft profile:

Input: First 20–30 pages of the manual (cover, TOC, introduction, first procedure section)

Prompt:

```
Analyze these pages from a vehicle service manual and generate a manual profile
in YAML format following this schema: [schema from §3.1]
```

```
Specifically identify:
1. The hierarchical structure (how many levels, naming convention, ID patterns)
2. Step numbering conventions
3. Figure reference format
4. Safety callout formatting
5. Internal cross-reference patterns
6. Page numbering format
7. Vehicle models and engine variants covered
8. OCR artifacts and quality issues observed in this specific scan
9. Maintenance schedule structure (if visible in these pages)

Output a complete YAML profile. Flag any fields where you're uncertain with a
comment "# NEEDS VERIFICATION".
```

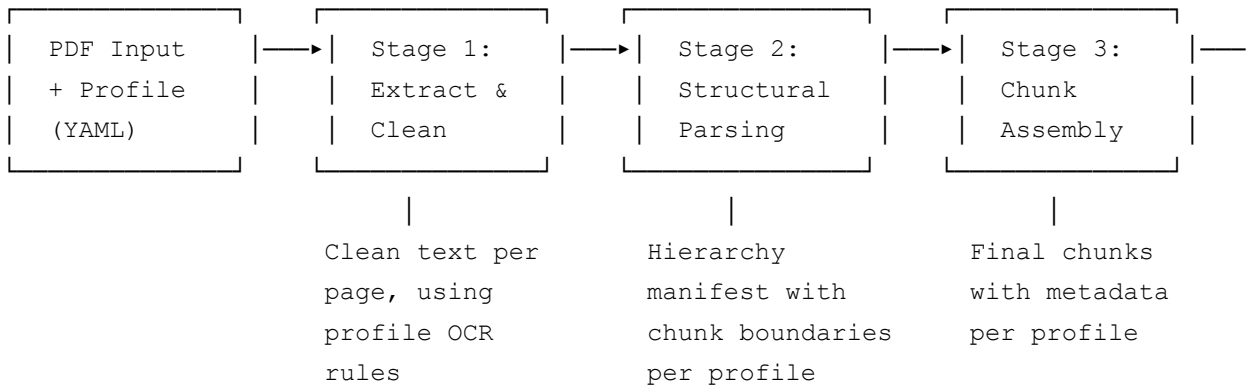
Model: Frontier model (Claude Sonnet or Opus) for profile generation — this is a one-time, high-judgment task. The profile is then reviewed and corrected by a human before the pipeline runs.

Estimated time: 15–30 minutes of LLM generation + 30–60 minutes of human review = under 2 hours to profile a new manual.

---

4. Pipeline Architecture

4.1 High-Level Flow



4.2 Stage 1: Text Extraction & OCR Cleanup (Profile-Driven)

Input: Source PDF + manual profile
Output: Clean text per page

4.2.1 PDF Text Extraction

Use `pymupdf` (fitz) for all manuals. Text extraction is format-independent.

4.2.2 OCR Cleanup (Profile-Driven)

The cleanup engine reads from the profile's `ocr_cleanup` section:

1. **Apply `known_substitutions`** — Simple find/replace pairs specific to this scan.
2. **Strip `header_footer_patterns`** — Regex patterns for page headers/footers. Capture page number metadata before stripping.
3. **Run `garbage_detection`** if enabled — Flag lines exceeding the profile's non-ASCII threshold.
4. **Universal cleanup** (applied to all manuals regardless of profile):
   - Smart quote → straight quote normalization
   - Ligature decomposition
   - Whitespace normalization (collapse multiple spaces, normalize line breaks)
   - Preserve structural markers (step numbers, figure refs, safety callouts — all identified by profile patterns)

4.2.3 OCR Quality Assessment

After cleanup, run a quality sampling pass on 50 random pages:

- % of words matching an English dictionary
- % of lines flagged by garbage detection
- Count of remaining suspected OCR errors

If quality is below threshold (`<85%` dictionary match rate), flag the manual for re-OCR (Tesseract 5 or commercial OCR) before proceeding.

4.3 Stage 2: Structural Parsing (Profile-Driven)

Input: Clean text pages + manual profile
Output: Hierarchical manifest defining chunk boundaries

4.3.1 Header Detection (Profile-Driven)

The parser reads the profile's `structure.hierarchy` to identify structural elements at each level:

```
# Pseudocode
for level_config in profile.structure.hierarchy:
    for line in page_text:
        if re.match(level_config.id_pattern, line):
            record_boundary(level=level_config.level,
                            id=extract_id(line, level_config.id_pattern),
                            title=extract_title(line, level_config.title_pattern))
```

If the profile includes `known_ids` for a level, validate detected boundaries against the ground truth and flag unrecognized IDs.

