PDF Extraction Flaws: The Four Horsemen Demonstration

The Post-Processing Pipeline Case Study

November 2025

The Need for Cleaning

This document is engineered to display the common problems encountered when naively extracting text from a PDF. A simple text extraction tool, which only cares about the physical coordinates of the characters, will fail to read the content in the correct logical order and structure.

Flaw 1: Unwanted Hyphenation

When the text extractor reads this line, it will often fail to intelligently combine the two parts of the split word. This happens even with the simplest of layouts. The word below is designed to be broken right at the column edge, simulating an un- broken hyphenation that should have been joined: “unintelligently” is split as “un-” and “intelligently”.

Flaw 2: Forced Line Breaks

A core document is designed for print, not for a continuous digital stream. This means that a single sentence, regardless of its length, is often broken into physical lines. When a naive extractor pulls the text, it treats every physical line break as a newline character ($\backslash n$), even if the sentence continues. The result is often choppy, incoherent, and extremely difficult for a large language model to process correctly. This paragraph is a perfect example of what a raw extractor would produce as one long string separated by meaningless $\backslash n$ breaks, failing to recognize that this is all part of one smooth, coherent narrative that should be joined with a single space. **(Raw Output Simulation: ”This paragraph is a perfect example of what a raw extractor would produce as one long string\n separated by meaningless \n breaks, failing to recognize that this is all part of one smooth, coherent narrative that\n should be joined with a single space.”)**

Flaw 3: Headers and Footers Intrusion

Look closely at the document flow. If an extraction tool pulls text based solely on vertical position on the page, the header at the top right, which reads **CONFIDENTIAL - DRAFT 1.0**, will inevitably land right in the middle of our extracted text block from this section title and the paragraph below. We must filter out this noise.

Example of Extraction Corruption:

```
Look closely at the document flow. CONFIDENTI
If an extraction tool pulls text based solely
position on the page, the header at the top r
```

Flaw 4: Multi-Column Mayhem

This section is the most notorious offender. Since this document uses a two-column layout, the text flows vertically down the first column and then jumps to the top of the second column. A naive extractor will read all the text in the first column, hit the bottom, and then jump to the top of the second column. If the text here discusses the initial setup of a large, independently-running system component, which happens to be split across the column boundary, the output will be nonsensical. The beginning of the sentence will be immediately followed by the end of the next logical paragraph.