## 📄 Assignment Statement

USB Power Delivery (USB PD) Specification Parsing and Structuring System

Title: Intelligent Parsing & Structuring of USB PD Document

Context:

USB Power Delivery (USB PD) specifications are complex technical documents containing a mix of structured and unstructured content including hierarchical sections (like chapters and subsections), figures, tables, and protocol/state machine descriptions.

To convert the raw PDF doc into structured, machine-readable formats (like JSONL), preserving their logical hierarchy and metadata.

Objective:

You are tasked with building a prototype system that parses a USB PD specification PDF file and produces structured JSONL output representing the Table of Contents (ToC) hierarchy with associated page numbers and metadata.

Deliverables:

* Python script(s) that:  
  + Extract the Table of Contents from the PDF
  + Define the json schema for ToC section and json schema for other sections in the doc
  + Generate a JSONL output file for ToC
  + Generate a JSONL output file for all other sections in the pdf
  + Generate validation report in xls format comparing the sections/subsections, tables Count in TOC and the actual output json file
    - Total sections in ToC vs parsed
    - Mismatches, order errors, gaps
* A sample JSONL files (named: usb\_pd\_spec.jsonl, usb\_pd\_toc.jsonl, usb\_pd\_metadata.jsonl) containing the parsed output
* README (or code comments) explaining how your script works

🧾 Sample Output Format (JSONL):

Each line in the file should look like this:

{"doc\_title": "USB PD Specification Rev X", "section\_id": "2.1.2", "title": "Power Delivery Contract Negotiation", "page": 53, "level": 3, "parent\_id": "2.1", "full\_path": "2.1.2 Power Delivery Contract Negotiation"}

## Support:

* If you have technical questions, reach out over email for a quick discussion.
* You can also email questions if asynchronous is preferred.
* Bonus points for creating reusable functions and robust error handling.

## Timeline:

Estimated effort: 2–5 working days

## 

Sample Prompt:

You are an expert document parser. Given the raw extracted front matter text from a technical specification (usually the first 3–10 pages), extract the Table of Contents in structured JSONL format and then extract all other sections of the document

## JSONL schema and a sample file

Refer below structure for a USB PD specification document — based on your provided Table of Contents.

This schema is optimized for:

* Retrieval, search, and hierarchical understanding
* Easy ingestion into vector stores or LLM-based document agents
* Consistency across all chapters, sections, and subsections

✅ JSONL Format Overview (Each line = one entry/section)

Each line (object) includes:

| **Field** | **Type** | **Description** |
| --- | --- | --- |
| section\_id | string | Hierarchical section identifier (e.g., "2.1.2") |
| title | string | Section title (without numbering) |
| page | integer | Starting page number of the section |
| level | integer | Depth level (chapter = 1, section = 2, etc.) |
| parent\_id | string/null | Immediate parent section (null for top level) |
| full\_path | string | Concatenation of section\_id and title (e.g., "2.1.2 Power Delivery Contract Negotiation") |
| doc\_title | string | Document name or version for reference |
| tags | list | Optional: semantic labels (e.g., ["negotiation", "contracts"]) |

—

📄 Sample JSONL File (usb\_pd\_spec.jsonl)

{"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2", "title": "Overview", "full\_path": "2 Overview", "page": 53, "level": 1, "parent\_id": null, "tags": []}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.1", "title": "Introduction", "full\_path": "2.1 Introduction", "page": 53, "level": 2, "parent\_id": "2", "tags": []}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.1.1", "title": "Power Delivery Source Operational Contracts", "full\_path": "2.1.1 Power Delivery Source Operational Contracts", "page": 53, "level": 3, "parent\_id": "2.1", "tags": ["contracts", "source"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.1.2", "title": "Power Delivery Contract Negotiation", "full\_path": "2.1.2 Power Delivery Contract Negotiation", "page": 53, "level": 3, "parent\_id": "2.1", "tags": ["contracts", "negotiation"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.1.3", "title": "Other Uses for Power Delivery", "full\_path": "2.1.3 Other Uses for Power Delivery", "page": 54, "level": 3, "parent\_id": "2.1", "tags": ["applications"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.2", "title": "Compatibility with Revision 2.0", "full\_path": "2.2 Compatibility with Revision 2.0", "page": 54, "level": 2, "parent\_id": "2", "tags": ["revision"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.3", "title": "USB Power Delivery Capable Devices", "full\_path": "2.3 USB Power Delivery Capable Devices", "page": 55, "level": 2, "parent\_id": "2", "tags": ["devices"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.4", "title": "SOP\* Communication", "full\_path": "2.4 SOP\* Communication", "page": 57, "level": 2, "parent\_id": "2", "tags": ["communication"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.4.1", "title": "Introduction", "full\_path": "2.4.1 Introduction", "page": 57, "level": 3, "parent\_id": "2.4", "tags": []}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.4.2", "title": "SOP\* Collision Avoidance", "full\_path": "2.4.2 SOP\* Collision Avoidance", "page": 57, "level": 3, "parent\_id": "2.4", "tags": ["collision", "avoidance"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.4.3", "title": "SOP Communication", "full\_path": "2.4.3 SOP Communication", "page": 57, "level": 3, "parent\_id": "2.4", "tags": ["communication"]}  
 {"doc\_title": "USB Power Delivery Specification Rev X", "section\_id": "2.4.4", "title": "SOP'/SOP'' Communication with Cable Plugs", "full\_path": "2.4.4 SOP'/SOP'' Communication with Cable Plugs", "page": 57, "level": 3, "parent\_id": "2.4", "tags": ["cable", "SOP'"]}

—

✅ Tips for Building This from PDF

* Extract text via pdfplumber or fitz
* Use regex to extract section IDs and titles:

Example regex:  
 ^(\d+(.\d+)\*)(\s+)([^\n.]+)(.+)\s+(\d+)$

* Use number of dots in section\_id to infer level
* Track parent\_id via section\_id truncation logic:  
  + 2.1.2 → parent = 2.1
  + 2 → parent = null