4.3.2 LLM-Assisted Parsing (Fallback)

For pages where heuristic parsing fails, feed 3–5 pages at a time to a local LLM with the profile's hierarchy definition as context. Model: Llama 3.1 8B or Qwen 2.5 7B — pattern matching with context, not frontier reasoning.

4.3.3 Manifest Output Format


```
{
  "manual_id": "xj-1999",
  "manifest": [
    {
      "chunk_id": "xj-1999::0::SP::JSP",
      "level": 3,
      "level_name": "procedure",
      "title": "Jump Starting Procedure",
      "hierarchy_path": ["0 Lubrication and Maintenance", "SERVICE PROCEDURES", "
      "content_type": "procedure",
      "page_range": { "start": "0-9", "end": "0-10" },
      "line_range": { "start": 1842, "end": 1923 },
      "vehicle_applicability": ["Cherokee XJ"],
      "engine_applicability": ["all"],
      "drivetrain_applicability": ["all"],
      "has_safety_callouts": ["warning", "caution"],
      "figure_references": ["Fig. 1"],
      "cross_references": ["Group 8A"],
      "parent_chunk_id": "xj-1999::0::SP",
      "children": []
    }
  ]
```

```
}
```

Chunk ID format: `{manual_id}::{level1_id}::{level2_id}::{level3_id}` — namespaces chunks globally across all manuals.

4.4 Stage 3: Chunk Assembly (Universal Engine + Profile Metadata)

Input: Clean text pages + manifest + profile
Output: Final chunk objects with text and metadata

4.4.1 Chunk Boundary Rules (Universal — All Manuals)

Rule	Description
R1: Primary chunk unit	One complete procedure/topic at the lowest meaningful hierarchy level
R2: Size targets	Min 200 tokens, target 500–1500, max 2000 (hard ceiling)
R3: Never split steps	A numbered/lettered step sequence stays in one chunk. Detect steps using `profile.structure.step_patterns`
R4: Safety callout attachment	Safety callouts (detected via `profile.safety_callouts` patterns) always stay with their governed procedure. If a callout precedes an entire procedure, duplicate it in any split continuation chunks
R5: Table integrity	Specification tables are never split. Override size ceiling if needed
R6: Merge small chunks	Chunks <200 tokens merge with next sibling, or upward into parent
R7: Cross-ref-only content merges	Lines matching `profile.structure.cross_reference.patterns` that constitute an entire “section” get merged into the parent chunk
R8: Figure reference continuity	Figure references stay with the text describing them

4.4.2 Content Type Special Handling (Profile-Driven)

Maintenance Schedules — varies by `profile.content_types.maintenance_schedule.structure`:

Structure Type	Chunking Strategy
`mileage-bands` (XJ)	Split into bands (0–30K, 30–60K, 60–90K, 90–120K) + overview chunk per schedule
`interval-table` (CJ)	Keep full table as one chunk; lubrication charts as separate spec chunks
`echelon-based` (TM 9-8014)	One chunk per echelon level (A/B/C/D); lubrication order as separate chunk

Specification Tables — varies by `profile.content_types.specification_tables.location`:

Location	Behavior
`inline`	Chunk with surrounding procedure context
`end-of-section`	Chunk as standalone spec unit with section header prepended
`dedicated-section`	Chunk by individual table within the section

4.4.3 Vehicle Applicability Tagging

For multi-vehicle manuals (CJ series, TM 9-8014), scan chunk text for model-specific callouts:

1. Match against `profile.vehicles[].model` names and aliases
2. Match against `profile.vehicles[].engines[].aliases`
3. If no specific model/engine is mentioned, tag as `"all"`

4.4.4 Chunk Text Composition

Each chunk's text:

```
{hierarchical_header}

{content_body}
```

Hierarchical header (generated, not from source):

```
{manual_title} | {level1_title} | {level2_title} | {procedure_title}
```

Examples:

- `1999 Jeep Cherokee XJ | Lubrication and Maintenance | Service Procedures | Jump Starting Procedure`
- `Jeep Universal CJ Series (1953-71) | Lubrication | B-4. Engine Lubrication System — Hurricane F4 Engine`
- `TM 9-8014 M38A1/M170 | Operating Instructions | Section III. Operation Under Usual Conditions | 42. Starting the Engine`

4.4.5 Final Chunk Schema


```
{
  "chunk_id": "xj-1999::0::SP::JSP",
  "manual_id": "xj-1999",
  "text": "1999 Jeep Cherokee XJ | Lubrication and Maintenance | Service Procedur
  "metadata": {
    "manual_id": "xj-1999",
    "manual_title": "1999 Jeep Cherokee (XJ) Factory Service Manual",
    "vehicle_models": ["Cherokee XJ"],
    "vehicle_years": "1999",
    "hierarchy_path": ["Lubrication and Maintenance", "Service Procedures", "Jump
    "level1_id": "0",
    "level1_name": "Lubrication and Maintenance",
    "level2_name": "Service Procedures",
    "procedure_name": "Jump Starting Procedure",
    "content_type": "procedure",
    "page_range": "0-9 to 0-10",
    "market": "domestic",
    "engine_applicability": ["2.5L", "4.0L"],
    "drivetrain_applicability": ["2WD", "4WD"],
    "safety_callout_levels": ["warning", "caution"],
    "figure_references": ["Fig. 1"],
    "cross_references": ["Group 8A"],
    "parent_chunk_id": "xj-1999::0::SP",
    "sibling_chunk_ids": ["xj-1999::0::SP::TR", "xj-1999::0::SP::HR"],
    "token_count": 847
  }
}
```

4.5 Stage 4: Embedding & Indexing

4.5.1 Embedding Input

```
{hierarchical_header}\n\n{first_150_words_of_body}
```

Front-loads semantic signal. Full text stored for retrieval; embedding from condensed input to reduce noise.

4.5.2 Embedding Model

Recommendation: `nomic-embed-text` v1.5 via Ollama for local deployment.

4.5.3 Unified Vector Store

Single Qdrant collection with all manuals. Filterable metadata indexes:

- `manual_id` — scope to a specific manual
- `vehicle_models` — filter by vehicle
- `vehicle_years` — filter by year/era
- `content_type` — procedure vs. specification vs. diagnostic
- `engine_applicability` — engine-specific filtering
- `drivetrain_applicability` — 2WD/4WD filtering
- `level1_id` — filter by major system/group/chapter

4.5.4 Secondary Metadata Index (SQLite)

Cross-manual lookup:

- `(manual_id, procedure_name)` → chunk IDs
- `(manual_id, level1_id)` → all child chunk IDs
- `(manual_id, figure_ref)` → chunk IDs
- `(manual_id, cross_ref_target)` → chunk IDs
- `vehicle_model` → all applicable chunk IDs

---

5. Query-Time Retrieval Strategy

5.1 Query Understanding

Before retrieval, identify:

1. **Vehicle scope** — Apply `manual_id` or `vehicle_models` filter
2. **System scope** — Apply `level1_id` filter
3. **Engine/drivetrain scope** — Apply appropriate filter
4. **Query type** — "how do I" (procedure), "what is the spec" (specification), troubleshooting (diagnostic)

If no vehicle specified with multiple manuals in store: search across all and include `manual_title` in context for disambiguation.

5.2 Retrieval Flow

1. Embed query → ANN search, top-k = 10, with metadata filters
2. Parent-chunk enrichment (section overview)
3. Sibling-chunk enrichment (adjacent procedures above threshold)
4. Cross-reference resolution via secondary index
5. Re-rank → top-3 to top-5 to the LLM

5.3 Safety Callout Handling

If any retrieved chunk has `safety_callout_levels` containing `"warning"` or `"caution"`:

- Surface warnings prominently in response
- Never summarize or paraphrase safety warnings
- Present before procedure steps

5.4 Multi-Manual Awareness

When chunks from multiple manuals are retrieved:

- Always attribute which manual a procedure is from
- Note that procedures from different eras may conflict
- Prefer the manual matching the user's stated vehicle

6. Quality Assurance

6.1 Chunk Validation (Universal)

Check	Rule	Action
Orphaned steps	No chunk starts mid-sequence (per profile `step_patterns`)	Flag for review
Split safety callouts	No safety callout at chunk start without preceding context	Re-attach to previous chunk
Size outliers	Flag chunks < 100 or > 3000 tokens	Review for merge/split
Empty metadata	Every chunk needs `manual_id`, `level1_id`, `content_type`	Flag for enrichment
Duplicate content	Cosine similarity > 0.95 within same manual	Review for dedup
Cross-ref validity	Every target resolves to a real chunk ID	Flag broken refs
Profile validation	All Level 1 IDs match `known_ids`	Flag unrecognized sections

6.2 Per-Manual Test Sets (50 queries each)

- Direct procedure lookups (15)
- Specification lookups (10)
- Maintenance schedule queries (10)
- Diagnostic/troubleshooting (10)
- Ambiguous/natural language (5)

6.3 Cross-Manual Test Set (20 queries)

Tests multi-manual disambiguation, model-specific routing, and handling of queries that span eras.

7. Technology Stack

Component	Tool	Deployment
PDF text extraction	`pymupdf` (fitz)	Python, local
OCR cleanup	Python + regex + profile rules	Local
Structural parsing	Heuristic parser + local LLM fallback	Docker on home server
Profile generation	Frontier LLM (Claude Sonnet/Opus)	API call (one-time)
Chunk assembly	Python, profile-driven	Local
Embedding	`nomic-embed-text` v1.5 via Ollama	Docker on home server
Vector store	Qdrant	Docker on home server
Secondary index	SQLite	Local file
Pipeline CLI	Python with profile path argument	Local

7.1 CLI Interface

```
# Process a single manual
python pipeline.py process --profile profiles/xj-1999.yaml --pdf data/xj-manual.p

# Bootstrap a profile from a new manual
python pipeline.py bootstrap-profile --pdf data/new-manual.pdf --output profiles/

# Validate a profile against its PDF
python pipeline.py validate --profile profiles/cj-universal.yaml --pdf data/cj-ma

# Run QA checks on an indexed manual
python pipeline.py qa --manual-id xj-1999 --test-set tests/xj-1999-queries.json
```

7.2 Docker Compose

```
services:
  qdrant:
    image: qdrant/qdrant:latest
    ports:
      - "6333:6333"
    volumes:
```

```
      - qdrant_data:/qdrant/storage

  ollama:
    image: ollama/ollama:latest
    ports:
      - "11434:11434"
    volumes:
      - ollama_models:/root/.ollama
    deploy:
      resources:
        reservations:
          devices:
            - capabilities: [gpu]

  pipeline:
    build: ./pipeline
    volumes:
      - ./profiles:/profiles
      - ./data:/data
      - ./output:/output
      - ./tests:/tests
    depends_on:
      - qdrant
      - ollama

volumes:
  qdrant_data:
  ollama_models:
```

---

8. Implementation Phases

Phase 1: Profile System & Core Parser (2–3 days)

- YAML profile schema and loader
- Profile-driven OCR cleanup engine
- Profile-driven structural parser (heuristic)
- Generate manifests for all three manuals
- Validate against known section lists

Phase 2: Chunk Assembly Engine (2–3 days)

- Universal chunk boundary rules (R1–R8)
- Profile-driven content type handlers
- Vehicle/engine applicability tagger
- Chunk validation suite
- Process all three manuals

Phase 3: Embedding & Unified Index (1 day)

- Hierarchical header embedding composition
- Ollama + nomic-embed-text integration
- Qdrant collection with all three manuals
- SQLite secondary index

Phase 4: Retrieval & QA (2–3 days)

- Query understanding layer
- Metadata-filtered retrieval pipeline
- Parent/sibling/cross-reference enrichment
- Build test sets (50 per manual + 20 cross-manual)
- Measure and iterate

Phase 5: Profile Bootstrap Tooling (1 day)

- LLM prompt for profile generation
- Profile validation tool
- Documentation for adding new manuals

Phase 6: Chatbot Integration (2 days)

- System prompt with safety callout handling
- Source attribution
- Multi-manual disambiguation UX
- End-to-end testing

---

9. Open Questions & Decisions

#	Question	Impact	Default Assumption
1	Single collection or per-manual collections?	Architecture	Single with `manual_id` filter
2	How to handle queries matching multiple manuals?	UX	Return all with clear attribution
3	Should profiles be versioned?	Maintenance	Yes, in git
4	Is re-OCR needed for TM 9-8014?	Quality	Try cleanup first; re-OCR if Recall@3 < 70%
5	Should pipeline support non-Jeep manuals?	Scope	Profile system is manufacturer-agnostic by design; validate with a non-Jeep manual before claiming generality
6	What chatbot LLM?	Response quality	TBD after retrieval accuracy measured
7	Should CJ lubrication charts get image treatment?	Quality	Text first; manual-verify critical numerics
8	How to handle CJ manual covering 18 years of changes?	Accuracy	Tag chunks with year-range applicability where noted; default to “all years”

---

10. Risks & Mitigations

Risk	Likelihood	Impact	Mitigation
Profile patterns don’t generalize to unseen formats	Medium	High	Explicit profiles over clever heuristics. LLM bootstrap + human review.
OCR quality varies within a single manual	High	Medium	Per-page quality scoring. Flag low-quality pages for review or re-OCR.

Military paragraph numbering collides with step numbering	Medium	Medium	Check hierarchy patterns in level order; disambiguate by context window.
Multi-vehicle manuals produce ambiguous applicability tags	High	Medium	Conservative: if chunk doesn't exclude a model, tag as "all". Over-retrieve > miss.
Spec tables lose numeric accuracy through OCR	High	High for specs	Critical specs need human verification. "Verified" flag in metadata.
Chunk IDs collide across manuals	Low	High	Namespaced IDs (`{manual_id}::{path}`) prevent by design.
Profile bootstrap LLM hallucinates structure	Medium	Low	All bootstrapped profiles require human review. `# NEEDS VERIFICATION` flags.
CJ manual's 18-year span has undocumented mid-production changes	Medium	Medium	Note in chatbot responses that specs may vary by production year within the covered range.